Selection and evolution of aggregation resistant proteins

Jessica Sarah Ebo

University of Leeds

Astbury Centre for Structural Molecular Biology

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

March 2021

Declaration

The candidate confirms that the submitted work is her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated overleaf. The candidate confirms the appropriate credit has been given within the thesis where reference has been made to the work of others. This copy has been supplied on the understanding that it is copyright material and that no quotation from this thesis may be published without proper acknowledgement.

Jointly authored publications

The candidate confirms that the work submitted is his/her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Data in this has been provided by Dr Janet Saunders who was previously affiliated with the University of Leeds and currently with AstraZeneca. The affiliation at which the work was performed is stated within the legend.

Chapters 3 and 4 include work from the following publication: Ebo JS*, Saunders JC*, Devine PWA, Gordon AM, Warwick AS, Schiffrin B, Chin SE, England E, Button JD, Lloyd C, Bond NJ, Ashcroft AE, Radford SE, Lowe DC, Brockwell DJ. An in vivo platform to select and evolve aggregation-resistant proteins. Nature Communications, 2020 (11) 1816.

* These authors contributed equally to this work.

For this publication, the author contribution statement reads:

J.S.E. and J.C.S designed and performed the in vivo assays and evolution assays. P.W.A.D. designed and performed cross-linking and mass spectrometry experiments. S.C., J.C.S. and J.S.E created mutagenic libraries. J.S.E., A.M.G. and A.S.W. performed the light chain evolution experiments. J.S.E, J.C.S., E.E., J.D.B. and C.L. performed the experiments on IgGs. J.S.E. and B.S. performed computational analysis of scFvs. S.E.R., N.J.B., A.E.A., D.C.L. and D.J.B. conceived and designed experiments. All authors contributed to manuscript preparation.

Acknowledgements

Firstly, I would like to thank my supervisors, Prof David Brockwell and Prof Sheena Radford, for their continued guidance and support throughout my PhD. Your enthusiasm and excitement kept me going, and I am extremely grateful for all your time and work that has gone into this project.

Thank you to the BBSRC and AstraZeneca for funding this project, and to everyone at AstraZeneca for their advice over the years, particularly Dr David Lowe and Dr Janet Saunders. Jan, thank you for your endless help throughout my PhD, this project would not have been possible without you!

It has been a pleasure to have been a member of the Radford and Brockwell groups and I am so thankful to every member that I have coincided with over the past four years. Thank you all for making it a joy to come to work and for always being there with biscuits, beer, or tequila. A special thanks go to Dr Paul White (for always singing and moaning with me), Sabine Ulamec (for your German honesty and bad influence) and Emily Byrd (for our common love of trash TV and Percy Pigs). Thank you to past members Drs Mathew Jackson, Esther Martin, Atenas Posada-Borbon, Anna Higgins and Julia Humes for all the trips to Caffe Nero/Old Bar over the years.

A huge thanks go to Nasir Khan; the lab would not have been the same place without you. Thank you for keeping the lab in order, and for dealing with 'Drama Queen Jess'. I will miss gossiping with you and I am eternally grateful for the hundreds of biscuits you have provided.

To the 'Radford Runners', I am so grateful for the motivation to run on winter evenings and for helping me to find the perfect stress relief. Thank you also to everyone in the Berry/Hemsworth groups for always being there for a gossip and a spare PCR machine!

Lastly, a huge thank you to the people who have supported me most. Oliver, thank you for being there on the good and bad days. Thank you for your patience during thesis writing and most importantly for distracting me with delicious food and beer. To my family, Mum, Dad and Rachel, thank you for always believing in me, I could not have done this without your love and support.

Table of contents

List of figures

List of tables

List of abbreviations

3D	Three dimensional
A3D	Aggrescan3D
AC-SINS	Affinity-capture self-interaction nanoparticle spectroscopy
AL	Light chain amyloidosis
APRs	Aggregation prone regions
ARM	Antibody-ribosome-mRNA
AUC	Analytical ultracentrifugation
BV	Baculovirus
CAT	Chloramphenicol acetyltransferase
CDR	Complementarity determining region
CHO	Chinese hamster ovary
CIC	Cross-interaction chromatography
Cryo-EM	Cryoelectron microscopy
DHFR	Dihydrofolate reductase
DLS	Dynamic light scattering
DNA	Deoxyribonucleic acid
DSC	Differential scanning calorimetry
DSF	Differential scanning fluorimetry
E. coli	*Escherichia coli*
ELISA	Enzyme-linked immunosorbent assay
Fab	Antigen binding fragments
FACS	Fluorescence activate cell sorting
Fc	Crystallisable fragment
FDA	Food and Drug Administration
Fv	Variable fragment
GCSF	Granulocyte colony-stimulating factor
GFP	Green fluorescent protein

GlcNAc	N-acetyl glucosamine
GS linker	Glycine-serine linker
HA	A haemagglutinin
HDX-MS	Hydrogen/deuterium exchange mass spectrometry
HEK293	Human embryonic kidney 293
HIC	Hydrophobic interaction chromatography
HTRF	Homogeneous time resolved fluorescence
Ig	Immunoglobulin
Im7	Immunity protein 7
KLD	Kinase, ligase and DpnI
LC	Light chain
mAbs	Monoclonal antibodies
MALS	Multi angle light scattering
MBP	Maltose binding protein
MCD_{GROWTH}	Maximal cell dilution allowing growth
MD	Molecular dynamic
MurNAc	N-acetylmuramic acid
NGF	Nerve growth factor
NGS	Next generation sequencing
PBPs	Penicillin binding proteins
PCA	Protein-fragment complementation assays
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
POI	Protein of interest
PTM	Post translational modification
RNA	Ribonucleic acid
RT-PCR	Reverse transcription PCR
S. cerevisiae	*Saccharomyces cerevisiae*
SAP	Spatial aggregation propensity

scFv	Single chain fragment variable
SLS	Static light scattering
SMAC	Stand-up monolayer adsorption chromatography
TAE	Tris-acetate-EDTA
TAP	Therapeutic antibody profiler
TEV	Tobacco etch virus
ThT	Thioflavin T
T_m	Melting temperature
TPBLA	Tripartite β-lactamase
UV	Ultraviolet
V_H	Variable heavy
V_L	Variable light

List of amino acids

A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartate
E	Glu	Glutamate
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine

Abstract

Over the last 40 years, proteins have emerged as highly effective therapeutics due to their endogenous specificity. While the structural and biophysical properties of protein scaffolds allow the formation of highly avid complexes, the inherent metastability of proteins can result in local or global unfolding that can lead to inactivation and/or protein aggregation. As a result, the production and formulation of biopharmaceuticals can be hindered by protein aggregation which can occur at every stage of the manufacturing process; ultimately jeopardising the successful development of promising candidates from becoming the next blockbuster biologic.

Aggregation compromises the quality, stability, and safety of a drug product, yet the ability to identify 'manufacturable' candidates with long-term stability during lead isolation and optimisation remains challenging. Similarly, the ability to predict the aggregation propensity of proteins associated with protein aggregation diseases is also arduous, and much remains unknown about the fundamental processes driving protein aggregation in these diseases. There is thus an important and currently unmet need to be able to identify protein sequences that may have undesired properties and to engineer their sequences to improve their properties.

Investigating protein aggregation and stability can be laborious, due to the difficulties in expression and purification for *in vitro* analysis and since aggregation can occur through a variety of mechanisms. The work presented in this thesis employs a tripartite β-lactamase platform to characterise the aggregation propensity of biopharmaceuticals that circumvents the need for recombinant expression and downstream analysis. This system can distinguish between aggregation and non-aggregation prone sequences, offering a powerful tool for assessing protein aggregation and stability earlier in the industrial pipeline.

This study also developed a directed evolution methodology that can be used as a novel strategy to modulate the aggregation propensity of protein therapeutics, to evolve 'manufacturable' biopharmaceuticals early during industrial development. Importantly, the approach does not require any structural knowledge or prior biophysical information about the protein of interest.

Finally, the application of this platform to disease-related proteins enabled the identification of hotpot residues that differ between germline and patient sequences in light chain amyloidosis, that may further the understanding of the processes that underpin aggregation diseases.

Overall, this platform provides a new approach for the rapid identification of aggregation resistant proteins and to provide insight into the molecular mechanism of aggregation.

Chapter 1
Introduction

1.1 Principles of protein folding

Proteins govern a multitude of biological pathways and processes that are crucial for life. The biological function of a protein is usually determined by its three dimensional (3D) native structure encoded by a string of amino acids, comprised from just 20 amino acid building blocks. Understanding the mechanism of protein folding lead to the concept of "the protein folding problem"[1], which encompasses three related enigmas: (i) the folding code, (ii) the folding mechanism and (iii) protein structure prediction. This problem has been a challenge in the field for over 60 years, during which theories have been developed to provide an unprecedented understanding of the principles of protein folding[2].

Anfinsen's experiments on ribonuclease A demonstrated that the protein can fold and refold after denaturation, without any biological machinery, to a thermodynamically stable and functional state by searching for the lowest negative free-energy state[3]. This led to the hypothesis that the primary amino acid sequence encodes all the information required for protein folding. However, from this notion arose Levinthal's paradox, where he calculated if a protein were to explore all possible conformations, folding to a global free energy minimum would be impossible on a biologically relevant timescale[4,5]. Levinthal proposed that folding is kinetically determined and must take place through defined mechanisms guided by the rapid formation of local interactions[4].

As research into the protein folding field evolved, the pathway from the unfolded polypeptide chain to the folded native state became represented by an energy landscape model[6]. This pathway characterises the energy-entropy relationship of the folding species. Proteins fold energetically downhill, towards a low-energy, low-entropy state until the native state is reached at the energy minimum[6] (Figure 1.1). Many proteins fold to their native state via folding intermediates, which often contain secondary structure but lack a packed hydrophobic core. These "on-pathway" intermediates give rise to a roughness of the energy landscape by creating low energy kinetic traps[7–9].

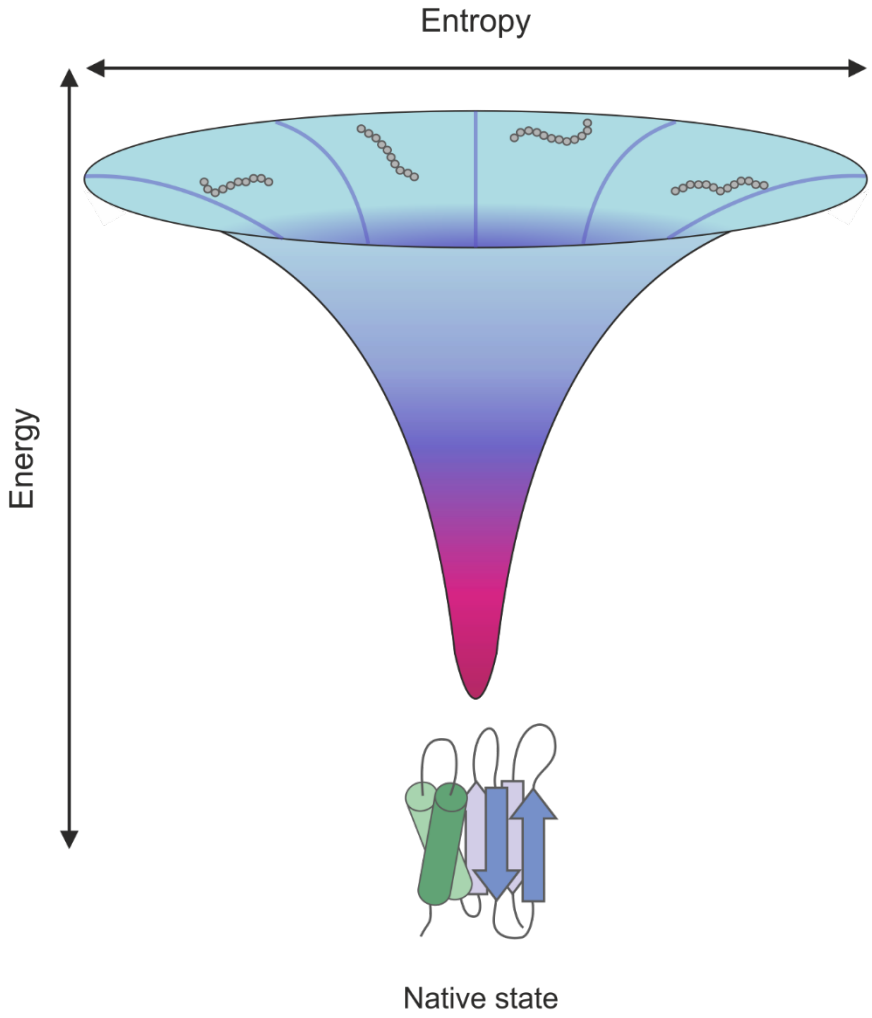


Figure 1.1 Idealised smooth energy landscape of protein folding. Internal free energy of the system is represented by the vertical axis and conformational entropy explored by the polypeptide chain is represented by the horizontal axis. Folding starts at the rim of the funnel, depicted as the grey polypeptide chains. As the number of intermolecular contacts increase, the internal free energy is lowered, and the conformational freedom is reduced until the native state is formed.

1.2 Protein misfolding

Protein folding is further complicated by the formation of "off-pathway" structures that hinder the formation of the native state[7,10]. Various properties can increase the potential of a protein to become trapped in non-native conformations. Destabilising factors include mutations to the amino acid sequence or changes in the cellular environment such as pH, temperature, or absence of ligands[11,12]. The misfolded state may be more stable than the native state causing the protein to become kinetically trapped and significant reorganisation of the protein is required to reach the native state[7,13]. These misfolded states create an additional problem of promoting self-assembly in a process called aggregation that consists of several different pathways and mechanisms (Figure 1.2)[14].

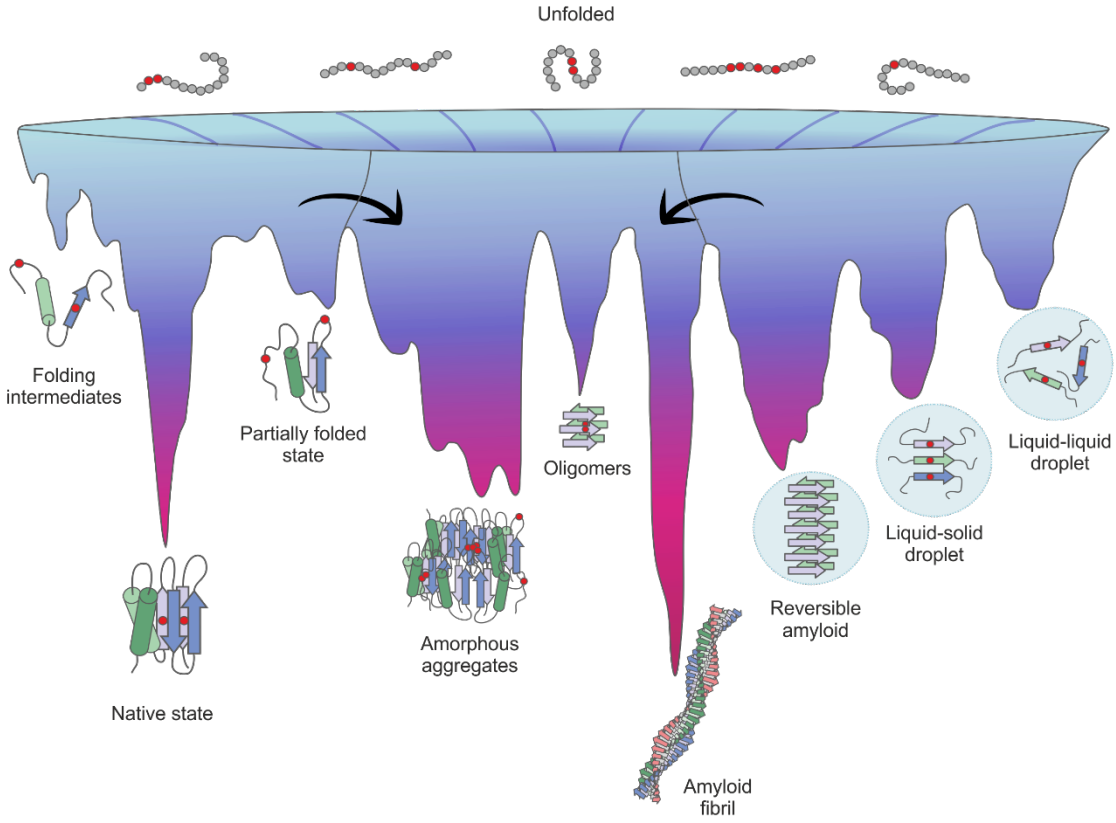


Figure 1.2 Energy landscape of protein folding and aggregation. Proteins sequences have inherent aggregation prone regions (red dots) that are buried in the hydrophobic core of the native state. Funnelling down the energy landscape, metastable conformations are sampled en route to the native state, that can be "on-pathway" or "off-pathway" folding intermediates. Exposure of aggregation prone regions can promote the formation of intermolecular contacts and form amorphous aggregates, oligomers and amyloid fibrils. Exposure of a certain polypeptide segment can trigger proteins to phase separate into membrane-less organelles, which can be classified as liquid-liquid or liquid-solid according to their properties. Higher order species like amyloid fibrils can form as a consequence of phase separation, which can be a dynamic reversible process.

1.2.1 Protein aggregation mechanisms

Aggregation is considered as a series of sequential and parallel events, which can occur from the unfolded, intermediate or native state[15,16]. Protein aggregation can be mediated by aggregation prone regions (APRs) in the protein sequence, which typically comprises hydrophobic residues[17,18]. In the native state APRs are protected from forming protein-protein interactions as they are buried in the hydrophobic core

of a protein, or reside in a protein-protein interface. APRs may exist in one or more places on the protein and can become exposed whilst the protein samples metastable conformations during (un)folding transitions[18] (Figure 1.2). The exposure of these hydrophobic regions can trigger an aggregation cascade producing non-native protein oligomers. These oligomeric species can vary in size from dimers to several monomers that forms a nucleus that serves as the structure for monomer addition (and dissociation), from which the aggregate can grow forming amorphous aggregates and higher order aggregates (Figure 1.3).

An alternative mechanism involves reversible association of the native monomer[15,16]. Here, aggregation occurs directly from the native state, through self-association of monomers to form reversible oligomers (Figure 1.3). This self-association can occur via a variety of mechanisms such as hydrophobic patches, electrostatic colloidal interactions, hydrogen bonding across β-strands or an exposed backbone[19–23]. As concentration of the protein aggregate increases and larger oligomers form, protein aggregation can become irreversible though the formation of covalent bonds such as disulfide bonds[12].

The most thermally stable, low energy protein aggregates known are amyloid fibrils. (Figure 1.2)[11,14,24]. The formation of amyloid fibrils generally occurs via nucleation-dependent oligomerisation[7,25]. In this mechanism oligomers formed have the tendency to rearrange to a more compact aggregation prone form, known as the nucleus (Figure 1.3). The formation of the nucleus is the rate limiting step of fibril formation and typically has increased β-sheet content, or a complete rearrangement of existing β-strands. This conformational rearrangement converts the aggregation process into energetically favourable polymerisation, around which further deposition of monomers occurs[7]. The nucleation model can also involve secondary nucleation events such as surface catalysed events and fibril fragmentation[26] (Figure 1.3). The amyloid state is not only detrimental as it leads to protein loss of function, but is also associated with toxic gain of function in a range of pathological conditions such as Alzheimer's disease, Parkinson's disease and type II diabetes[27,28]. Despite their potential for production of toxic species, amyloids can also be produced as a natural protein fold, whereby the fibrils perform an array of physiological functions[29,30]. Examples of functional amyloids include the storage of peptide hormones[31] and semenogelin proteins for the removal of damaged sperm[32].

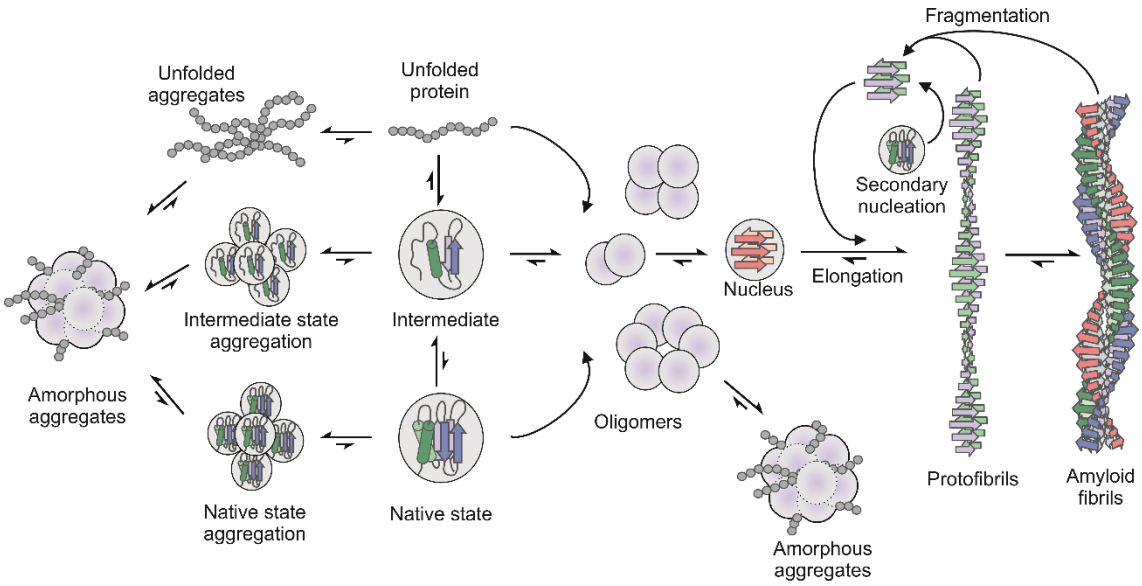


Figure 1.3 Mechanisms of protein aggregation. Aggregation precursors may be the unfolded, partially folded, or native state of a protein. These precursors can assemble to form amorphous aggregates. During amyloid formation, oligomeric species formed from the initial aggregation-prone monomer, can then assemble further to form higher-order oligomers, one or more of which can form a nucleus, which, by rapidly recruiting other monomers, can nucleate assembly into protofibrils and amyloid fibrils. As fibrils grow, they can fragment, yielding more fibril ends that are capable of elongation by the addition of new aggregation-prone species.

1.3 Biopharmaceuticals

Biopharmaceuticals can be defined as the application of biomolecules produced in living systems by biotechnology[33]. This encompasses the use of engineered proteins or nucleic acid based substances for the use as therapeutics or *in vivo* diagnostic purposes. Biopharmaceuticals fall into the category of a biologic which is a broader term defined as any therapeutic agent manufactured in living systems. This includes blood components, vaccines and toxins from natural (non-engineered) sources[33].

Proteins used as therapeutic entities have emerged as a major new class of pharmaceuticals. These modalities have become increasingly popular due to their inherent specificity, and therefore they rarely have adverse effects compared to small molecules[34]. Since proteins naturally regulate many biological processes, they make ideal candidates to treat a broad range of diseases which would be difficult to mimic using synthetic compounds. Additionally, as many diseases result from genetic mutations, proteins can be used as treatment through replacement.

1.3.1 History of biopharmaceuticals

The origins of biotechnology are built upon pivotal advancements in the field of genetic engineering (Figure 1.4). Following the discovery site specific restriction enzymes, Paul Berg developed the "cut-and-splice" method in 1972, combining the SV40 monkey virus with lambda bacteriophage DNA to create the first recombinant DNA molecule[35].

Shortly after in 1973, Boyer and Cohen described the revolutionary method of recombinant DNA technology showing that genetically engineered DNA plasmids are biologically functional when transformed into *E. coli* cells[36]. They created a plasmid with resistance to tetracycline that contained the *EcoRI* restriction site. At this site they added the gene for kanamycin resistance and demonstrated that after transformation into *E. coli,* subsequent generations of *E. coli* had resistance to both tetracycline and kanamycin. They built upon this work by introducing genes from the toad *Xenopus laevis* into the plasmid and demonstrated that these genes were active in multiple generations of *E. coli.*

It became clear that this approach could be used on a large scale to manufacture genes rapidly and in quantity. In 1976 Boyer founded Genentech, where they became the first to successfully express a human gene (somatostatin) in bacteria[37]. The first recombinant protein therapeutic generated using this technology approved by the US Food and Drug Administration (FDA) was human insulin in 1982[38]. Within several years, new techniques for mapping and rapidly sequencing genes were developed and genetic engineering became the basis for an explosion in biotechnology.

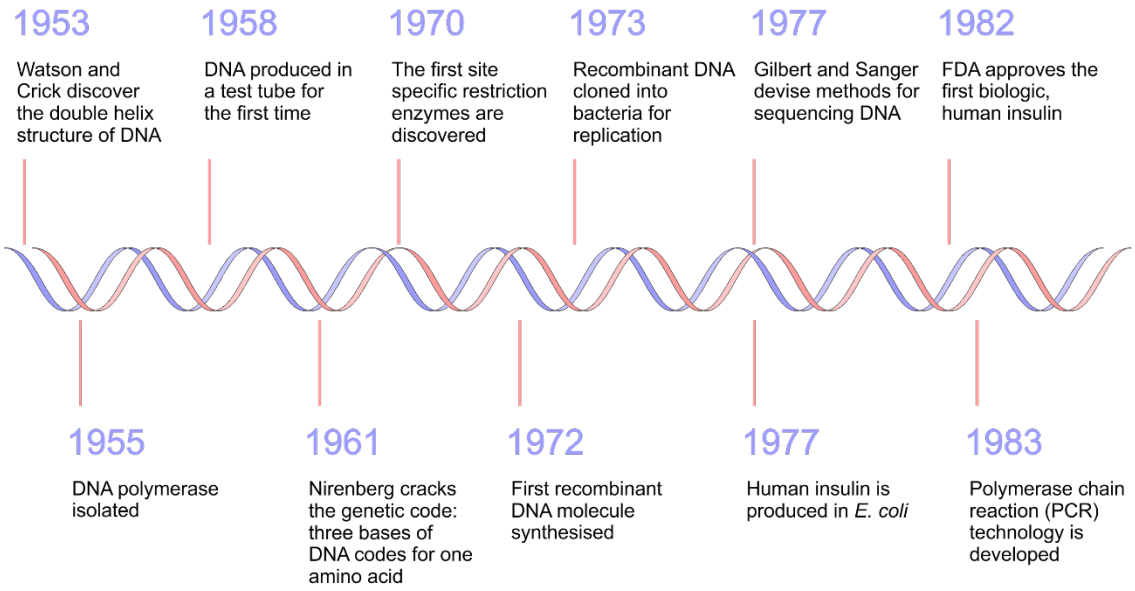


Figure 1.4 Genetic engineering methods that transformed biotechnology.

Innovative methods were also being developed to generate monoclonal antibodies (mAbs) as therapeutics. In 1975 Köhler and Milstein created a fusion of immunised myeloma cells with immortalised myeloma cells to produce a hybridoma cell line that secreted a single type of antibody[39]. Following their discovery they were awarded the Nobel Prize in 1984, and soon after, the first therapeutic antibody for human use, muromonab-CD3 (Orthoclone OKT3), was approved by the FDA in 1985 for the treatment of acute transplant rejection[40]. Today, 45 years after this pioneering work, mAbs are now the highest grossing class of biopharmaceuticals, contributing 65 % of total biopharmaceutical global sales in 2018[41]. This recent success is partly owed to the advancements in the methodologies used for their generation and production.

1.3.2 Antibody therapeutics

Antibodies were first discovered in 1890 by Behring and Kitasato, where they discovered that the transfer of serum from animals immunized against diphtheria to animals suffering from it could cure the infected animals[42]. They identified that the serum contained a specific "anti-toxic activity" (antibodies) that could confer short-lived protection against diphtheria, for which they won the first Nobel prize in physiology or medicine in 1901. Understanding how these molecules were created and how diversity was introduced to recognise almost any foreign substance took scientists many years.

The substances that elicit the body's immune response to foreign substances are called antigens as they elicit antibody generation[43]. Antigens bind to receptors on the surface

of B lymphocytes (B cells) which stimulates the B cell to proliferate and differentiate into plasma cells. Plasma cells secrete millions of antibodies into the blood and lymphatic system where they bind to their specific antigen enabling it to be cleared from circulation through neutralisation, opsonisation or complement activation.

There are five different isotypes of antibodies that differ by the heavy chain constant region: IgM, IgD, IgG, IgE and IgA[44]. The constant region of the heavy chain determines the functional properties of the molecule such as the functional activity (e.g., neutralisation or activation of complement system) and the distribution of the molecule in the body. The IgG isotype is the most common in humans and in the biopharmaceutical sector and can be broken down into four further subclasses: IgG1, IgG2, IgG3 and IgG4[44,45]. These are named in the order of antibody abundance in the serum, with IgG1 being the most abundant. Subclasses mainly differ in the length of the hinge region and number of inter-heavy chain disulfide bonds[45].

IgGs are large (~150 kDa) proteins, which consist of four polypeptide chains connected by disulfide bonds, forming a flexible Y-shaped structure[44]. There are two identical heavy chains (~50 kDa) and two identical light chains (~25 kDa). Each chain folds into structural immunoglobulin (Ig) domain that consist of two β-sheets arranged in an Ig fold; the heavy chain is made up of four Ig domains, whereas the light chain contains two Ig domains. The domains associate to form larger globular domains which assemble to form three equal-sized regions: the crystallisable fragment (Fc) and two antigen binding fragments (Fab). Within the Fab domain is the variable fragment (Fv), which has a diverse range of amino acids in the complementarity determining region (CDR) loops (three per variable domain). The variation within the sequences of the CDRs gives rise to the unique binding specificity of the antibody to its antigen and is responsible for the vast diversity in antigen-recognition by mAbs.

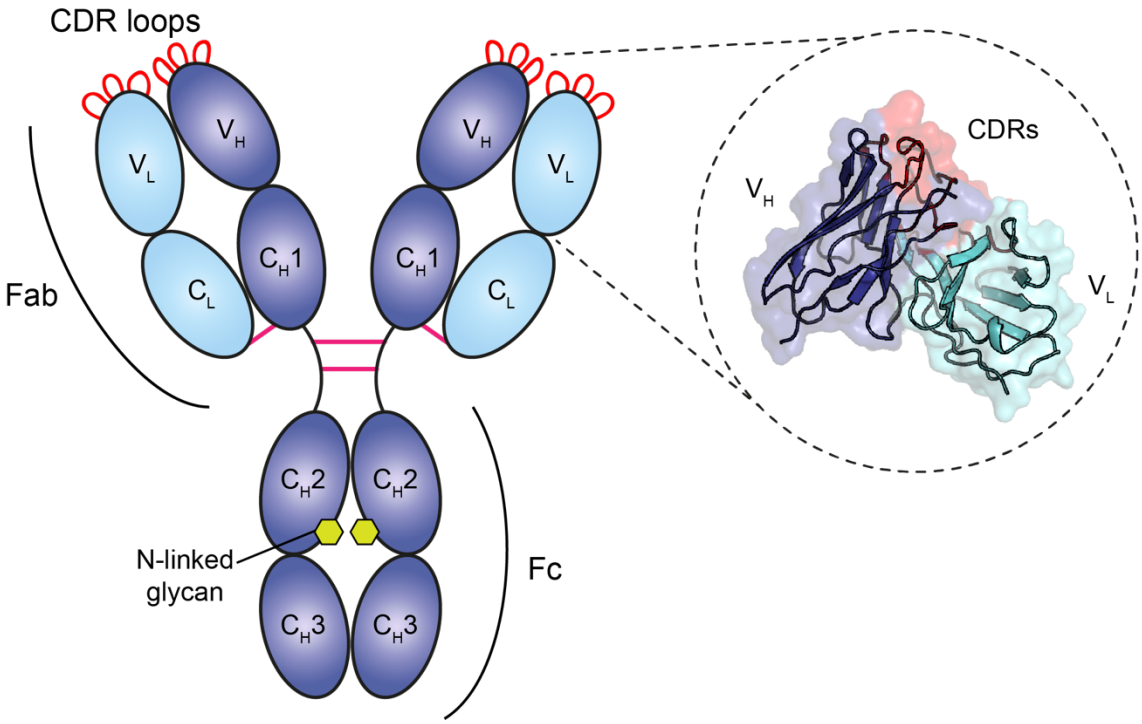


Figure 1.5 Structure of an IgG1 antibody. Schematic representation of an IgG1 domain shows the heavy chains in dark blue and light chains in light blue. The top half of the molecule represents the Fab fragment and the bottom half of represents the Fc fragment. CDR loops are shown in red, disulfide inter-molecular bonds in pink and glycosylation site in green. Figure inset shows the crystal structure of the V_H and V_L domains from a Fab fragment (PDB 5JZ7)[46].

1.3.2.1 Antibody fragments

The modular nature of antibodies has resulted in the generation of different antibody-based scaffolds over the past 30 years[47,48]. Since the binding function of the antibody is determined by the CDRs on the variable domains, fragment molecules have been created to mimic the effect of an IgG in a smaller format, by removing the Fc domain. This was initially carried out by proteolytic digestion of full-length antibodies resulting in two fragments: Fab and Fc[49]. The choice of enzyme for cleavage can generate different Fab fragments. For example, papain cleavage produces a single monovalent Fab fragment (Figure 1.6b), whereas if pepsin is used to digest the IgG bivalent $F(ab)'_2$ is produced (Figure 1.6a), keeping the hinge region between two Fab domains intact.

The engineering of a single chain fragment variable (scFv) was first described in 1988, where the variable heavy (V_H) domain and variable light (V_L) domains are linked together by a flexible glycine-serine liker[50] (Figure 1.6c). This produced a single chain fragment that maintained the affinity for its target antigen, comparable to the full

length mAb[50]. scFvs have become the building block for other antibody fragments such as diabodies and other multivalent formats[51,52].

The single domains of variable regions have also been extensively researched for the use as therapeutics, known as single domain antibodies (dAbs) (Figure 1.6d). This research began with the isolation of mouse V_H dAbs that were shown to have affinity to lysozyme[53]. It became apparent that although the V_H was able to bind to the target, the removal of the V_L exposed a hydrophobic surface, usually protected by the V_H-V_L interface, that caused the dAb to be insoluble and aggregation prone[53]. It was later discovered that camelid organisms[54] and sharks[55] produce antibodies that lack a light chain, triggering further research for the use of dAbs. One noticeable difference between the domains from camelids (VhH/nanobodies) is that they contain long CDR3 loop[56], larger than those observed in conventional murine and human antibodies, and have a complete hydrophilic surface therefore enhancing their solubility and aggregation propensity compared to dAbs[57].

Antibody fragments can offer several advantages over full-length mAbs, such as the ability to be produced in *E. coli* resulting in faster cultivation, higher yields and lower production costs[58,59]. The small size of fragments allows them to penetrate tissues and access cryptic epitopes, making them ideal candidates for tumour penetration in cancer immunotherapy[60–62]. However, the removal of the Fc domain prevents FcR-mediated recycling and as a result the fragments have shorter half-lives[63]. The Fc domain also serves to stabilise the antibody and thus removal results in low thermostability compared to the parent mAb and a greater propensity to aggregate and therefore increases the risk of immunogenicity[51]. Despite these disadvantages, three Fabs, one scFv and one dAb have been FDA approved, with many more candidates in the clinic[51,64].

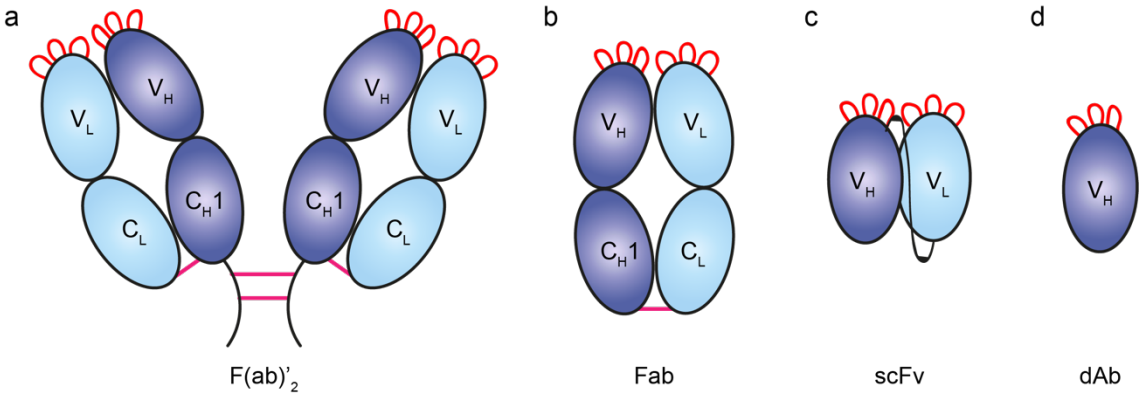


Figure 1.6 Antibody fragments. Cleavage or engineering of the full length IgG can produce a range of antibody fragments such as: a) F(ab)'$_2$ b) Fab, c) scFv and d) dAb.

1.3.3 Generation of antibodies

Different methods can be employed to generate high-affinity antibodies to a range of different antigens. The early stages of biopharmaceutical development involve basic research to identify the aetiology of the disease, such as a signalling pathway or protein to be targeted by the therapeutic to be developed. Once the target has been identified, monoclonal antibodies or antibody fragments that bind to the antigen can be created and lead candidates selected using *in vivo* or *in vitro* methods explained in the following sections.

1.3.3.1 Mouse hybridoma technology

Polyclonal antibodies generated by the natural immune system described in section 1.3.2, are a mixture of molecules with heterogeneity and therefore aren't suitable for the use as biopharmaceuticals. As described earlier, Köhler and Milstein devised a technique to produce a homogenous population of antibodies with known antigenic specificity, known as mouse hybridoma technology (Figure 1.7)[39]. Mouse spleen cells from an immunised mouse are fused to mouse myeloma cells. The spleen cells provide the ability to make the specific antibody whereas the myeloma cell (abnormal plasma cell) provides the ability to grow indefinitely and to continuously secrete the immunoglobulin molecules.

The selection of hybridomas in culture utilises HAT (hypoxanthine-aminopterin-thymidine) supplemented medium[65,66]. Aminopterin in this medium inhibits dihydrofolate reductase (DHFR), preventing endogenous DNA synthesis[67]. When this pathway is blocked, cells utilise the salvage pathway[68] as an alternative method

which requires hypoxanthine and thymidine also supplemented in the medium. The unfused myeloma cells, however, lack the HGPRT (Hypoxanthine-guanine phosphoribosyltransferase) which prevents the production of purine nucleotides required for the purine salvage pathway[69,70]. Therefore, only the hybrid myeloma cell lines (hybridomas) survive in HAT medium as the parental myeloma cells die and the unfused parental spleen cells have a naturally short lifetime[39]. The hybridomas with the desired specificity are identified and cloned by re-growing cultures from single cells (clonal expansion) to create monoclonal antibodies[39].

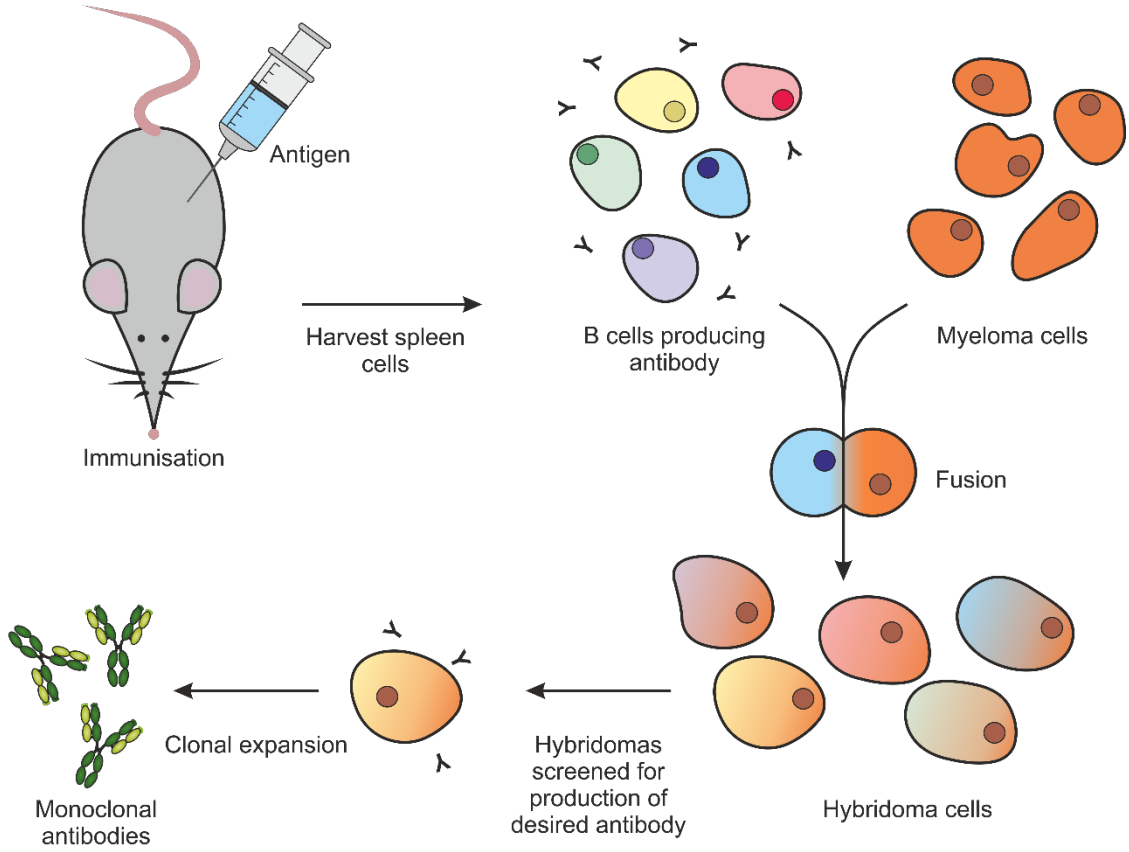


Figure 1.7 Mouse hybridoma technology. A mouse is immunised with an antigen, from which spleen cells are isolated to harvest the B cells producing the antibody raised against the antigen. The B cells are fused with myeloma cells to create hybridomas. Hybridomas are screened to identify the desired antibody for clonal expansion to produce monoclonal antibodies in bulk.

Despite the success of hybridoma technology generating mAbs, the murine linage resulted in immunogenic effects in patients. The consequences of this led to the development of engineering methods to make mAbs more 'human-like', to carry a lower risk of immune reactions. Chimeric antibodies were first engineered where the

murine constant domain was replaced with human constant domain[71,72], this retains the original antibody's antigen specificity and affinity whilst reducing immunogenicity[73]. This idea evolved further by grafting the mouse CDRs onto a human scaffold, generating 'humanised antibodies'[74].

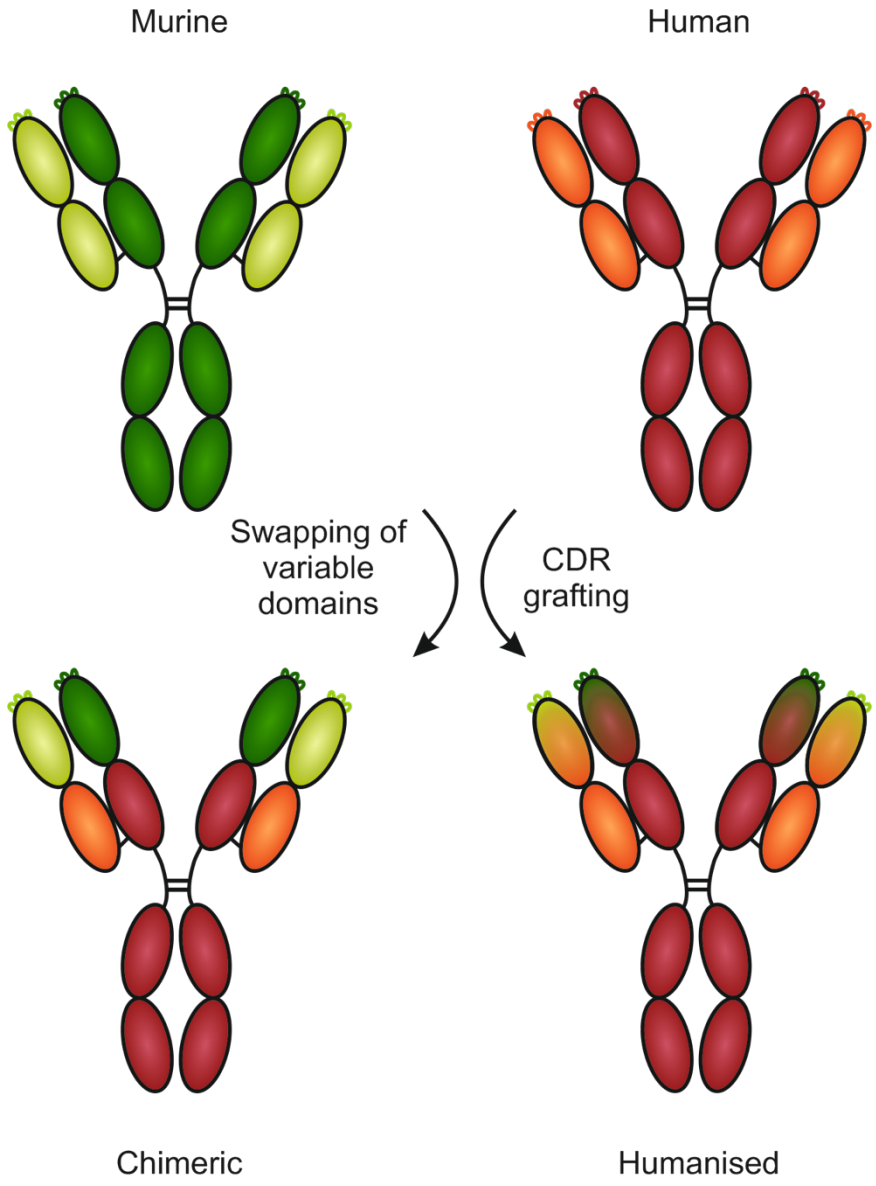


Figure 1.8 Chimeric and humanised antibodies. Murine sequences are illustrated in green and human sequences in red. Lighter colours represent light chains and darker colours represent heavy chains. Figure redrawn from Chames *et al.*[75].

1.3.3.2 Phage display

Following the success of humanised antibodies described in section 1.3.3.1, *in vitro* methods were established to create fully-human mAbs, such as phage display. This technology was first reported in 1985 by George Smith to present a peptide fragment from *EcoRI*, for the enrichment by polyclonal antibodies specific to *EcoRI* endonuclease[76]. The method was further developed and improved for the display of proteins, such as antibodies, for therapeutic purposes[77]. Using a display-based approach during screening circumvents the need of directly panning for binders by exploiting a link between a protein (phenotype) to its cognate gene (genotype) through a phage.

The sequence of interest, such as an scFv or Fab, is cloned into the phage DNA causing a fusion of the antibody fragment and coat protein III gene of a filamentous bacteriophage, such as M13[76]. The resulting antibody::pIII fusion protein is displayed on the surface of phage, to create a 'phagemid'. A diverse phagemid library of 10^6-10^{11} clones can be generated using this method[78]. The library is added to a well in which the antigen is coated to the surface, and phages displaying high affinity antibody fragments bind to the antigen[79]. Sequences that are either non-specific or weak binders to the target antigen are removed in the wash steps before eluting the hits from the panning process. The specific binders are then enriched in *E. coli* with the use of a helper phage to aid production[79]. During two to three rounds of further panning, the conditions are often altered, such as pH, temperature or binding competitors, to enable selection of proteins with increased affinity for the antigen. Following the amplification for increased affinity, the sequences are analysed by enzyme-linked immunosorbent assay (ELISA) for binding specificity[78].

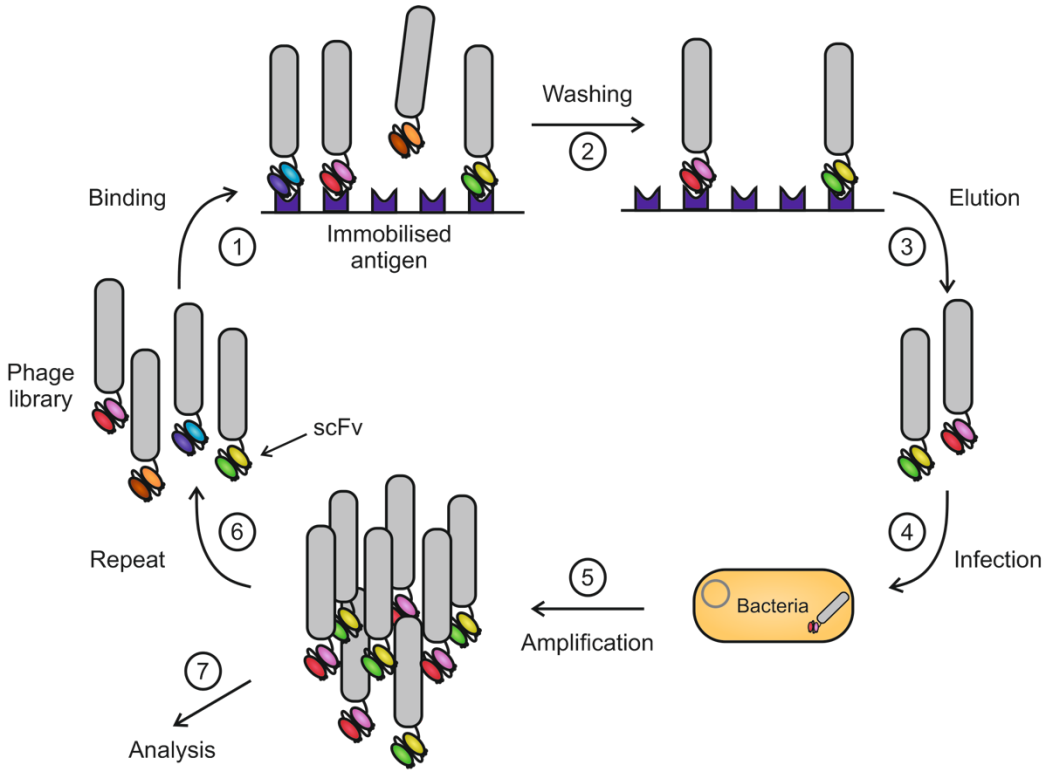


Figure 1.9 Phage display technology. 1) Phage library displaying an scFv is added to the immobilised antigen. 2) The phages displaying the highest affinity scFvs will bind to the antigen, while unspecific binders will be removed by washing. 3) Antigen-specific phages are eluted. 4) *E. coli* are transformed with eluted phages and 5) amplified. 6) A new panning round can further the selection for increased affinity. 7) Sequences are selected for further analysis.

The first approved antibody developed using phage display was Humira (adalimumab)[80,81], which was selected for binding against TNFα. Humira inhibits the acute phase of inflammatory immune responses[82] and is currently marketed for the treatment of nine diseases including rheumatoid arthritis and Crohn's disease. Humira generated the highest sales of pharmaceutical products of 2019 ($19.7 billion)[83], highlighting the success of phage display technology.

1.3.3.3 Ribosome display

An alternative technique to select for high affinity antibodies is ribosome display[84]. This technique utilises a cell-free transcription and translation system and can generate large libraries of 10^{12} - 10^{14} clones[85]. *In vitro*, each sequence in the DNA library is transcribed and translated, however constructs do not contain a stop codon therefore preventing release factors from binding and triggering disassembly of the

translational complex[86]. The antibody fragment, typically a scFv, is now displayed out of the ribosome exit tunnel forming a stable antibody-ribosome-mRNA (ARM) complex[87]. Similar to phage display, this method then utilises a panning procedure to select ARM complexes from the translation mixture via binding of the scFv to the target antigen. Following washing of non-specific sequences, the remaining bound ARM complexes can be dissociated from the antigen and the mRNA is isolated from the complex. Subsequently using reverse transcription PCR (RT-PCR) the mRNA from the initial hits can then be reverse transcribed into cDNA, which is then used for the next cycle of enrichment. Sequencing then reveals which scFvs form the tightest binding molecules. Ribosome display poses challenges to successful screening as the nascent protein still attached to the mRNA and ribosome must be capable of folding into its native state in this constrained format[88]. Furthermore, the intrinsic instability of mRNA reduces the stringency and of selection conditions that can be applied to protein variants for the recovery of the successful sequences[89].

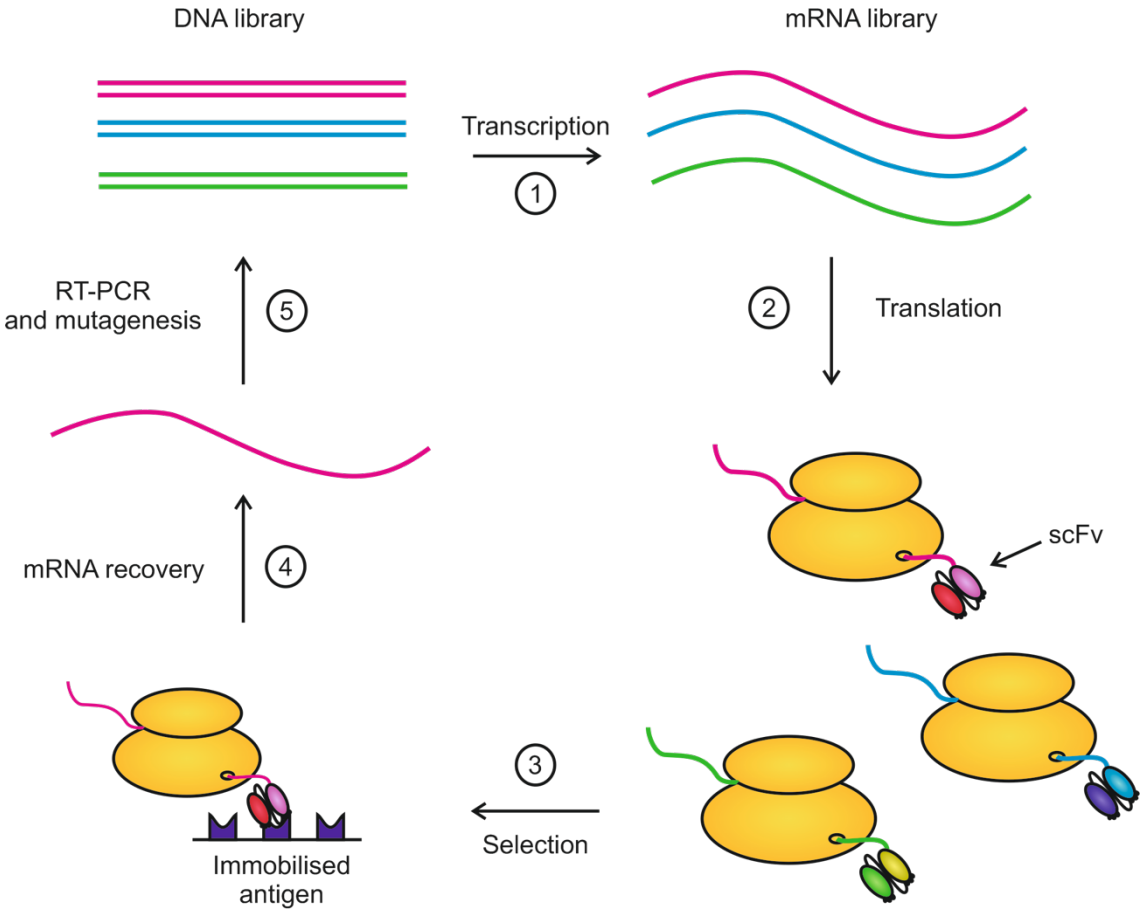


Figure 1.10 Ribosome display. 1) The DNA library is transcribed to mRNA library lacking a stop codon .2) Translation of the mRNA library forms an antibody-ribosome-mRNA (ARM) library. 3) Selection of ARM bound to an antigen. 4) Isolation of mRNA from ARM. 5) RT-PCR and mutagenesis to create DNA library for the next round of selection. Figure adapted from He and Taussig[87].

1.3.3.4 Yeast display

Yeast display utilises the *Saccharomyces cerevisiae* adhesion receptor, a-agglutinin, to display protein fragments on the cell surface[90]. A-agglutinin consists of two subunits, Aga1p which is linked to the cell wall and Aga2p which is covalently bound to Aga1p via a disulfide bond. The antibody fragment is fused to the C-terminus of agap2, along with two epitope tags: a haemagglutinin (HA) and c-myc[90]. After transformation and expression, yeast cells are incubated with a biotinylated antigen, and then methods such as fluorescence activate cell sorting (FACS) can be performed to enrich binders by detecting the antigen with a secondary reagent such as streptavidin conjugated to a flurophore[91].

The expression of the molecules on the surface can be measured through immunofluorescence labelling of either the HA or c-myc tag[92]. On average, the more thermodynamically stable the variant, the larger the number of molecules displayed on the yeast surface[93]. Higher stringency conditions such as higher temperature can also be incorporated to evolve stable antibodies[94]. Another advantage of yeast display is that as the proteins are synthesised in the endoplasmic reticulum of yeast and are therefore subject to the eukaryotic quality control processes and enables the addition of post-translational modifications. A major limitation of yeast display in comparison to phage and ribosome display, is the transformation efficiency of *S. cerevisiae* is limited to 10^7 variants[91].

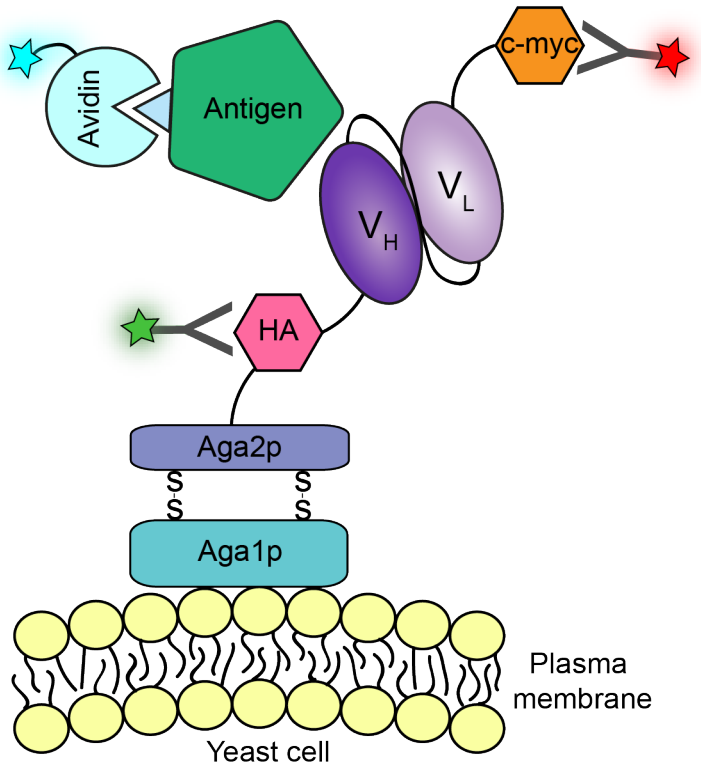


Figure 1.11 Yeast display schematic. The scFv (purple) is displayed as a fusion protein to Aga2p on the surface of yeast. Binding to a biotinylated antigen can be detected with fluorescent avidin. Expression can be detected using fluorescent antibodies binding to hemagglutinin (pink) or c-myc (orange) epitope tags. Figure redrawn from Chao *et al*[p1].

1.3.4 Biopharmaceutical development

The lead candidates that are identified by one of the antibody discovery strategies outlined in section 1.3.3 are taken forward for further development, summarised by the drug development pipeline shown in Figure 1.12. Candidates selected from initial panning experiments can be engineered to further enhance properties such as binding affinity, improved stability, solubility and reduced the aggregation of the protein[95], these approaches are discussed in further detail in section 1.6.2. Other optimisation approaches include formulation screening to identify optimal buffer conditions for the protein before the final candidates are extensively characterised[96]. Lead candidates can then enter pre-clinical trials in animal models to demonstrate the quality, safety and efficacy of the product[33]. Successful therapeutics then enter clinical trials in humans which consist of different phases[97]. Phase I involves safety testing in healthy human volunteers (20-80 people) before moving into phase II where the safety and efficacy are tested on a small number of patients (100-300)[33]. During phase III a large-scale efficacy and safety testing is performed in 1000-3000 patients[33].

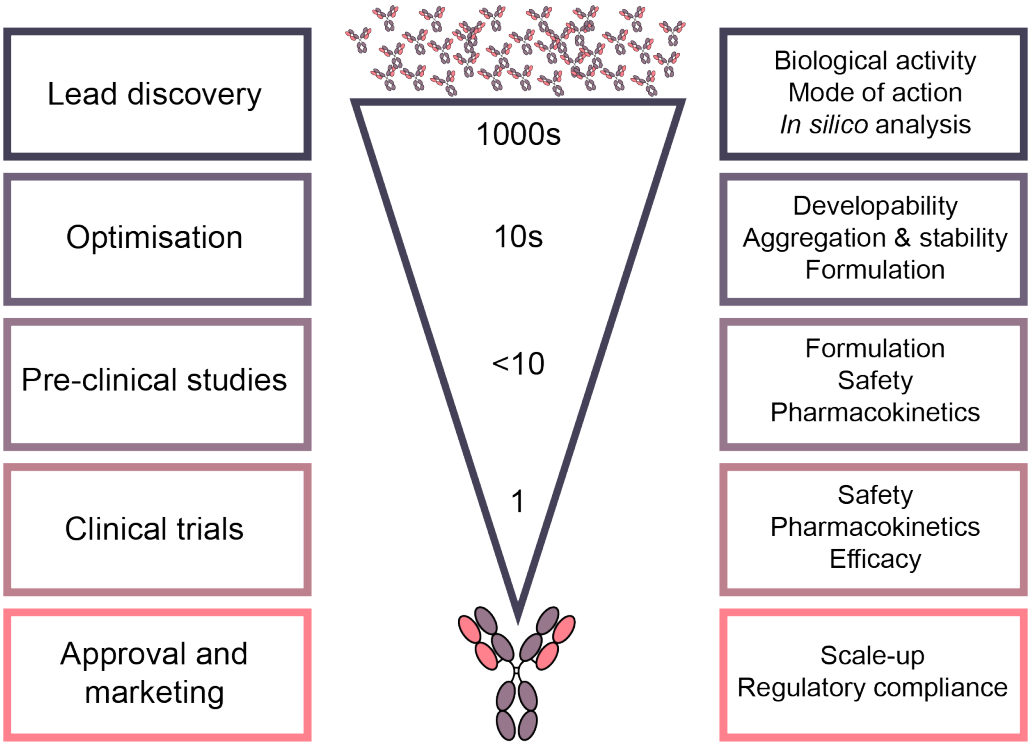


Figure 1.12 Overview of the drug discovery pipeline Screening libraries identifies 1000s of lead candidates from which *in silico* analysis can evaluate diversity and sequence liabilities before expression. Biophysical characterisation is initiated as soon as binding and activity data meets the minimum requirement for the target profile. Optimisation characterises the developability of the therapeutics, which is an iterative process to be during protein engineering. <10 candidates are taken forward into pre-clinical trials to assess pharmacokinetics and safety before clinical trials. Figure adapted from Bailly *et al.* 2020[98].

1.3.4.1 Upstream and downstream processing

The manufacturing of biopharmaceuticals can be divided into upstream and downstream processes (Figure 1.13)[99]. Upstream processing involves expression of the desired product. There are numerous production systems that can be used for the recombinant protein expression of biopharmaceuticals, depending on the characteristics desired[100] as summarised in Table 1.1.

Protein expression system	Cost	Timescale	Expression levels	Post translational modifications (PTMs)
Bacteria	Low	Fast	10-30 g/L	No
Yeast	Low	Fast	Up to 30 g/L	Yes, however differences to human PTMs
Insect cells	High	Medium	Up to 500 mg/L	Yes, however differences to human PTMs
Mammalian cells	High	Slow	5-10 g/L	Yes
Cell free	High	Fast	1-3 mg	Yes, but limited

Table 1.1 Comparison of cell-based expression systems for biopharmaceuticals. Information collated from Tripathi 2019[100], Walsh 2003[33] and Anderson 2002[101].

The most common expression system employed in a biopharmaceutical setting are mammalian cells due to the correct post translational modifications (PTMs) and the correct folding and product assembly which is important for the functional activity of the protein[102]. The benefits of this system however are coupled with slow growth, demanding culture conditions and high production costs due to the defined culture conditions and optimisation requirements.

Typically, Chinese hamster ovary (CHO)[103] or human embryonic kidney 293 (HEK 293)[104] cells are used for mammalian expression and can be generated as stable cell lines that can be used over several experiments or transient production that can generate large quantities of proteins in one or two weeks[105]. The cells are transfected with an expression plasmid for the protein of interest (POI) that can then be used to inoculate growth medium for the production-scale bioreactor[33,105]. At the end of the fermentation process the crude product is harvested from the medium through centrifugation and filtration to remove cells and cell debris prior to downstream manufacturing[99].

Downstream processing ensures that contaminants are removed from to obtain the final product that satisfies purity and quality regulatory requirements (Figure 1.13). A platform process that consists of a well-defined sequence of unit operations is usually employed involving centrifugation, filtration, precipitation and chromatography steps[99,106,107].

Various types of chromatography (affinity, ion exchange and hydrophobic), can be employed to purify the protein[33]. For mAbs, protein A affinity chromatography is the most widely used capture process[108]. This affinity chromatography uses protein A from *Staphylococcus aureus,* that is composed of five homologous Ig binding domains[108]. The domains are independently able to bind to the Fc region hence this chromatography resin is commonly used for its ability to capture IgGs[108]. The specificity enables host cell proteins, DNA and other impurities to be separated from the IgG, providing >98 % purity in a single step[99]. Elution of the IgG occurs in the presence of low pH, which also aids removal of any viruses not cleared in previous steps[99].

Following protein A chromatography additional polishing steps are performed such as ion exchange or hydrophobic interaction chromatography (HIC) to remove any residual impurities and aggregates[109]. The final processing step is ultrafiltration/diafiltration to formulate and concentrate the product[106]. From this, the product can then be stored in vials/prefilled syringes or lyophilised before transportation[110].

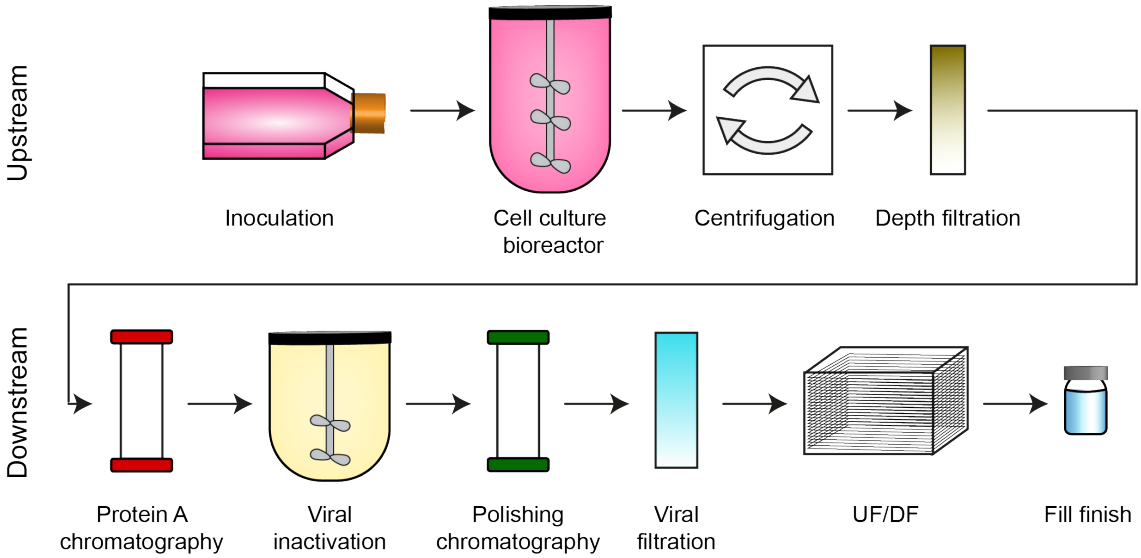


Figure 1.13 Upstream and downstream processing. Cell cultures are inoculated to overexpress the protein from which fermentation is upscaled to a cell culture bioreactor. Cells are separated from the product by centrifugation and depth filtration. The product is then moved to the downstream phase of bioprocessing during which protein A is used to elute the mAb and any remaining viral particles are inactivated. Polishing chromatography steps such as ion exchange or HIC are performed to remove any residual impurities and a further viral filtration step ensures removal of virus particles. Ultrafiltration/diafiltration (UF/DF) concentrates the product and buffer exchanges it into the desired formulation. The product can then be frozen or stored in vials before transportation and administration. Figure based on figures and information from Jozola *et al*[111] and Shukla and Thömmes[99].

1.4 Biopharmaceutical aggregation

Proteins for the use as biopharmaceuticals are subject to the same inherent property to aggregate as naturally occurring proteins as described in section 1.2. Aggregation in biopharmaceutical settings is arguably one of the most challenging factors during research and development. This difficulty stems from the lack of understanding of aggregation mechanisms and the different classes of aggregates that are formed: soluble/insoluble, covalent/non-covalent, reversible/non-reversible or native/denatured[112].

Elimination of aggregation is essential for the safety of a drug. To enter clinical trials a product requires full characterisation to meet quality specifications. If aggregation is present in the end product this too will be administered to the patient which can potentially cause adverse effects such as an immune response. In patients, antidrug antibodies are generated in response to therapeutic protein aggregates that can be categorised into two types: neutralising and non-neutralising[113,114]. Neutralising antidrug antibodies bind to the therapeutic antibody resulting in loss of efficacy of the product. Additionally, they can also inhibit the function of endogenous proteins leading to life-threatening conditions[114]. Non-neutralising bind to the therapeutic protein but do not inhibit the function but may still affect the efficacy of the drug by increasing the rate of clearance from the body[115].

Aggregation can be induced at any stage of biopharmaceutical development (Figure 1.14). The native sequence of the protein may have inherent APRs or hydrophobic/charged patches on the protein surface that can drive aggregation in the early stages of development during protein expression[16,18,46,116]. During upstream and downstream processing, proteins are subject to many different manufacturing stresses that challenge protein stability and can enhance protein misfolding and aggregation[117,118]. These external/environmental factors include freezing, thawing, pH jumps, filtration and agitation[117].

Changes in temperature can alter the conformation of a protein through temperature-induced unfolding that can expose APRs promoting aggregation. The increase in temperature may also increase the rate of aggregation through increasing collision rates[119]. Temperature increases accelerate chemical reactions such as oxidation and deamination which can lead to chemically-induced aggregation[120]. To avoid temperature induced aggregation proteins generally are stored at 2-8 °C and during processing (e.g. fermentation) temperatures are kept below their melting temperature (T_m)[117].

The solution environment such as pH, ionic strength, excipients and contact materials may also induce the formation of aggregates[118]. A change in pH/ionic strength alters the electrostatic interactions through charge distribution on the protein surface influencing aggregation[121]. The contact surfaces such as stainless steel[122] or silicone[123] have all been shown to induce aggregation[124] along with adsorption at air-water interfaces[125,126].

Bioprocessing steps involved in the manufacture of protein therapeutics also involves various dynamic fluid processes, including centrifugation, pumping and filtration. The hydrodynamic forces generated during these processes can bring about unfolding of the protein triggering aggregation[116,127].

Finally, the final product is required to be stored high concentration (~50-150 mg/mL) to reduce the dose volume but since aggregation is usually a concentration dependent process this raises potential issues. The high concentration increases the aggregation potential by increasing macromolecular crowding[128], exceeding the critical concentration for the aggregation nucleus formation[118] and proteins may be near their solubility limit[129]. Moreover, high concentration can increase the viscosity, making injection into the patient slow and painful[63,130].

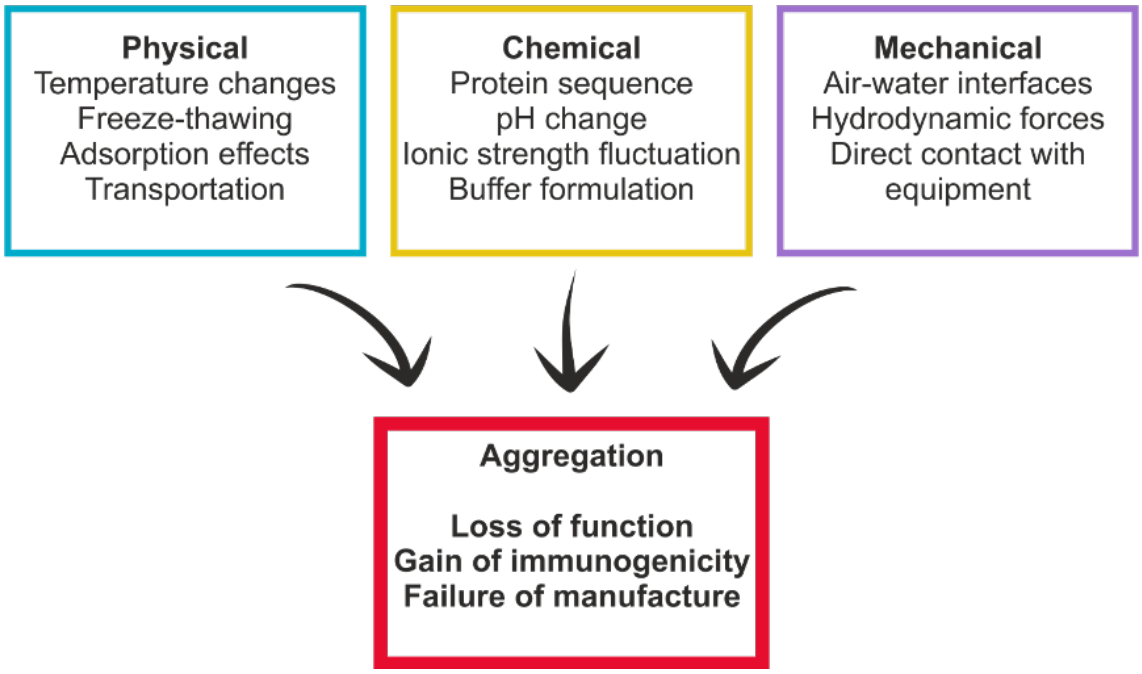


Figure 1.14 Factors that induce aggregation. mAbs encounter a range of stresses during development such as physical, chemical and mechanical that have the potential to induce protein aggregation. Figure adapted from Willis 2018[131].

1.5 Techniques employed to detect aggregation

Since aggregation can jeopardise the developability of a biopharmaceutical, there have been multiple methods developed to predict and characterise aggregation during product development[132]. The physicochemical nature of biopharmaceutical aggregates, in addition to their relative abundances in a formulation, dictates which biophysical methods are suitable for their quantification and characterisation[117], which will be discussed in further detail in this section.

1.5.1 Predicting aggregation *in silico*

There has been an increasing demand to identify any detrimental properties of a protein as early as possible in the development to minimise the risk of late-stage failure and to reduce overall production costs. Computational methods have become valuable tools in both biopharmaceutical and academic settings to identify aggregating, insoluble and/or destabilising regions in polypeptides and proteins. In principle, these tools can be used to identify residues or hotspots on proteins to guide rational design or to screen a large libraries of protein variants. No single algorithm has been shown to be the superior *in silico* tool and each has different advantages depending on the driving force of aggregation.

Many prediction algorithms have been trained on datasets from amyloid/β-rich aggregates, particularly for the detection of APRs[18]. TANGO[17] for example, predicts cross-β aggregating segments by incorporating energetic contributions from hydrogen bonding, side chain–side chain interactions, electrostatics, hydrophobicity, solvation energetics, Van der Waals contacts, and entropy cost[17]. The algorithm was benchmarked against peptides compiled from the literature along with 71 peptide segments from human-disease related proteins[17]. The application of this tool has enabled a deeper understanding of aggregation, identifying that 20 % of all residues in a typical globular domain reside within APRs[133], which are typically flanked by charged amino acids that protect or slow down the aggregation process[133–135]. TANGO however assumes that the polypeptide sequence is fully denatured and solvent exposed, and therefore APRs that are usually buried in the core of the protein are still detected by this software.

FoldX force field was developed to calculate the free energy of a molecule based on the input PDB structure[136]. This server can be used to calculate the free energy of unfolding of a protein, this can therefore be used to calculate the predicted effect that mutations will have on a protein's stability[136]. The force field has also been incorporated into many algorithms to energetically minimise the input structure.

TANGO and FoldX have been combined to create Solubis[137]. This tool can redesign the protein of interest by introducing 'gatekeeper' mutations that disrupt the APRs detected by TANGO, whilst preserving or even improving its intrinsic stability as calculated by FoldX[138]. This has been applied for the screening of mAbs as many APRs reside within the CDRs. Distinguishing between structural APRs (APR exposed upon denaturation) and critical APRs (APR that can trigger aggregation under native conditions) enabled the antibodies to be reengineered to reduce aggregation whilst maintaining function[139].

AGGRESCAN is another prediction software that can identify aggregation hotspots based on an aggregation-propensity scale for natural amino acids derived from *in vivo* experiments that utilise green fluorescent protein (GFP) as a folding reporter[140]. Mutants of Aβ42 were fused to GFP and *in vivo* fluorescent levels were measured, if the substitution resulted in lower fluorescence then that amino acid was proposed to have caused increased aggregation. This original approach assumes that the protein is partially unfolded and therefore to overcome this AGGRESCAN3D (A3D) was created[141]. Again, this utilises the FoldX forcefield to energetically minimise the input structure. A dynamic mode can also be applied where CABS-flex[142] simulations are performed on the energy minimised protein structure, that results in an ensemble of structures with the highest A3D score (most aggregation prone) presented as the output. In A3D 2.0, an automated mutations feature identifies high scoring residues and suggest protein variants with optimised solubility[143].

To screen for solubility *in silico* CamSol can be applied to either predict intrinsic (sequence based) or structurally corrected solubility[144]. The intrinsic method has been used to rapidly screen libraries of sequences, which could be applied during lead candidate selection[145,146]. CamSol can also be used to rationally design protein variants by analysing the structurally corrected profile to identify suitable sites for amino acid substitutions or insertions and then systematically screening thousands of mutations *in silico*. Alongside this, residues can be eliminated from the analysis to maintain protein function. Protein-Sol is another tool that can also be used to predict and evolve proteins with higher solubility[147,148]. Similar to CamSol, this can be used to screen the intrinsic amino acid sequence, or as a structure-based method.

Patterns of hydrophobic residues will form aggregation hotspots when they are clustered on the surface of a protein, and therefore it is important that dynamic 3D information is incorporated into *in silico* predictions. Spatial aggregation propensity (SAP)[149] addresses this by incorporating molecular dynamic (MD) simulations to simulate normal protein fluctuations[150] and has guided engineering studies to improve stability and reduce aggregation of proteins[149,151].

It is clear that lots of generic factors are important in predicating aggregation, and so recently a set of developability guidelines were derived from clinical stage therapeutics[152,153]. Therapeutic antibody profiler (TAP)[152] calculates the total CDR length, patches of hydrophobicity, patches of positive and negative charges and the structural Fv charge (net charge of V_H and V_L) and raises a flag if the antibody screened has nonconforming properties.

1.5.2 Detecting aggregation *in vitro*

A range of analytical methods are employed to characterise biopharmaceuticals. The physiochemical properties that they investigate include molecular weight, conformation, size and shape and extent of aggregation. One of the challenges for studying aggregation is that no single analytical method exists to cover the entire size range in which aggregates appear. To overcome this issue, several routine analytical technologies are implemented to characterise the product, summarised in Table 1.2.

Size exclusion chromatography (SEC) has become an essential tool for the purification and analysis of proteins. SEC separates proteins based on their size (molecular weight and volume) and shape, based on their ability to permeate through a porous matrix e.g., Sepharose. Coupled with high performance liquid chromatography (HPLC) it offers a short analysis time (~15 min) for the rapid separation of macromolecules in a molecular weight range of roughly 5-1000 kDa. An advantage of SEC is that proteins such as antibodies can be separated to detect both higher order species (oligomers and aggregates) and lower order species (unpaired chains or fragments)[154]. Insoluble aggregates however are not characterised using HP-SEC as they are removed during filtration before loading onto the column. Furthermore, soluble aggerates can dissociate reversibly during dilution that occurs during the chromatography process. Another disadvantage of SEC is that non-specific interactions between the sample and the matrix may occur which can increase the elution time of the sample. Other chromatography methods are also employed to study a range of biophysical properties such as hydrophobic interaction chromatography (HIC) to study hydrophobic interactions[155], cross-interaction chromatography (CIC) to identify low specificity[156] and stand-up monolayer adsorption chromatography (SMAC) to investigate colloidal stability[157].

Analytical ultracentrifugation (AUC) can also be used to separate macromolecular species with different densities[158]. During a sedimentation velocity experiment the increasing force applied will separate low- and high- molecular weight species. A time dependent concentration profile can be detected by absorbance or interference detectors to provide a size-distribution analysis of the sample[159]. AUC can be performed at high protein concentrations in the sample formulation buffer and thus

allows a direct measurement of protein aggregation under various solvent conditions to ensure the correct formulation for the product. Although AUC provides a high-resolution analysis of protein aggregates, the technique is low throughput and time consuming[160].

Light scattering techniques, such as static light scattering (SLS) and dynamic light scattering (DLS), are employed to detect and characterise soluble aggregates. The principles of light scattering are based on the properties of particles in solution correlating with the amount of light reflected into the detector. SLS uses Raleigh scattering, in which the electrons of a particle that has been hit by light re-emits radiation at the same frequency in all directions. Larger molecules scatter more light than smaller molecules, and the intensity of the light is proportional to the molecules molecular weight. SLS detectors such as multi angle light scattering (MALS) can be used in combination with SEC to determine the absolute mass and radius of gyration.

DLS measures fluctuations in scattered light that arises from Brownian motion of molecules in the sample, the smaller the molecules the faster the diffusion[161]. Analysis of the fluctuation in the scattered light yields a diffusion coefficient that can be used to calculate the hydrodynamic diameter of the molecule using the Stokes-Einstein equation. Measurements are sensitive to temperature and viscosity and so these conditions must be kept constant for reliable results. DLS is low resolution technique and cannot differentiate between monomer or dimer species[162]. The presence of large aggregates, dust or bubbles can also bias the results due to interference from the particles. Samples must therefore be filtered before analysis, which may alter the particle distribution by removing aggregates.

Other approaches have been implemented to predict if aggregation will occur, by detecting aggregation prone states or structural conformers which are present in transient low concentrations that may nucleate aggregation. A simple way that is commonly used in both industry and academia is to measure the T_m to probe the dynamic nature of protein unfolding. Differential scanning calorimetry (DSC) or differential scanning fluorimetry (DSF) are both implemented in an industrial setting. DSC measures heat capacity as a function of temperature[163]. The protein is exposed to increasing temperature to initiate unfolding during which the heat capacity of the cell increases (T_{onset}). At the temperature at which 50 % of the protein is in its native conformation, and 50 % is denatured, the heat capacity will reach its maximum value (T_m). As each domain of the protein denatures, a peak is formed, and when analysed the T_m for CH2, CH3 and Fab domains can be determined. At the end of the DSC experiment all of the protein will be in its unfolded conformation from which the unfolding enthalpy (ΔH) can also be calculated[163]. One of the limitations of DSC is

that it requires large amounts of material and so when sample is limited DSF is often used. DSF is a fluorescence emission spectroscopy based measurement that uses an extrinsic dye to calculate an apparent T_m[164]. For example, fluorescence of SYPRO orange increases upon protein unfolding due to binding of the fluorophore to newly exposed hydrophobic regions on the protein[165]. The midpoint of the unfolding transition can be determined from the fluorescence emission intensity versus temperature plot. DSF and DSC have been shown to produce comparable results and therefore DSF can be a robust alternative for assessing the stability of many different variants[166].

Biopharmaceuticals have to stay stable in a formulation throughout the products shelf life. The shelf life can be estimated from real-time and accelerated stability studies. In real-time testing, the product is stored at the recommended storage conditions and then monitored for two years, or until it fails product specifications. In accelerated stability studies, the product is stored at elevated stress conditions e.g. 40 °C over two weeks – several months depending on the study[154]. Degradation can then be assessed by monomer loss quantification using HP-SEC and the presence of aggregates can be screened by the methods detailed above. Although accelerated stability studies rapidly enhance the time scale of the study, there is little data that shows a direct correlation between their ability to predict stability at the intended storage conditions (2-8 °C) vs elevated temperature conditions[167].

A high throughput method for early stage antibody discovery has been developed to identify self-interaction of IgGs. Affinity-capture self-interaction nanoparticle spectroscopy (AC-SINS) can be used with low purity samples at a low concentration and therefore is ideal for screening panels of antibodies[168,169]. AC-SINS uses gold nanoparticles to locally cluster antibodies from dilute antibody solutions to achieve high local concentrations (>60 mg/mL) for detecting antibody self-interactions[170]. The gold nanoparticles are coated with an anti-human IgG Fc antibody that captures the test IgGs. Interactions between the immobilised mAbs bring the particles into close proximity[171] which results in a change in colour of the gold colloid solutions[172]. This can be simply quantified via the change in the wavelength of maximum absorbance (plasmon wavelength) using a standard plate reader.

Other methods have also been reported that can be applied at the early stages of lead candidate discovery such as the baculovirus (BV) ELISA[173]. This assay is predictive of non-specific cross-interactions. The BV particles provide a large collection of representative surfaces that an antibody may encounter in human serum. Weak interactions with BV particles detected by an ELISA is indicative of polyspecificity of the antibody which increases the risk of clearance from serum[173].

Method	Parameter measured
SEC	Molecular mass
HIC	Surface hydrophobicity
CIC	Specificity
SMAC	Colloidal stability
AUC	Molecular mass
SLS	Molecular mass and radius of gyration
DLS	Molecular mass and hydrodynamic radius
SEC-MALS	Molecular mass and radius of gyration
DSC	Thermal stability
DSF	Thermal stability
Accelerated stability	Stability and shelf life
AC-SINS	Self-association
BV-ELISA	Non-specific binding

Table 1.2 *In vitro* techniques to characterise protein aggregation

Clearly, there is an arsenal of different biophysical methods that can be employed to study protein aggregation in a biopharmaceutical setting. A landmark study by Jain *et al.* recently selected a panel of twelve biophysical assays to assess the developability criteria of 137 clinical stage antibodies from which the relationship between the many employed 'developability' assays was delineated[153]. Whilst the majority of antibodies performed favourably in the assays, it was apparent that some mAbs failed one or more assays. To understand the relationship between each assay, the data were clustered and analysed to produce a 'family tree' of assays (Figure 1.15)[153]. The clustering identified that the twelve biophysical assays could be divided into five distinct groups that report on: (i) the expression and thermal stability (pink branch), (ii) hydrophobicity (green branch), (iii) loss of monomer at elevated temperature (blue branch), (iv) non-specific interactions (orange branch), and (v) polyspecific or self-interaction (purple branch).

Although this paper has been influential in the field by providing the community a large dataset of antibodies and understanding the relationship between assays, it is not without its flaws. The V_H and V_L domains were all grafted into a common IgG1

scaffold and all assays were performed in HEPES-buffered saline. Evidently the change in scaffold and formulation buffer will have altered the characteristics of the protein to that of the optimised conditions during its original development.

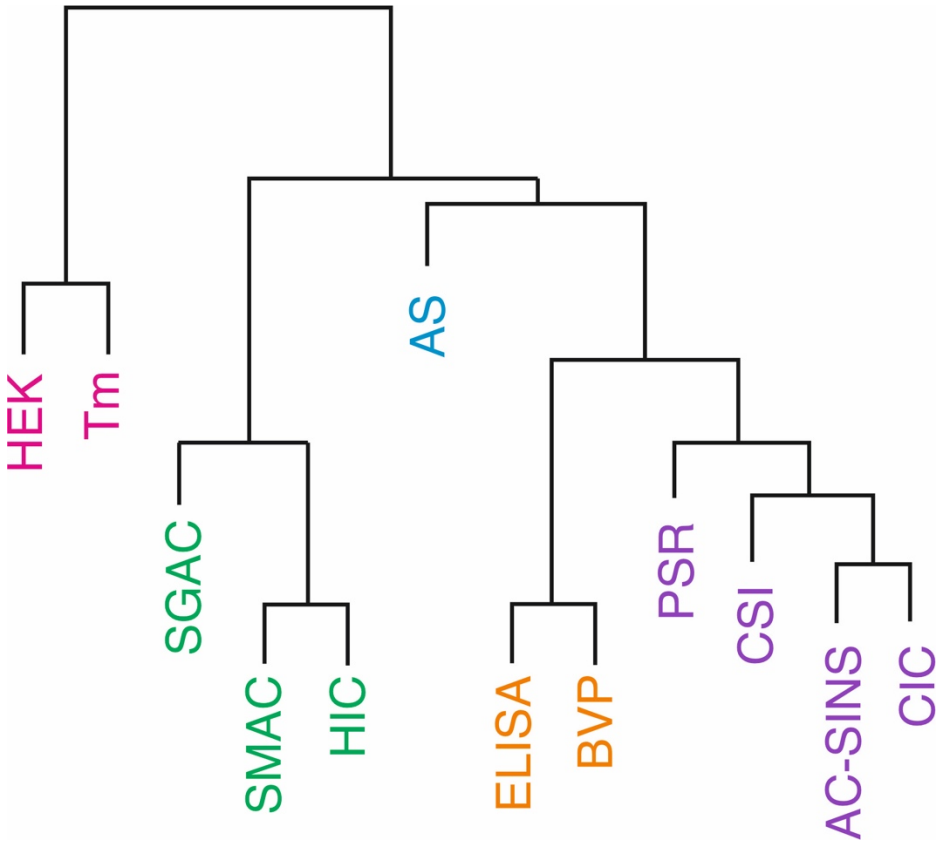


Figure 1.15 Family tree of biophysical assays. Hierarchical clustering of biophysical properties from Jain *et al.* 2017[153]. The assays are grouped together by colour that are statistically related. Assays used: HEK cell titre (HEK), thermodynamic stability (Tm), salt-gradient-affinity-capture self-interaction nanoparticle spectroscopy (SGAC), standup monolayer adsorption chromatography (SMAC), hydrophobic interaction chromatography (HIC), accelerated stability (AS), ELISA panel of commonly used antigens (ELISA), baculovirus particle (BVP), poly-specificity reagent (PSR), clone self-interaction (CSI), affinity capture self-interaction nanoparticle spectroscopy (AC-SINS) and cross-interaction chromatography (CIC).

1.6 Methods employed to reduce aggregation

There are many attributes of an antibody that must be collectively optimised for the generation of a successful therapeutic however, optimising one property can lead to deleterious impacts on others (Figure 1.16)[95]. Importantly the mAb needs to be able to bind with high affinity and specificity to its target which is determined by the CDRs. Altering the sequence to optimise binding specificity can be improved by increased hydrogen bonding, electrostatic or hydrophobic interactions, which may result in increased aggregation or reduced stability. Likewise, engineering approaches that eliminate stretches of hydrophobic amino acids that contribute to aggregation will likely disrupt the folding of the protein. As such, multiple factors must be accounted for when attempting to prevent aggregation. Common approaches taken to enhance conformational (folding) stability, colloidal stability (solubility), and reduced aggregation to improve developability and manufacturing include altering the formulation of the product and an extensive range of protein design and engineering approaches.

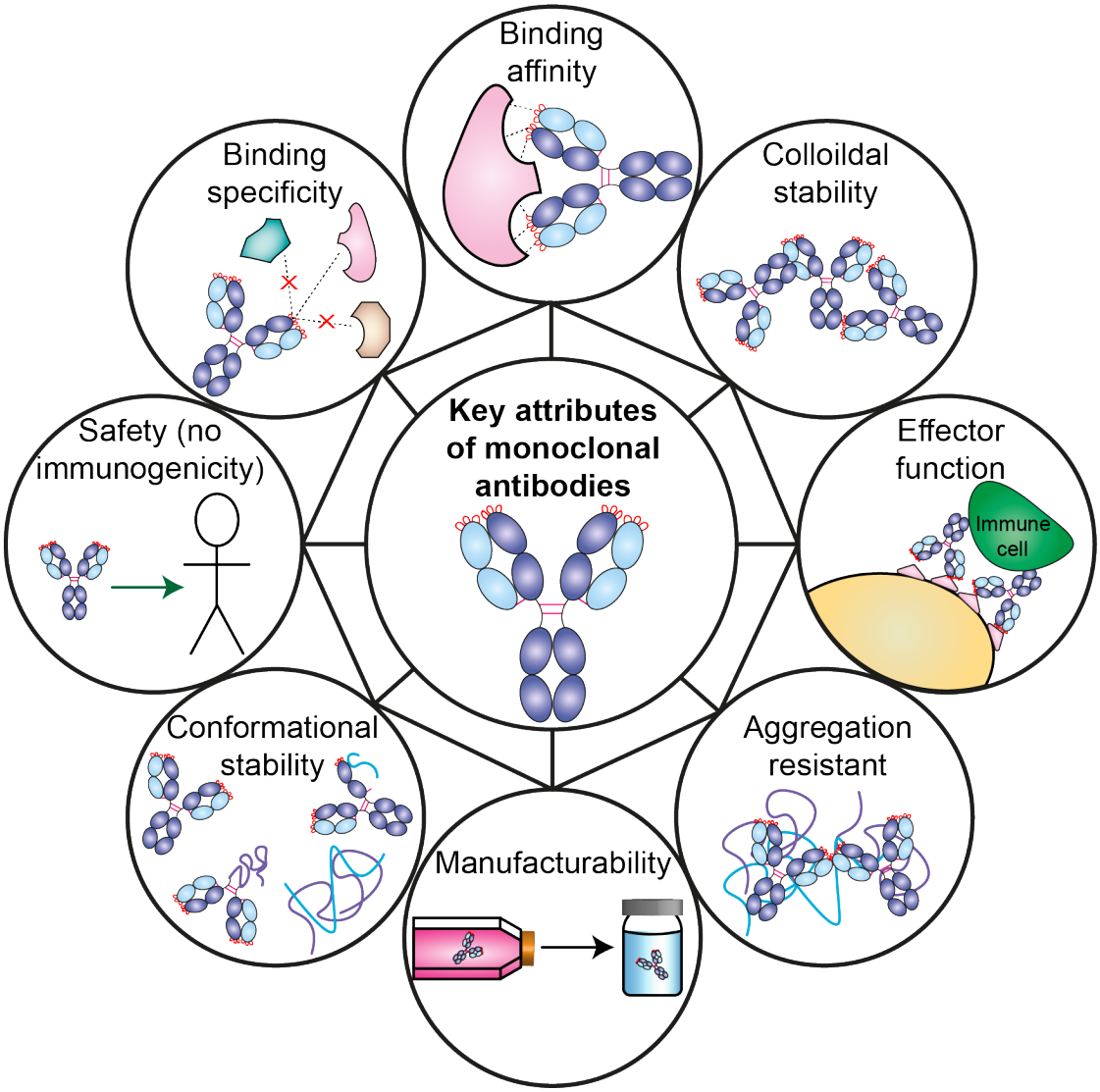


Figure 1.16 Key properties optimised during antibody design. All attributes must be collectively optimised to generate effective IgGs for the clinic, however optimising one property can have deleterious impacts on other properties. Lines represent interdependence, optimisation of any one property can lead to deleterious impacts on others. Figure redrawn and adapted from Rabia *et al.*[174]

1.6.1 Formulation

A common method to address mAb aggregation and increase conformational and colloidal stability is to change the formulation of the product. Excipients such as sugars, polyols and amino acids stabilise the native state conformation[175,176]. Buffering agents are also carefully selected to control the pH and ionic strength of the solution.

Arginine is widely used during protein purification for the refolding of proteins from inclusion bodies through its ability to supress aggregation, yet the mechanism is unclear[177]. This amino acid is commonly used in protein formulations to supress

aggregation through neutralising opposite charges and masking hydrophobic regions[178–180].

Sugars and polyols such as sucrose and glycerol do not physically interact with the protein, rather they are excluded from the protein surface in preference for water[175,181]. This increase in free energy of the system is proportional to the surface area and therefore the unfolded state is unfavoured, pushing the equilibrium of the system to the folded native state[182].

Typically, a therapeutic mAb will be prepared in a range of different formulations following which a structural and stability analysis will be performed (using techniques such as those described in section 1.5.2) to explore the optimal formulation condition. Formulation studies can be low-throughput and product consuming, therefore there is a drive to develop a high-throughput analysis method[183].

1.6.2 Protein engineering

A variety of protein engineering approaches can be taken to redesign a protein to have reduced aggregation. These approaches can range from screening large mutational variants using directed evolution approaches, rational design mutagenesis or isotype switching. Protein engineering has not only been useful for therapeutic aggregation, but also for understanding the fundamental molecular mechanisms of aggregation[184].

1.6.2.1 Isotype switching and reformatting

One engineering approach that can be taken is to alter the antibody scaffold through switching the isotype, subtype or reformatting the antibody (Figure 1.17). For example, switching from an IgG to an IgM can improve the avidity or valency of an antibody. Swapping the subtype of antibodies can be employed to reduce aggregation as although the IgG subclasses are similar in tertiary structure, they differ in the location and number of interchain disulfide bonds. There have been several studies comparing the IgG isotypes in terms of stability and aggregation and the results appear to be differ per study. For example, it has been found that under denaturing conditions the IgG2 format is more prone to aggregation than the IgG1, due to the increased number of free-cysteines in the IgG2 scaffold upon unfolding[185–187]. However, other studies have shown that swapping the framework from an IgG1 to either an IgG2 or IgG4 has enhanced the colloidal stability of the mAb[188].

The reformatting of mAbs to smaller antibody fragments (Figure 1.17) may also prevent aggregation. Although the V_H and V_L domains contribute to biophysical properties of the antibody due to their variability, they can be reformatted into an antibody fragment such as a scFv or Fab as the full length mAb aggregation may be

induced by unfolding of the Fc region[189]. It has also been shown that the mutations introduced into the CDRs to prevent aggregation are dependent on the antibody scaffold[190].

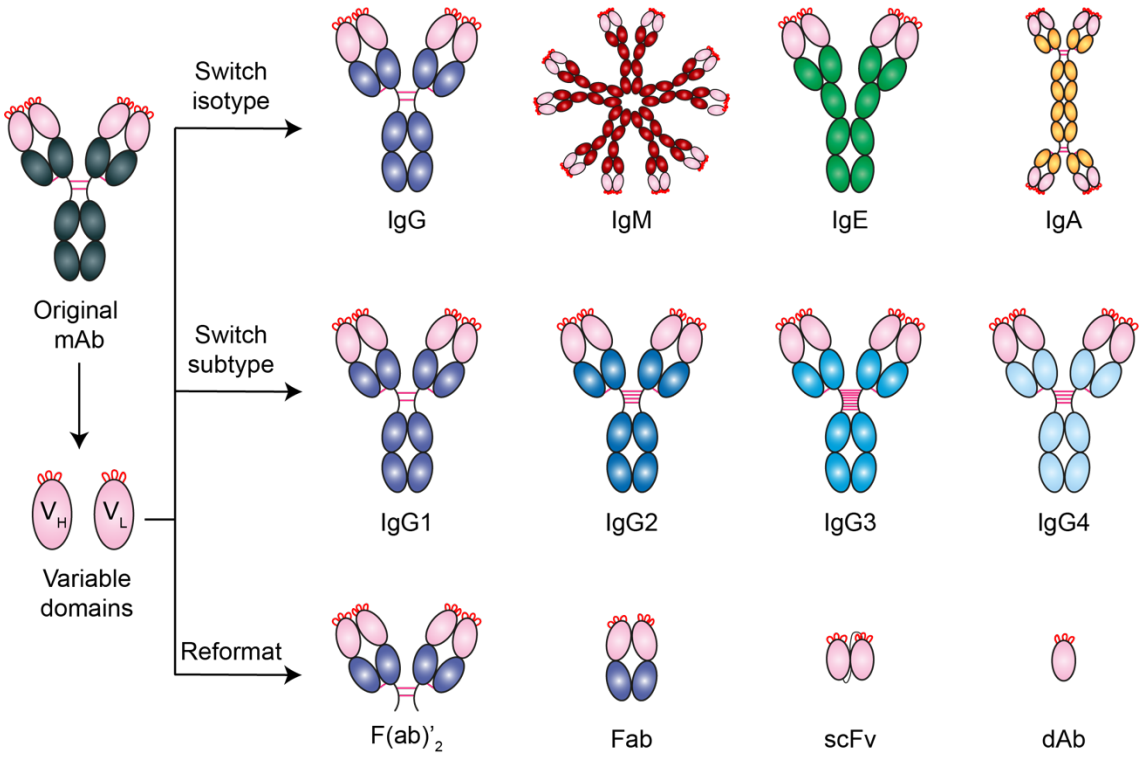


Figure 1.17 Antibody engineering. Engineering the parent antibody (black and pink) into different formats can prevent aggregation. This can involve switching to a different Ig molecule, switching the subtype of the IgG, or reformatting the variable domains into an antibody fragment. Figure redrawn from Absolute Antibody[191].

1.6.2.2 Rational design

Rational design involves the substitution of a small number of residues in a protein sequence to improve the physiochemical or spatial properties. This approach can be taken when there is prior knowledge of the mechanism of aggregation, such as the protein-protein aggregation interface or hotspots identified by *in silico* algorithms (discussed in section 1.5.1).

The use of high-resolution structural information can be useful to generate a hypothesis for rational design. For example, crystallisation of a Fab fragment revealed a symmetrical tetramer that formed as a result of packing contacts[192,193]. A triad of aromatic residues in the V_H-CDR3 were identified as the aggregation hotspot in this structure. Mutation of the aromatic residues to a triple alanine mutant yielded a highly soluble Fab, however unsurprisingly, the Fab lost affinity for the antigen[193]. To overcome this, an N-linked glycosylation moiety was introduced to V_H-CDR2 to

shield the aggregation hotspot. This variant had improved solubility and bound with similar affinity.

Electrostatic interactions can modulate aggregation by changing the probability of protein-protein interactions through electrostatic repulsion. This has been demonstrated by Meisl *et al.* whereby modulating the intermolecular interactions of Aβ42, linked to Alzheimer's disease, resulted in significantly varied aggregation behaviour of the peptide[194]. 'Supercharging' proteins by introducing an excess of acidic or basic residues has also been shown to reduce colloidal aggregation of proteins[195–198]. Similarly, introducing defined clusters of specific charged residues have been shown to control protein stability[190,196,199].

As outlined in section 1.5.1 *in silico* predictors of aggregation are valuable tools to identify regions for protein engineering studies. Since each tool differs in the attribute that it measures such as stability, solubility, or aggregation, using the algorithms in combination can be a powerful approach to identify a key area for rational design. This approach has recently been employed to understand the aggregation the IDP α-synuclein, linked to Parkinson's disease, to identify a seven-residue region (^{36}GVLYVGS^{42}) that controls the aggregation of α-synuclein[200]. Although deleting/substituting the seven residues at position 36-42 prevents aggregation, they found this region was required for the function of α-synuclein, emphasising how proteins have a balance between function and aggregation[200].

Other *in silico* approaches to guide rational design include the use of molecular dynamics (MD) to allow protein motion and flexibility to be simulated to identify a relationship between conformational changes and functional activity of the protein[201]. MD has been useful for identifying highly flexible regions for the introduction of mutations to increase protein stability/reduce aggregation[202–204] and to detect the allosteric and epistatic effects mutations can have on a protein[205,206].

1.6.2.3 Directed evolution

Directed evolution mimics natural evolution by imposing a 'survival of the fittest' selection process. The approach was developed by Frances Arnold and colleagues to evolve the protease subtilisin E to function in a highly non-natural environment[207]. This approach is now employed across multiple industries to enhance the catalytic properties of enzymes[208]. Arnold won the 2018 Nobel prize in chemistry for her work on directed evolution along with George Smith and Gregory Winter for evolution of high affinity binders by phage display (described in section 1.3.3.2).

As a process, directed evolution can be separated into two parts. Firstly, diversity must be introduced into the gene of interest using a mutagenesis strategy[209] followed by a

screen employed to link the genotype to phenotype to identify variants with improved characteristics[210,211].

1.6.2.4 Library generation

A number of genetic diversification techniques can be employed to generate libraries of gene variants that accelerate the exploration of a gene's sequence space. Focused mutagenesis can be to maximize the likelihood that a library contains improved variants, provided that amino acid positions that are likely determinants of the desired function are known. In the absence of known structure–function relationships, random mutagenesis can provide a greater chance of accessing functional library members. Examples of these genetic diversification approaches include error prone PCR (epPCR)[212], DNA shuffling[213], mutator strains of *E. coli*[214], chemical mutagenesis[215] or site saturation mutagenesis[216], which are summarised in Table 1.3. Successful strategies often integrate both random and focused mutagenesis approaches.

Approach	Examples	Random or focused?	*In vivo* or *in vitro*?	Advantages	Disadvantages
Chemical mutagenesis	ethyl methanesulfonate, nitrous acid, ultraviolet irradiation and bisulfite	Random	*In vitro* and *in vivo*	Dose-dependent mutation rates	Low mutation rates; uneven mutational spectrum; hazardous chemicals
Mutator strains	XL1-red *E. coli,* mutagenesis plasmid (PACE) and yeast orthogonal replication	Random	*In vivo*	Easy to use	Low mutation rates; uneven mutational spectrum
epPCR	Taq supplemented with Mg^{2+}, Mn^{2+} and/or unequal dNTPs; proprietary enzyme mixes (Mutazyme)	Random	*In vitro*	Permits high mutation rates; easy to use commercial formulations; relatively even mutational spectrum	Random mutagenesis at the nucleotide level but does not evenly sample amino acid codon space; amplification bias
Site-directed saturation mutagenesis	NNK and NNS codons (where N can be any of the four nucleotides, K can be G or T, and S can be G or C) on mutagenic primers	Focused	*In vitro*	Fully samples amino acid repertoire; focus on functionally relevant residues increases library quality	Requires structural or biochemical knowledge; excess of inactive clones within simultaneous saturation libraries
Homologous recombination	DNA shuffling, family shuffling, heritable recombination and synthetic shuffling42	N/A	*In vitro* or *in vivo*	Can identify beneficial combinations of mutations or eliminate passenger mutations; can also shuffle sequences of orthologous proteins to repurpose functional diversity from nature	Rely heavily on sequence homology; evolved clones and natural orthologues can be divergent in nucleotide sequence

Table 1.3 Summary of library diversification approaches. Comparison of the different approaches employed for library synthesis. Examples for each method are listed along with advantages and disadvantages for each approach. Table from Packer and Liu 2015[209]

1.6.2.5 Genotype-phenotype screens

Following the construction of a library of mutational variants a screen is employed to identify the successful mutants. Several screens have previously been developed for the evolution of reduced aggregation, enhanced solubility, expression or thermal stability.

GFP has been utilised as a folding reporter to assess the folding and solubility of test proteins[210,217]. Here, the POI is expressed as an N-terminal fusion with GFP and expressed in *E. coli*. Cells expressing GFP fused to a folded, soluble protein will produce the GFP chromophore and fluorescence is detected. If, however the POI forms inclusion bodies or aggregates this is directly related to the reduction in fluorescence of GFP[217]. This screening platform has been used in combination with FACS to select for improved solubility for a variety of proteins including Aβ42[218,219], tobacco etch virus (TEV) [220] and ferritin[217]. Enzymes can also be used as fusion tags, such as dihydrofolate reductase (DHFR)[221] or chloramphenicol acetyltransferase (CAT)[222], to identify soluble expression of proteins, that link *E. coli* survival to enzyme activity.

Reporter fusion tags however, can generate high levels of false positives due to truncation or cleavage of the proteins, moreover the addition of a fusion tag may alter the solubility of the protein. To overcome this, protein-fragment complementation assays (PCAs) have been developed[223–225]. For example, GFP is separated into two parts: strands 1-10 and strand 11, which alone do not fluoresce[223,226]. The POI is fused to GFP strand 11 and upon soluble and stable expression of the POI the GFP-S11 is available for complementation by the independently expressed GFP S1-10 fragment, resulting in fluorescence[226].

For many directed evolution approaches, a potential disadvantage of optimising the sequence for solubility[210] or stability[227] is that selection for function is removed resulting in proteins with reduced activity. To counter this, Wang *et al.* describe a soluble expression phage assisted continuous evolution method[228]. Here, the POI is linked in-frame to the N-terminal fragment of a split T7 RNA polymerase (to select for soluble POIs) and the omega subunit of RNA polymerase (RNAP) to select for POIs with high target binding affinity. Linking expression of soluble and functional POI to these distinct polymerases allowed both traits to be selected simultaneously by only allowing expression of the minor coat protein III required for progeny phage upon expression and complementation of N- and C-terminal fragments of an intein transcribed by RNAP and T7 polymerase respectively. Using this approach allowed the isolation of scFvs with five-fold enhancement of expression but unchanged target affinity of for cytidine deaminase.

Other evolution approaches for the selection of aggregation-resistant biopharmaceuticals include enhancing the stringency during phage display selection (1.3.3.2)[229–231]. In this method proteins from the library are selected for binding and then aggregation is induced by heating the phage to >80 °C, following which proteins were selected for their binding to protein A after cooling, which should only bind refolded domains[229]. This approach was used for the evolution of an aggregation prone single domain antibody Dp47d. When isolated as soluble fragments, the evolved variants resisted aggregation upon heating and reversibly refolded, however the T_m of the evolved sequences were either lower or similar to Dp47d[229].

Although the evolution methods developed have been successful in engineering properties of proteins, each method targets a specific mechanism of aggregation, such as aggregation initiated by the unfolded state[229] or solubility[228]. Another flaw of the evolution screening systems described is that they are cytoplasmic-based and therefore cannot usually be applied to proteins that require disulfide bond formation for folding.

1.7 Periplasmic system for identifying aggregation prone proteins

The work in this thesis aims to overcome the flaws of previously developed evolution screens by developing an evolution assay that requires no prior knowledge of the mechanism of aggregation. The assay used in this study utilises β-lactamase as a reporter protein. This requires the activity of β-lactamase to be modulated when aggregation occurs to the POI inserted into a Gly/Ser linker between two domains of β-lactamase.

1.7.1 β-lactamase enzyme

TEM-1 β-lactamase confers resistance to β-lactam antibiotics, such as the penams, cephalosporins and carbapenems in Gram-negative bacteria by catalysing the hydrolysis of the amide bond in the β-lactam ring (Figure 1.18). β-lactam antibiotics interfere with the synthesis of peptidoglycan, an essential component of the bacterial cell wall that contributes to the maintenance of the cell shape and serves as a scaffold for anchoring other cell envelope components[232].

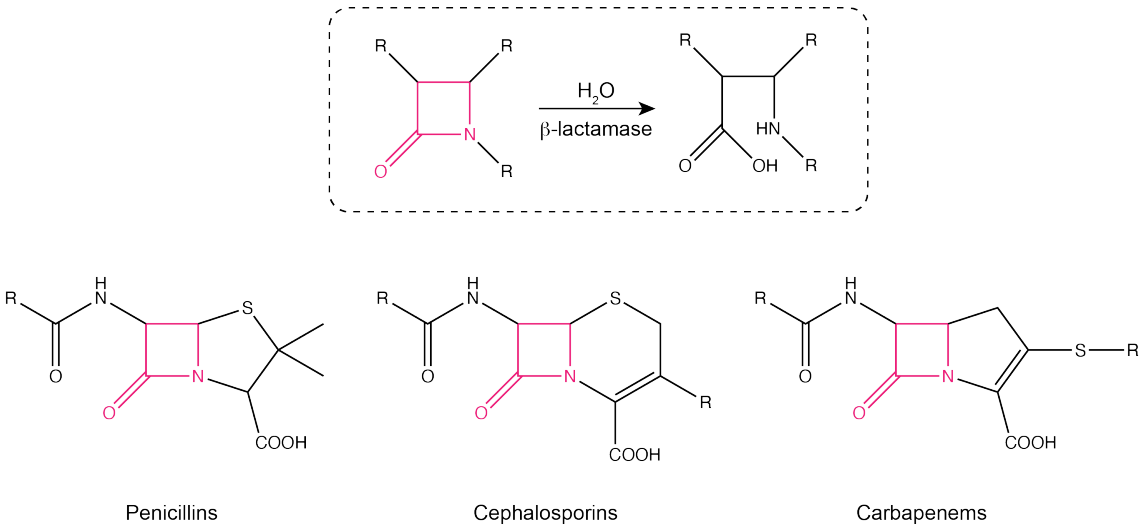


Figure 1.18 Hydrolysis of β-lactam antibiotics. β-lactam antibiotics share a common β-lactam ring, shown in pink. The hydrolysis of the β-lactam ring catalysed by β-lactamase is shown in the top box. Examples of the core β-lactam antibiotics are shown below.

Peptidoglycan is composed of repeating disaccharide units of *N*-acetyl glucosamine (GlcNAc) and *N*-acetylmuramic acid (MurNAc). These strands are cross-linked to each other through peptide side chains of the subunits, via transpeptidation catalysed by penicillin binding proteins (PBPs) (Figure 1.19a). β-lactam antibiotics target this process by binding to PBPs as they have a similar chemical structure to the D-Ala-D-

Ala dipeptide at the terminus of the peptide component of the peptidoglycan (Figure 1.19b and c). This irreversible binding renders the PBP inactive, disrupting cell wall synthesis resulting in cell lysis. The presence of β-lactamase, however, prevents this inhibition of PBPs through the irreversible hydrolysis of β-lactam antibiotics.

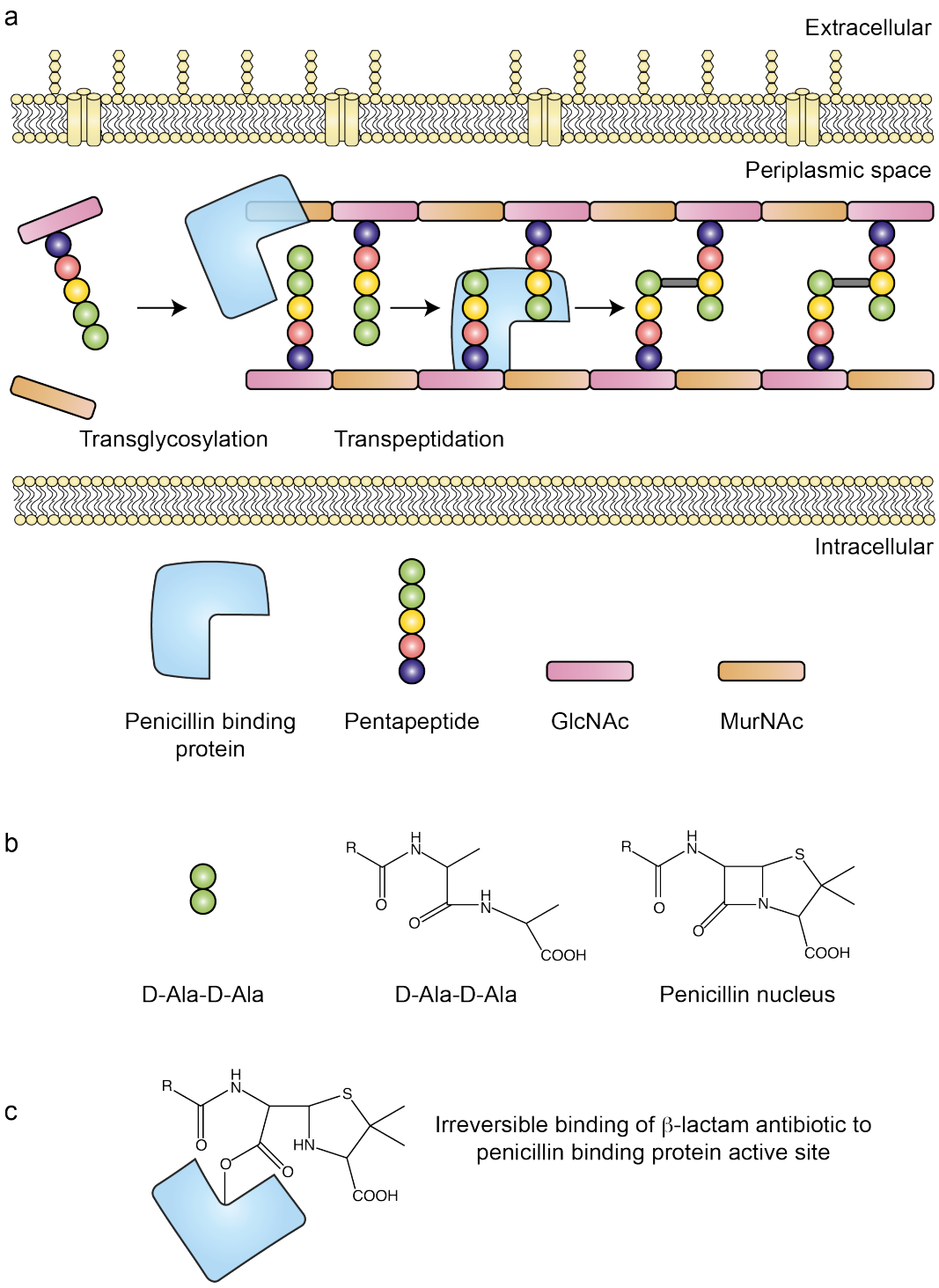


Figure 1.19 Biosynthesis of peptidoglycan and its inhibition by β-lactam antibiotics. a) Peptidoglycan is synthesised by the transglycosylation and transpeptidation of GlcNAc and MurNac by PBP. b) β-lactam antibiotics are structural analogues of the dipeptide terminus D-Ala-D-Ala. c) β-lactam antibiotics irreversibly bind to PBPs, rendering the enzyme inactive. Figure redrawn from Saunders *et al.*[233].

The β-lactamase employed as a reporter in this study is the class A TEM-1 β-lactamase (E.C. 3.5.2.6.)[234]. This 29 kDa protein can be organised into two domains: the αβ domain (comprised of 5 β-sheets and 3 α-helices) and the α domain (comprised of 8 α-helices and several loops)[234] (Figure 1.20). The two domains form the binding cleft on the surface of the protein, in which hydroxyl oxygen of Ser70 serves as the nucleophile for the attack on the carbonyl carbon of the amide bond[235]. The opposite surface to the active site has a region with no secondary structure, providing a suitable site (at residues 196-197) to split the protein in half, that would cause both domains to be inactive on their own, or for the insertion of a POI into which it is still topologically feasible for the protein to fold[236].

Previous studies have investigated the tolerability of domain insertion in TEM-1 β-lactamase[237–239]. For example, pentapeptide scanning randomly inserted 15 bp into the DNA of TEM-1 β-lactamase[237] or domain insertion of cytochrome b_{562}[238,239] has been employed following which β-lactamase activity was assessed to determine the impact of the insertions. For both approaches, tolerated insertions were mapped to two protruding loops which are distant from the catalytic site. Insertions which conferred intermediate levels of ampicillin resistance were found either in different regions of secondary structure which are not directly involved in the substrate binding cavity, in one of two hinge regions of the protein, or in a loop whose C-terminus forms the left border of the catalytic site.

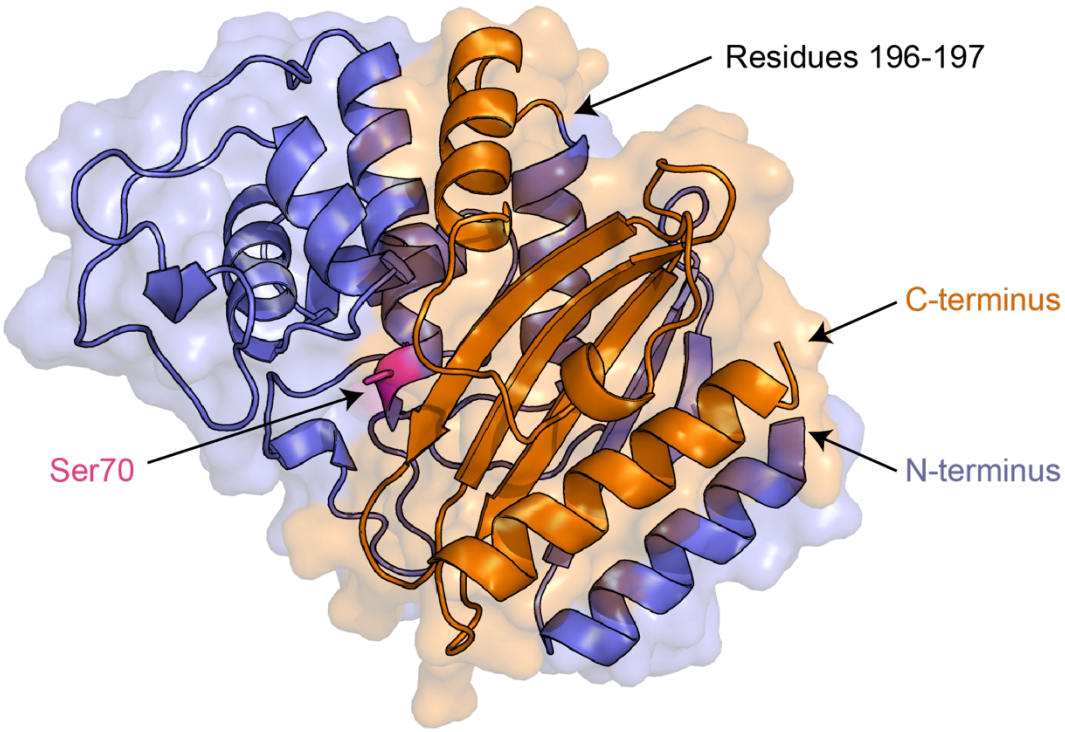


Figure 1.20 Structure of TEM-1 β-lactamase from *E. coli.* The αβ domain is shown in orange and the α domain in blue. Ser70 at the active site is highlighted in pink and the region for POI insertion at residues 196-197 is labelled opposite the active site. Figure created using PDB 1BTL[234] and PyMOL 2.3.2 (Schrödinger)

1.7.2 β-lactamase as a reporter protein

A range of characteristics of β-lactamase make it an ideal candidate as a reporter protein such as being small size (29 kDa), monomeric in nature, well characterised structurally and functionally, easily expressed, and not toxic to prokaryotic and eukaryotic cells. This, combined with the enzymatic read out of antibiotic resistance *in vivo*, or colorimetric assays using chromogenic substrates have been utilised by other groups to develop β-lactamase as an *in vivo* sensor.

β-lactamase was first developed as a PCA to study protein-protein interactions such as the homodimerization of leucine zippers and heterodimerisation of apoptotic proteins Bcl2 and Bad[236]. The two domains of β-lactamase are separated, and each is attached to a POI. If the two POIs interact, then the two domains of β-lactamase are brought into close proximity and can form an active enzyme site. This PCA used the substrate, cephalosporin nitrocefin, that changes from yellow (380 nm) to red (492 nm) once hydrolysed by β-lactamase, that can be detected visually or through change in absorbance. Since this development the β-lactamase PCA has been developed for

filtering genetic opening reading frames[240] and screening for binding from scFv libraries[241].

A tripartite fusion sensor was developed by Foit *et al.* to analyse protein stability[242]. This reporter also incorporates a 23 residue N-terminal signal peptide, to target the protein to the periplasm through the general secretory pathway[243]. Once the signal sequence exits the ribosome, it associates with the cytosolic chaperones SecA and SecB, that maintain β-lactamase in an unfolded state and target it to the Sec YEG translocon. The unfolded polypeptide is then translocated into the periplasm and the signal peptide is cleaved. The mature polypeptide can then fold to the native state in the oxidising periplasm allowing the formation of disulfide bonds.

The tripartite construct took advantage of the domain arrangement of β-lactamase and inserted a POI into the loop on the surface opposite the active site as described in section 1.7.1 (Figure 1.20). Correct folding of the POI allows the two domains of β-lactamase to come into close proximity to form a functional enzyme, capable of hydrolysing β-lactam antibiotics. The *E. coli* expressing a protein that is stable can therefore survive in the presence on β-lactam antibiotics. If, however a mutation is introduced that compromises the proteins stability, the two halves of β-lactamase are unable to associate due to proteolytic cleavage.

Foit *et al.* generated β-lactamase constructs containing four different proteins: the immunity protein 7 (Im7, bla'-Im7-'bla), granulocyte colony-stimulating factor (GCSF, bla'-GCSF-'bla), maltose binding protein (MBP, bla'-MBP-'bla) and cytochrome b_{562} (bla'-cytb_{562}-'bla). Using 62 mutants of these structurally different proteins, the tripartite fusion assay was used to determine a correlation between antibiotic resistance and thermal stability. Interestingly, the evolution of Im7 using this assay identified the evolutionary compromise between function and stability. Any mutations identified to enhance the thermodynamic stability of Im7 mapped to the surface involved in binding to its natural binding partner colicin E7. The method developed by Foit *et al.* has been applied to assess the foldability of designed *de novo* proteins and to evolve mutants with enhanced folding close to the original designs[244,245].

The tripartite β-lactamase sensor has also been developed by Saunders *et al.* to differentiate between aggregation-prone and aggregation-resistant variants such as $A\beta_{1-42}$, amylin, β_2-microglobulin[246]. The system has also been used to screen for small molecule inhibitors of amyloid formation[246] and for the selection of excipients able to supress aggregation[247].

1.8 Aims of the study

The tendency of proteins to aggregate causes significant obstacles during the development of biopharmaceutical development. Despite the array of techniques employed to understand aggregation, these are currently employed late on during the industrial pipeline and require large quantities of purified protein.

Currently, the evolution approaches that are utilised to engineer enhanced biophysical properties focus on a specific mechanism of aggregation. In order to decide on which evolution approach should be performed, prior research must be undertaken to identify the cause of aggregation.

It is evident that there is a critical need to identify and re-engineer aggregation-prone sequences at an early stage of development. Moreover, a method is needed that does not require any prior knowledge of the protein's structure or mechanism of aggregation. Therefore, the aims of this study are to apply an *in vivo* periplasmic screen to identify aggregation-prone biopharmaceuticals. The objectives to achieve this are:

- To assess the applicability of the screen using a range of test proteins selected from the literature with aggregation-prone and aggregation-resistant counterparts.
- To demonstrate the sensitivity of the assay, single- or double-point mutations will be screened *in vivo* and compared to *in vitro* biophysical behaviour.
- To develop a directed evolution approach to engineer proteins with enhanced *in vivo* growth.
- To exploit the data from directed evolution experiments to identify hotspots of protein aggregation in the protein sequence.
- To use an array of biophysical techniques to investigate the enhanced biophysical properties *in vivo*.
- To apply the evolution method developed to biopharmaceutical and disease relevant proteins to identify mutation profiles for different immunoglobin scaffolds.

Chapter 2
Materials and methods

2.1 Materials

2.1.1 Technical equipment

Equipment	Manufacturer
Centrifuges	
Avanti J-26 XP centrifuge	Beckman Coulter, CA, USA
Bench top centrifuge 5418	Eppendorf, NY, USA
Incubators and shakers	
Gallenkamp economy incubator size 1	Sanyo, UK
SI500 orbital incubator	Stuart, UK
SI600 orbital incubator	Stuart, UK
SWB water bath	Stuart, UK
Gel electrophoresis	
PowerPac Basic	Bio-Rad, CA, USA
Slab Gel Electrophoresis Chamber AE-6200	ATTO, Japan
Varigel casting tray	Fisher scientific, UK
Varigel Modular Horizontal Gel Tank Unit	Fisher scientific, UK
Spectrophotometers	
NanoDrop 2000 UV-Vis spectrophotometer	Thermo Scientific, MA, USA
Ultrospec2100 UV/Visible spectrophotometer	GE healthcare, UK
Microplate readers and plates	
384-well polystyrene UV transparent plate	Thermo Scientific, MA, USA
48-well suspension plate with lid	Greiner Bio-one, UK
96-well flat bottom assay plate	Corning, Germany
96-well sterile transparent plate	Greiner Bio-one, UK

96-well white well PCR plate	Bio-Rad, CA, USA
Adhesive sealing film	Sigma Life Sciences, MO, USA
CLARIOstar	BMG Labtech, Germany
EnVision	PerkinElmer, MA, USA
FLUOstar optima	BMG Labtech, Germany
Lunatic	Unchained Labs, CA, USA
SPECTROstarNano	BMG Labtech, Germany
Thermocyclers	
CFX96 Real-Time PCR system	Bio-Rad, CA, USA
T100 thermal cycler	Bio-Rad, CA, USA
Chromatography	
Agilent 1,100 series HPLC	Agilent, CA, USA
ÄKTA pure	GE healthcare, UK
HiLoad SuperdexTM 75 26/60 gel filtration column	GE healthcare, UK
HiTrap Q HP 5 mL anion exchange column	GE healthcare, UK
TSK-GEL G3000SW$_{XL}$ HPLC column	Tosoh, Japan
Other	
1 mm pathlength cuvette	Hellma Analytics, Germany
ChirascanTM plus CD Spectrometer	Applied Photophysics, U.K.
MACS MultiStand magnetic rack	Miltenyi Biotec, Germany
MicroPulser electroporator	Bio-Rad, CA, USA
Orion versa star pro pH meter	Thermo Scientific, MA, USA
PD-10 columns	GE healthcare, UK
Pipetman M 8 channel 50-1200 µL	Gilson, UK
Pipetman M 812 channel 1-20 µL	Gilson, UK
Q9 Alliance	Uvitech, UK
Qubit 4 Fluorometer	Thermo Scientific, MA, USA

Series 2100 media autoclave	Prestige Medical, UK
Siliconized 1.5ml tubes	VWR, PA, USA
SnakeSkin Pleated Dialysis Tubing; 3,500 MWCO	Thermo Scientific, MA, USA
Tecnai T12 electron microscope	FEI company, OR, USA

2.1.2 Reagents

Reagent	Manufacturer
A	
Acetic acid, glacial	Fisher Scientific, UK
Acrylamide 30 % (*v*/*v*)	Severn Biotech, UK
AffiniPure goat anti-human IgG Fcγ Fragment specific	Jackson ImmunoResearch, PA, USA
Agar	Melford Laboratories, UK
Agarose	Melford Laboratories, UK
Ammonium chloride, NH_4Cl	Fisher Scientific, UK
Ammonium persulfate	Sigma Life Sciences, MO, USA
Ampicillin	Formedium, UK
Anti-β-lactamase (CSB-PA352353YA01ENL)	Cusabio, TX, USA
Anti-rabbit goat IgG horseradish peroxidase conjugate	New England Biolabs, UK
Arabinose	Sigma Life Sciences, MO, USA
B	
Bovine serum albumin, BSA	Sigma Life Sciences, MO, USA
Bromophenol blue	Sigma Life Sciences, MO, USA
C	
Calcium chloride, $CaCl_2$	Melford Laboratories, UK
ChromePure Goat IgG, whole molecule	Jackson ImmunoResearch, PA, USA
Citrate-stabilized 20nm gold nanoparticles	Expedeon, UK
D	
Deoxynucleoside triphosphates, dNTPs	Promega, UK
Dimethyl sulfoxide, DMSO	Sigma Life Sciences, MO, USA
Dithiothreitol, DTT	Melford Laboratories, UK
DNA ladder, 1 kb	New England Biolabs, UK

DNA ladder, 100 bp	New England Biolabs, UK
Dylight650	Thermo Scientific, MA, USA
E	
Ethanol	Fisher Scientific, UK
Ethidium bromide, EtBr	Sigma Life Sciences, MO, USA
Ethylenediamine tetra acetic acid, EDTA	Sigma Life Sciences, MO, USA
G	
Gel loading dye, purple (6×)	New England Biolabs, UK
Glucose	Fisher Scientific, UK
Glycerol	Fisher Scientific, UK
H	
Hydrochloric acid, HCl	Fisher Scientific, UK
I	
Instant blue stain	Expedeon, UK
Isopropanol	Fisher Scientific, UK
K	
Kanamycin	Formedium, UK
L	
α-Lactose	Fisher Scientific, UK
LB broth, granulated	Fisher Scientific, UK
M	
Magnesium sulfate, $MgSO_4$	Fisher Scientific, UK
Manganese chloride, $MnCl_2$	Sigma Life Sciences, MO, USA
MOPS	Sigma Life Sciences, MO, USA
N	
Nerve growth factor, NGF	R&D systems, MN, USA
Nitrocefin	Calbiochem, CA, USA
P	
Phosphate buffered saline, PBS	Sigma Life Sciences, MO, USA

Poly(ethyleneglycol) 10,000	Sigma Life Sciences, MO, USA
Poly(ethyleneglycol) methyl ether thiol (2000 MW)	Sigma Life Sciences, MO, USA
Potassium acetate	Sigma Life Sciences, MO, USA
Potassium phosphate, KH_2PO_4	Sigma Life Sciences, MO, USA
Potassium fluoride, KF	Sigma Life Sciences, MO, USA
Potassium hydroxide, KOH	Fisher Scientific, UK
Precision plus dual xtra protein marker	Bio-Rad, CA, USA
Q	
Qubit dsDNA BR Assay Kit	Thermo Scientific, MA, USA
R	
Rubidium chloride, RbCl	Sigma Life Sciences, MO, USA
S	
Skim milk powder	Serva, Electrophoresis, Germany
SOC outgrowth medium	New England Biolabs, UK
Sodium azide, NaN_3	Sigma Life Sciences, MO, USA
Sodium chloride, NaCl	Fisher Scientific, UK
Sodium phosphate monobasic, Na_2HPO_4	Sigma Life Sciences, MO, USA
Sodium phosphate dibasic, NaH_2PO_4	Sigma Life Sciences, MO, USA
Sodium sulfate, Na_2SO_4	Sigma Life Sciences, MO, USA
Sodium dodecyl sulphate	Sigma Life Sciences, MO, USA
Sodium hydroxide, NaOH	Fisher Scientific, UK
Streptavidin Europium cryptate	CisBio, France
Sucrose	Sigma Life Sciences, MO, USA
SuperSignal western pico chemiluminescent substrate	Thermo Scientific, MA, USA
SYBR safe	Invitrogen, UK
SYPRO Orange protein gel stain	Invitrogen, UK
T	

Tetramethylethylenediamine (TEMED)	Sigma Life Sciences, MO, USA
Tetracycline	Sigma Life Sciences, MO, USA
Thioflavin T	Sigma Life Sciences, MO, USA
Tris-Tricine-SDS (10×)	Thermo Scientific, MA, USA
Tris	Fisher Scientific, UK
Tryptone	Fisher Scientific, UK
Tween 20	Sigma Life Sciences, MO, USA
U	
Uranyl acetate	Sigma Life Sciences, MO, USA
Y	
Yeast extract	Fisher Scientific, UK

2.1.3 Molecular biology enzymes and kits

Enzyme	
Antarctic phosphatase	New England Biolabs, UK
*Bam*HI-HF restriction endonuclease	New England Biolabs, UK
T4 DNA ligase	Promega, UK
Vent DNA polymerase	New England Biolabs, UK
*Xho*I restriction endonuclease	New England Biolabs, UK
Molecular biology kits	
Diversify PCR random mutagenesis kit	Takara Bio, Japan
GeneMorph II random mutagenesis kit	Agilent Technologies, CA, USA
Golden gate assembly kit	New England Biolabs, UK
NEBNext multiplex oligos for Illumina (Dual Index)	New England Biolabs, UK
NEBNext Ultra II DNA library prep kit for Illumina	New England Biolabs, UK
Q5 site directed mutagenesis kit	New England Biolabs, UK
QuikChange lightning multi site-directed mutagenesis kit	Agilent Technologies, CA, USA
DNA purification kits	
PureYield plasmid midiprep system	Promega, UK
QIA gel extraction kit	Qiagen, UK
QIA PCR purification kit	Qiagen, UK
QIAprep spin miniprep kit	Qiagen, UK

2.1.4 Buffers

2× SDS loading buffer

50 mM Tris.HCl, pH 6.8, 100 mM DTT, 2 % (*w*/*v*) SDS, 0.1 % (*w*/*v*) bromophenol blue, 10 % (*v*/*v*) glycerol

Electrophoresis anode buffer

400 mM Tris.HCl, pH 8.8

Electrophoresis cathode buffer

200 mM Tris.HCl, pH 8.25, 200 mM tricine, 0.2 % (*w*/*v*) SDS

Tris-acetate-EDTA (TAE)

40 mM Tris-HCl, pH 8, 20 mM acetic acid (glacial), 1 mM EDTA, pH 7.5

Tris-EDTA (TE)

10 mM Tris-HCl, pH 8, 1 mM EDTA, pH 8

Transformation buffer 1 (TFB1)

30 mM potassium acetate, 10 mM $CaCl_2$, 50 mM $MnCl_2$, 100 mM RbCl, 15 % (*w*/*v*) glycerol. Adjust pH to 5.8 with acetic acid

Transformation buffer 2 (TFB2)

10 mM MOPS, 75 mM $CaCl_2$, 10 mM RbCl, 15 % (*w*/*v*) glycerol. Adjust pH to 6.5 with KOH

2.1.5 Media

Autoinduction

464 mL 2ZY, 1 mL $MgSO_4$, 10 mL 50× Lac, 25 mL 20× NPSC

2ZY

1 % (*w*/*v*) Yeast extract, 2 % (*w*/*v*) Tryptone

50× Lac

25 g (*v*/*v*) Glycerol, 2.5 g (*w*/*v*) glucose, 10 g α-lactose, 100 mL H_2O

20× NPSC

26.75 g (*w*/*v*) NH_4Cl, 16.1 g (*w*/*v*) $NaSO_4$, 34 g (*w*/*v*) KH_2PO_4, 35.5 g (*w*/*v*) Na_2HPO_4, 500 mL H_2O, pH 6.75

2.1.6 Bacterial strains

Strain	Supplier	Genotype
E. coli BL21 (DE3)	New England Biolabs (UK)	*fhu*A2 *[lon] omp*T *gal (λ DE3) [dcm]* Δ*hsdS*
E. coli DH5α	New England Biolabs (UK)	*fhu*A2 Δ(*arg*F-*lac*Z)U169 *pho*A *gln*V44 ϕ80Δ(*lac*Z)M15 gyrA96 *rec*A1 *rel*A1 *end*A1 *thi*-1 *hsd*R17
E. coli SCS1	Agilent (USA)	*rec*A1 *end*A1 *gyr*A96 *thi*-1 *hsd*R17 (rκ- mκ+) *sup*E44 *rel*A1
E. coli TG1	Lucigen (UK)	F' [*tra*D36 *pro*AB *lacIqZ* ΔM15] *sup*E *thi*-1 Δ(*lac-pro*AB) Δ(*mcr*B-*hsd*SM)5(*rK - m*K -)

Table 2.1 Bacterial strains used in this study

2.1.7 Origin of Plasmids

Details of plasmids obtained for this thesis are outlined in Table 2.2.

The plasmid containing β-lactamase with a 28-residue GS linker (pMB1-βla-linker) was kindly provided by Professor J. Barwell (University of Michigan, USA).

Plasmids containing β-lactamase with WFL, WFT, WTL, WTT, SFL, SFT, STL and STT scFvs inserted into the GS linker were kindly provided by Dr Janet Saunders and Dr Paul Devine (University of Leeds, UK).

Plasmids containing β-lactamase with dp47d and HEL4 dAbs inserted into the GS linker were kindly provided by Dr Janet Saunders (University of Leeds, UK).

The pET23a plasmid containing GCSF was kindly provided by Dr Rhys Thomas (University of Leeds, UK).

The plasmid containing β-lactamase prepared for Golden Gate assembly (pMB1-βla-GG) was kindly provided by Romany Mclure (University of Leeds).

Plasmids containing the genes encoding IGLV1-44-germline, IGLV1-44-patient, IGLV6-57 and IGLV6-57-patient were synthesised by Twist Bioscience (CA, USA) in a pTwist vector.

Plasmids for protein expression, pET29a- IGLV6-57 and pET29a-IGLV6-57-patient were synthesised by Twist Bioscience (CA, USA) that included a N-terminal pelB periplasmic signal sequence.

Plasmid maps for pMB1-βla-linker, pMB1-βla- WFL and pET29a-IGLV657 are available in Appendices.

Plasmid	Insert	Promoter	Vector backbone	Antibiotic resistance
pMB1-βla-linker	β-lactamase 28 GS linker	pBAD	pMB1	Tetracycline
pMB1-βla-WFL	βla-scFv-WFL	pBAD	pMB1	Tetracycline
pMB1-βla-WFT	βla-scFv-WFT	pBAD	pMB1	Tetracycline
pMB1-βla-WTL	βla-scFv-WTL	pBAD	pMB1	Tetracycline
pMB1-βla-WTT	βla-scFv-WTT	pBAD	pMB1	Tetracycline
pMB1-βla-SFL	βla-scFv-SFL	pBAD	pMB1	Tetracycline
pMB1-βla-SFT	βla-scFv-SFT	pBAD	pMB1	Tetracycline
pMB1-βla-STL	βla-scFv-STL	pBAD	pMB1	Tetracycline
pMB1-βla-STT	βla-scFv-STT	pBAD	pMB1	Tetracycline
pMB1-βla-Dp47d	βla-Dp47d	pBAD	pMB1	Tetracycline
pMB1-βla-HEL4	βla-HEL4	pBAD	pMB1	Tetracycline
pET23a-GCSF	GCSF	T7	pBR322	Ampicillin
pMB1-βla-GG	βla-GG_{stop}	pBAD	pMB1	Tetracycline
pTwist-IGLV1-44	IGLV1-44	N/A	pMB1	Ampicillin
pTwist-IGLV1-44-patient	IGLV1-44-patient	N/A	pMB1	Ampicillin
pTwist-IGLV6-57	IGLV6-57	N/A	pMB1	Ampicillin
pTwist-IGLV6-57-patient	IGLV6-57-patient	N/A	pMB1	Ampicillin
pET29-IGLV6-57	pelB-IGLV6-57	T7	pMB1	Kanamycin
pET29-IGLV6-57-patient	pelB-IGLV6-57-patient	T7	pMB1	Kanamycin

Table 2.2 Plasmids obtained for this thesis

2.2 Molecular biology methods

2.2.1 Polymerase chain reaction

The polymerase chain reaction (PCR) was performed to amplify a specific region of DNA *in vitro*. The sequences and the purpose of the oligonucleotide primers designed to amplify the desired genes from select plasmids are shown in Table 2.3. Oligonucleotides were purchased from Eurofins Genomics, Germany.

Primer name	Sequence (5' → 3')	Use
GCSF Forward	GCTAGAATAGCCTCGAGCATGACTCCTCTCGGTCCTGCATC	Addition of *Xho*I restriction site 5' of GCSF gene for cloning into β-lactamase linker plasmid
GCSF Reverse	GCATACATAGCGGATCCCGGTTGCGCCAAATGGCGCAG	Addition of *Bam*HI restriction site 3' of GCSF gene for cloning into β-lactamase linker plasmid
Li33 Forward	GATGCTGAGATGCTCGAGCGAAGTGCAGCTGCTG	Addition of *Xho*I restriction site 5' of Li33 gene for cloning into β-lactamase linker plasmid
Li33 Reverse	GATGCTGAGATGGGATCCTTTAATTTCCACTTTGGTGC	Addition of *Bam*HI restriction site 3' of Li33 gene for cloning into β-lactamase linker plasmid

Table 2.3 Oligonucleotides used for PCR in this study. Restriction site *Xho*I is shown in blue and *Bam*HI in orange.

The following components were prepared in a 0.2 mL PCR tube on ice:

dsDNA template	100 ng
Upstream primer	0.5 μM
Downstream primer	0.5 μM
dNTPs	0.25 mM
$MgSO_4$	0, 2, 4 or 6 mM
ThermoPol reaction buffer	1×
Vent DNA polymerase	1 U
Nuclease-free water	to 50 μL

The components were gently mixed and transferred to a thermocycler. The temperature cycles for a typical reaction are shown in Table 2.4. The theoretical melting temperature (T_m) of primers was calculated from Equation 2.1 where n_{AT} corresponds to the number of AT nucleotide base pairs and n_{CG} corresponds to the number of CG nucleotide base pairs.

Step	Temperature (°C)	Time (s)
Initial denaturation	95	300
Denaturation	95	30
Annealing	5 below T_m	30
Elongation	72	60 per kb
Repeat denaturation, annealing and elongation (20-30 cycles)		
Final extension	72	300

Table 2.4 Temperature cycle for a typical PCR.

$$T_m = (n_{AT} \times 2) + (n_{GC} \times 4)$$

Equation 2.1 Calculation of theoretical melting temperature

The products from PCR were visualised by gel electrophoresis (2.2.2) and excised from the gel using a scalpel. DNA extraction was performed using QIAquick Gel Extraction Kit, as described by the manufacturer's instructions.

2.2.2 Agarose gel electrophoresis

The gel was prepared by dissolving 1.5 % (*w*/*v*) agarose in 1× Tris-acetate-EDTA (TAE) buffer (Section 2.1.4). The solution was heated using a microwave until the agarose had dissolved fully. Once cooled to < 50 °C, 0.5 μg/mL of ethidium bromide was added and the solution mixed. The gel was then poured into a 12 × 15 cm gel tray with a comb and allowed to set before use. Once set, the gel was transferred to the electrophoresis unit and the gel box was filled with 1× TAE until the gel was covered.

DNA samples were diluted in 6× Purple gel loading dye prior to loading to wells in the gel along with 5 μL of 1kb and 100 bp DNA ladders to allow size determination. Electrophoresis was performed at 100 V until the DNA fragments were suitably resolved. Gels were visualised using ultraviolet (UV) transillumination and photographed using Alliance Q9 Advanced gel doc system.

2.2.3 Restriction digests

Restriction digests of plasmids or PCR products were carried out using enzymes listed in 2.1.3. The following restriction digest reaction was prepared on ice:

Plasmid DNA or purified PCR product	1 μg
10× CutSmart buffer	1×
Enzyme 1 (20 U/μL)	20 U
Enzyme 2 (20 U/μL)	20 U
Nuclease-free water	to 50 μL

Control reactions were also performed containing single enzyme and enzyme free samples. Reactions were incubated at 37 °C for 1 h, followed by enzyme inactivation at 65 °C for 20 min.

Agarose gel electrophoresis (section 2.2.2) was performed to separate DNA fragments and remove restriction enzymes and unwanted by-products of the digestion. The required DNA fragments were excised using a scalpel and extracted from the gel using QIAquick Gel Extraction Kit, as described by the manufacturer's instructions.

2.2.4 Dephosphorylation of restriction endonuclease digests

To prevent re-ligation of the plasmid DNA, the 5' ends were dephosphorylated with Antarctic phosphatase as per the following reaction:

DNA	1 pmol DNA ends
Antarctic phosphatase reaction buffer	1×
Antarctic phosphatase	5 U
Nuclease-free water	to 20 µL

The reaction mixture was incubated for 30 min at 37 °C followed by enzyme inactivation at 80 °C for 2 min.

2.2.5 DNA ligation

Ligation of DNA fragments was performed by setting up molar ratios 3:1, 1:1 and 1:3 vector to insert with 1 U T4 DNA ligase in a 20 µL reaction with 1× T4 ligase buffer. A control reaction with no digested insert was also assembled. Reactions were incubated overnight at 16 °C and then kept on ice prior to transformation into DH5α supercompetent cells (section 2.2.9).

2.2.6 Site directed mutagenesis

Site directed mutagenesis was performed to introduce mutations into β-lactamase constructs containing GCSF, WFL and Li33 using Q5 mutagenesis. Primers were designed using the NEB online tool (http://nebasechanger.neb.com) using the most commonly used codon in *E. coli* and purchased from Eurofins Genomics.

2.2.6.1 Exponential amplification

Q5 hot start high fidelity DNA polymerase was used with the template DNA and the mutagenic primers in the following reaction:

Q5 Hot start high-fidelity master mix	1×
Forward primer	0.5 µM
Reverse primer	0.5 µM
Template DNA	25 ng
Nuclease-free water	to 25 µL

The reaction mixture was transferred to a thermocycler and the PCR cycling conditions outlined in Table 2.5 were performed.

Step	Temperature (°C)	Time (s)
Initial denaturation	98	30
Denaturation	98	10
Annealing	T_a	30
Elongation	72	30 per kb
Repeat denaturation, annealing and elongation (25 cycles)		
Final extension	72	120

Table 2.5 Temperature cycle for PCR for site directed mutagenesis.

2.2.6.2 Kinase, ligase and *Dpn*I (KLD) treatment

Following PCR, the amplified product is subject to treatment with kinase, ligase and *Dpn*I (KLD) enzymes to allow efficient phosphorylation, intramolecular ligation of plasmid DNA and to remove template DNA respectively. The KLD reaction outlined below was assembled and incubated at 25 °C for 5 min. Following incubation, 5 μL of the reaction was transformed into DH5α competent cells (section 2.2.9).

Q5 PCR product	1 μL
2× KLD reaction buffer	5 μL
10× KLD enzyme mix	1 μL
Nuclease-free water	3 μL

2.2.7 Golden Gate assembly

Golden Gate assembly utilises the simultaneous digestion with Type IIS restriction enzymes and ligation by a DNA ligase to enable scarless assembly (Figure 2.1). For cloning of V_L genes into β-lactamase, genes were synthesised by Twist Bioscience containing *Bsa*I restriction sites complimentary to those within the β-lactamase vector previously prepared for Golden Gate cloning (βla-GG_{stop}, Section 2.1.7). For DNA library synthesis applications, genes were amplified by epPCR with primers that introduce *Bsa*I restriction sites 5' and 3' to the gene as described in Section 2.4.2.

DNA concentrations were quantified using a Qubit fluorometer and the assembly reaction was set up with a 2:1 molar ratio of insert to 75 ng destination plasmid (βla-GG_{stop}), along with 1 μL NEB Golden Gate Assembly mix and 1× T4 DNA ligase buffer in a 20 μL reaction.

For cloning of one insert, the reaction was incubated at 37 °C for 5 min followed by enzyme inactivation at 60 °C for 5 min. Following incubation, 2 μL of the reaction

was transformed into DH5α competent cells (Section 2.2.9). For the construction of DNA libraries, this reaction was incubated for 1 min at 37 °C then 1 min at 16 °C for 45 cycles followed by 5 mins at 60 °C.

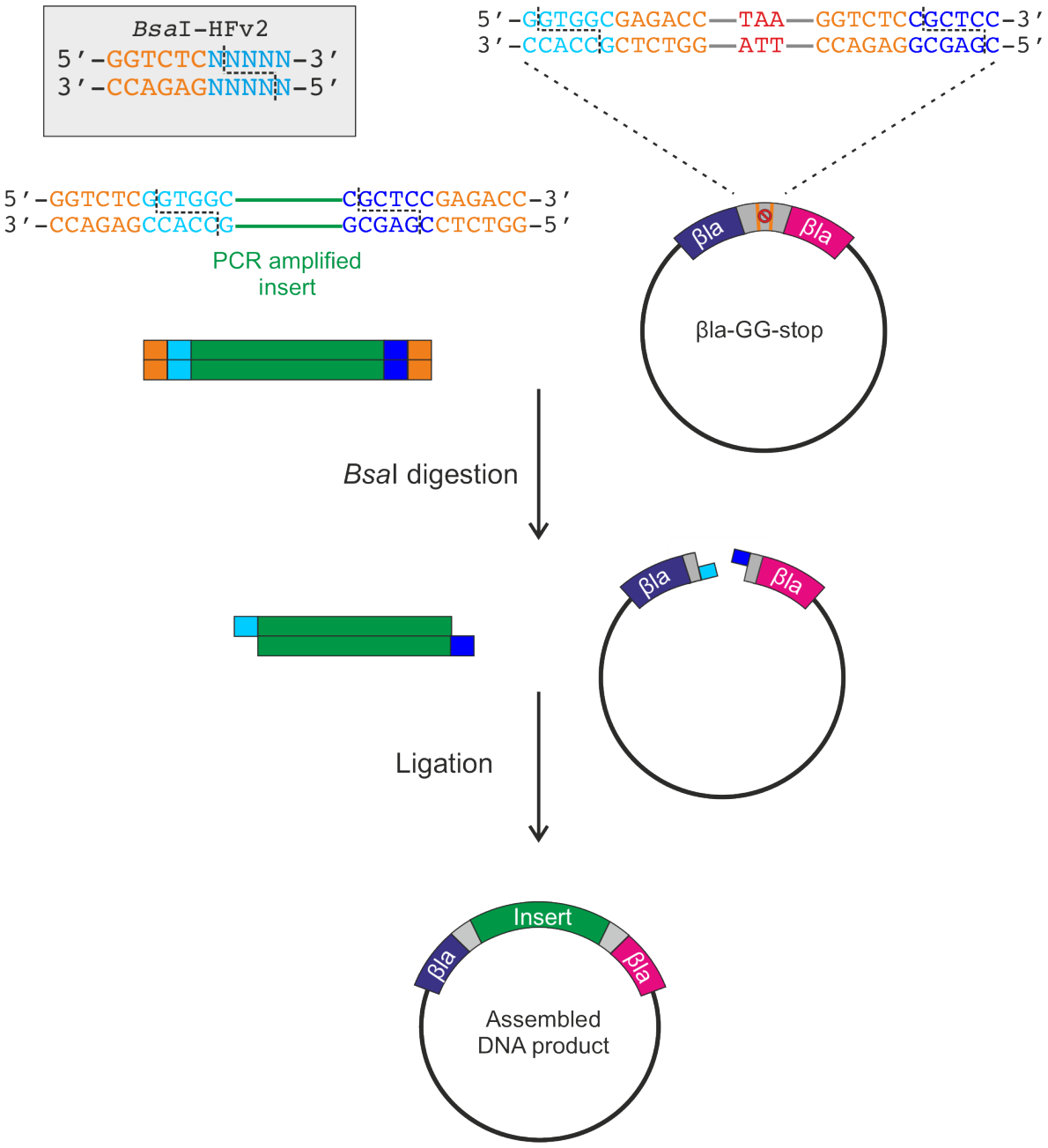


Figure 2.1 Overview of Golden Gate assembly. Golden Gate assembly uses type II restriction enzymes that cut outside of their recognition sequence. The recognition sequence for *Bsa*I-HFv2 is shown in the grey box, where orange text is the *Bsa*I recognition site and the blue text/dashed lines is the cut site. The gene of interest (green) is flanked 5' and 3' with *Bsa*I restriction sites (orange) and 5 bases complimentary to those in βla-GG$_{stop}$ (light and dark blue). Digestion of the vector and insert with *Bsa*I produces four base pair complementary overhangs (light and dark blue), that are then ligated resulting in scarless cloning. βla-GG$_{stop}$ includes a premature stop codon (red) so that any template carried over during library synthesis will produce a non-functional β-lactamase.

2.2.8 Preparation of competent cells

25 mL of LB medium was inoculated with a single colony from LB plate and incubated at 37 °C, 200 rpm for 12 – 14 h. 5 mL of the culture was used to inoculate 500 mL LB in a baffled flask. Cells were grown at 37 °C, 200 rpm until an OD_{600} of 0.4 was reached. Cells were harvested by centrifugation at 4,500 ×g (JLA 16.25 rotor) for 5 min at 4 °C. The supernatant was discarded and the cells were gently resuspended in 100 mL ice cold TFB1 solution (Section 2.1.4) and incubated on ice for 5 min. Cells were pelleted by centrifugation at 4,500 ×g (JLA 16.25 rotor) for 5 min at 4 °C. The supernatant was discarded, and cells were gently resuspended in 10 mL ice cold TFB2 solution (Section 2.1.4). Cells were kept on ice and 50 µL aliquots were pipetted into prechilled Eppendorf tubes. Cells were quickly frozen in liquid nitrogen and stored at −80 °C.

2.2.9 Transformation

Competent cells (section 2.1.6) were thawed on ice for 10 min before the addition of 100 ng plasmid DNA. Cells were incubated on ice for 30 min and then heat shocked at 42 °C for 45 s. The cells were incubated on ice for a further 2 min before the addition of 950 µL SOC medium. The cultures were incubated at 37 °C, 200 rpm for 1 h. 100 µL of the transformation reaction was spread onto LB agar plates containing the appropriate antibiotic. For ligation reactions, the transformation reaction was pelleted at 3,000 ×g for 3 min. The pellet was resuspended in 100 µL SOC and plated out onto the LB agar plate containing the selection antibiotic. Transformation plates were incubated overnight at 37 °C.

2.2.10 Plasmid DNA purification

Single colonies were picked from antibiotic selection agar plates and grown overnight (37 °C, 200 rpm) in 10 mL LB containing the appropriate antibiotic. Cells were pelleted at 4,000 rpm for 10 min and plasmid DNA was extracted using QIAprep Spin Miniprep Kit according to the manufacturer's instructions. The plasmids were eluted in nuclease-free water and the concertation was calculated using a nanodrop 2000 spectrophotometer using the optical density at 260 nm (A_{260}) (concentration (µg/mL) = 50 µg/mL × A_{260}). The plasmid DNA was diluted to 100 ng/µL for storage and sequencing. DNA for stock maintenance at −80 °C was stored in TE buffer (Section 2.1.4).

2.2.11 DNA sequencing to confirm cloning

To confirm the success of cloning, plasmid DNA was sequenced by Eurofins genomics, Germany. Sequencing of the β-lactamase *Xho*I and *Bam*HI cloning site was carried out using the primers in Table 2.6.

Primer	Sequence (5' → 3')
β-lactamase-linker-Forward	CGGAGCTGAATGAAGCCATACC
β-lactamase-linker-Reverse	TCACCGGCTCCAGATTTATCAGC

Table 2.6 β-lactamase sequencing primers.

2.3 Tripartite β-lactamase assay

2.3.1 Preparation of 48-well agar plates

LB agar was autoclaved at 121 °C, 15 psi for 20 min. Once cooled to less than 50 °C tetracycline and arabinose were added to give a final concentration of 10 μg/mL and 0.075 % (*w*/*v*) respectively. 300 μL of agar was then added to the first column of wells in the 48 well plate. The required volume of ampicillin was added to the agar and mixed thoroughly before pipetting into the next column on the plate (Table 2.7). This process was repeated giving 8 columns of increasing in ampicillin concentration. The plates were left to set in a sterile environment.

Ampicillin (μg/mL)	Agar volume (mL)	Ampicillin (μL)
0	100	0
20	96.4	193
40	92.8	186
60	89.2	178
80	85.6	171
100	82	164
120	78.4	157
140	74.8	150

Table 2.7 Preparation of 48-well agar plates with an ampicillin range between 0-140 μg/mL. Plates are prepared with a 10 mg/mL ampicillin stock to produce the 20 μg/mL ampicillin increments

2.3.2 Culture inoculation and induction

A single colony from *E. coli* SCS1 cells transformed with the appropriate plasmid was used to inoculate 100 mL sterile LB containing 10 μg/mL tetracycline. Cultures were incubated overnight at 37 °C, 200 rpm. 1 mL of overnight culture was used to inoculate 100 mL sterile LB medium containing 10 μg/mL tetracycline, and grown at 37 °C, 200 rpm until and OD_{600} of 0.6 was reached. Expression of the β-lactamase construct was then induced with 0.075 % (*w*/*v*) L-arabinose. Cultures were further incubated for 1 h at 37 °C, 200 rpm. Serial dilutions were performed on the induced culture in 10-fold increments into sterile 170 mM NaCl. From each dilution, 3 μL was pipetted into each column of the agar plates prepared in section 2.5.1. Plates were left to dry in a sterile environment and then incubated overnight at 37 °C for 18 h. The maximal cell dilution allowing growth (MCD_{GROWTH}) was then determined for each ampicillin concentration by visual inspection. The area under the antibiotic survival curve is calculated as a sum of the areas of seven trapezia (Equation 2.2), where x_i and y_i are the x and y axis values at one concentration of ampicillin (i).

$$A_{curve} = \sum_{i+1}^{7} \frac{y_i + y_{i+1}}{2} \times (x_{i+1} - x_i)$$

Equation 2.2 Sum of the areas of seven trapezia

2.3.3 Western and dot blot

2.3.3.1 Sample preparation

Cultures were grown as described in section 2.3.2, however 10 mL of culture was removed at OD_{600} = 0.6 before protein induction with L-arabinose for use as an uninduced control. Following the induction of β-lactamase expression for 1 h at 37 °C, 200 rpm, 10 mL sample was removed for the induced sample. The 10 mL cultures were harvested by centrifugation at 4,000 ×g for 10 min at 4 °C. Both sets of cell pellets were resuspended in phosphate buffered saline (PBS) to obtain an OD_{600} of 2.5.

2.3.3.2 Western blot

The sample was combined with loading dye (Section 2.1.4) and separated by SDS-PAGE (see section 2.7.6). The gels were transferred to a Bio-Rad 0.2 μm polyvinylidene fluoride membrane using a Trans-Blot Turbo Semi-Dry (Bio-Rad Ltd). The membrane was incubated overnight at 4 °C with the anti-β-lactamase antibody diluted 1:10,000 in 5 % (*w*/*v*) milk powder in TBST. The membrane was washed for 3 × 10 min in TBST at room temperature with agitation. The membrane was then incubated for 1 h with goat anti-rabbit IgG horseradish peroxidase conjugate diluted

1:10,000 in TBST. The membrane was then washed again with TBST for 3 × 10 min with agitation, before incubation with SuperSignal™ western pico chemiluminescent substrate. The emitted signal was detected using Q9 Alliance, chemiluminescence.

2.3.3.3 Dot blot

Nitrocellulose membrane (Amersham, 0.45 μm pore) was soaked in PBS prior to loading 50 μL sample using a SCIE-PLAS dot-blotting manifold. The membrane was blocked for 1 h with 5 % (*w*/*v*) milk powder TBST. The membrane was incubated overnight at 4 °C with the anti-β-lactamase antibody diluted 1:10,000 in 5 % (*w*/*v*) milk powder in TBST. The membrane was washed for 3 × 10 min in TBST at room temperature with agitation. The membrane was then incubated for 1 h with goat anti-rabbit IgG horseradish peroxidase conjugate diluted 1:10,000 in TBST. The membrane was then washed again with TBST for 3 × 10 min with agitation, before incubation with SuperSignal™ western pico chemiluminescent substrate. The emitted signal was detected using Q9 Alliance, chemiluminescence.

2.3.4 Nitrocefin activity assay

Nitrocefin was dissolved in DMSO to make a stock solution of 5 mg/mL. The solution was protected from light and stored in aliquots at −20 °C. Bacterial cultures expressing β-lactamase constructs were grown as described in section 2.3.2 and samples were prepared by correcting the final OD_{600} of cultures to 0.6. 180 μL bacterial culture was added to the wells of a flat-bottomed 96-well plate in triplicate. Nitrocefin stock was diluted in PBS to 1 mg/mL and 20 μL was added to each well. The plate was sealed with transparent, hydrophobic and gas permeable plastic films to prevent evaporation. The plate was incubated in a CLARIOstar microplate reader at 37 °C with agitation at 200 rpm and absorbance was measured at 486 nm every 60 s overnight.

2.4 DNA library synthesis

Two methods of library synthesis were used in this thesis. For both WFL and Li33 scFv proteins the megaprimer method was utilised. For V_L domain proteins IGLV1-44-germline, IGLV1-44-patient, IGLV6-57 and AL55 the golden gate method was performed.

2.4.1 Construction of mutant library using megaprimer method

The Diversify PCR Random Mutagenesis Kit was used to synthesise a scFv megaprimer (error rate of 8.1 (WFL) and 2.7 (Li33) mutations per 1000 bp) according to the manufacturer's instructions for each condition. Primers were used that anneal

to the Gly/Ser linker regions up- and down-stream of the scFv sequence (Table 2.8) to ensure mutations were only introduced into the target protein and not the β-lactamase gene (Figure 2.2a).

Primer	Sequence (5' → 3')
epPCR-Forward	GTGGTGGTGGCTCGA
epPCR-Reverse	AACCGCTCCCGGATC

Table 2.8 Primers used in epPCR for megaprimer synthesis.

For scFv-WFL the following components were assembled on ice:

1 ng/μL pMB1- βla-WFL	1 μL
10× Titanium *Taq* buffer	5 μL
8 mM $MnSO_4$	4 μL
2 mM dGTP	5 μL
50× Diversify dNTP mix	1 μL
10 μM Forward primer	0.5 μL
10 μM Reverse primer	0.5 μL
Nuclease free water	41 μL

The components were gently mixed and transferred to a thermocycler and the thermal cycling conditions shown Table 2.9 in were performed.

Step	Temperature (°C)	Time (s)
Initial denaturation	94	30
Denaturation	94	30
Annealing/Extension	68	60
Repeat denaturation, annealing and extension (25 cycles)		
Final extension	68	60

Table 2.9 Temperature cycle for Diversify epPCR.

The product was purified on a 1.5 % (*w*/*v*) agarose gel (as described in section 2.2.2) and the desired band was excised and purified using QIAquick Gel Extraction Kit, according to the manufacturer's instructions.

To prevent expression of the WT protein, a 'stop template' was created by introducing two stop codons into the WT gene by site directed mutagenesis (section 2.2.6). This stop template, was used to build the mutant library using the QuikChange Multi Lightning kit, for which the megaprimer synthesised in the step above was used to introduce the mutations from epPCR into the plasmid, and simultaneously, revert the stop codons back to the WT sequence (or a mutation) (Figure 2.2b).

The following reaction was set up in triplicate, along with a no megaprimer control sample:

10× QuikChange lightning muti buffer	5 μL
Stop template plasmid	0.33 μg
Megaprimer	0.42 μg
dNTP mix	1 μL
QuikSolution reagent	1.5 μL
QuikChange lightning enzyme blend	1 μL
Nuclease free water	to 50 μL

The reaction mixtures were transferred to a to a thermocycler and the PCR conditions were performed as per manufacturer's instructions (Table 2.5).

Each reaction was then incubated with 1 U *Dpn*I for 1 h to remove the stop template plasmid (Figure 2.2c) and the products were purified using QIAquick PCR purification kit and transformed as described in Section 2.4.3.

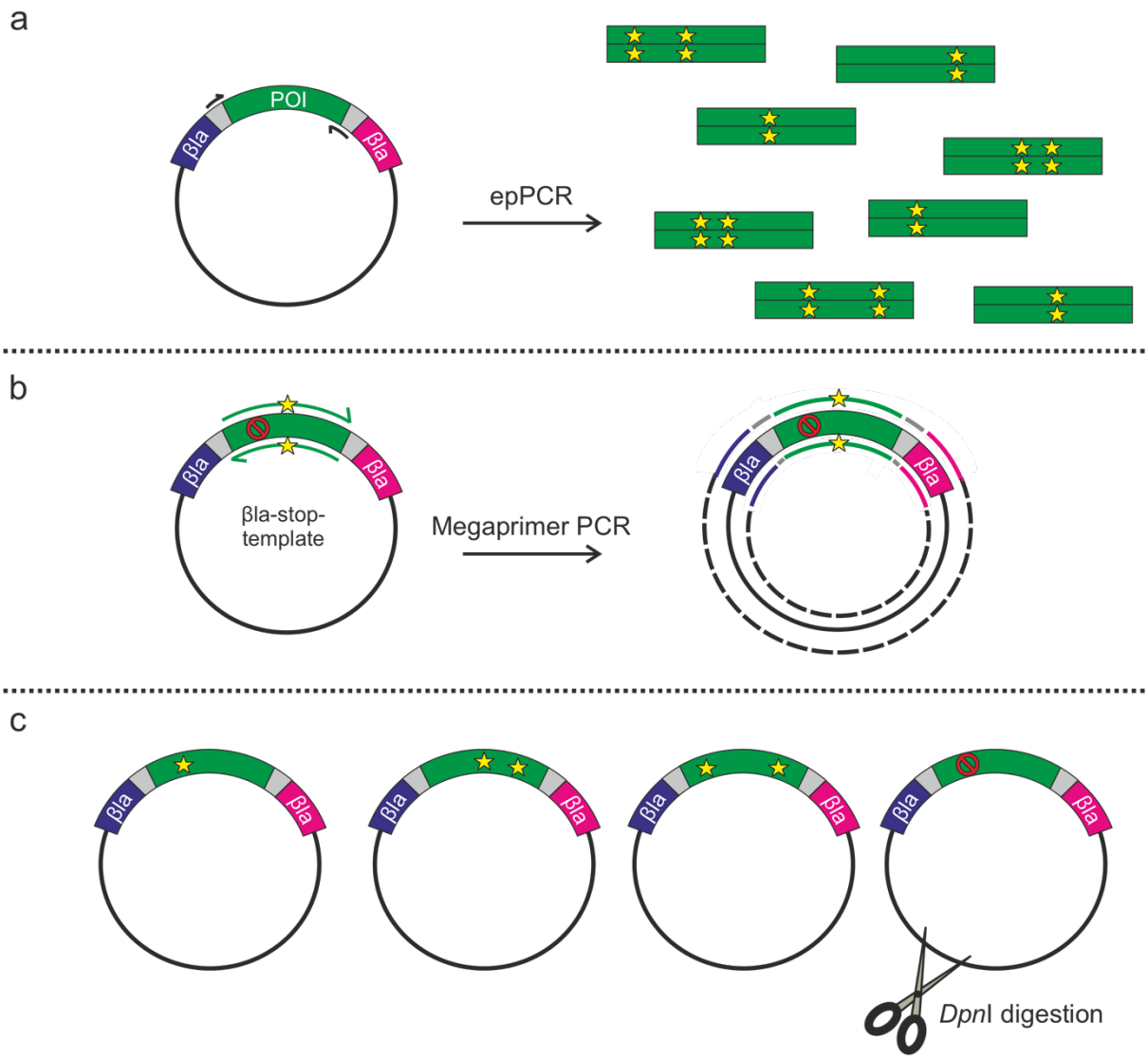


Figure 2.2 Overview of megaprimer method for library creation. a) epPCR is performed on the protein of interest (green) to introduce random mutations into the gene (yellow stars). b) The epPCR product from a) is then used as a megaprimer for PCR. The megaprimer binds to the full WT gene sequence in the stop template plasmid to introduce the mutations into the gene. Through PCR extension the full plasmid is amplified (dashed lines). The stop codon (red) in the stop template will be mutate back to the WT codon. c) *Dpn*I digestion removes the stop template DNA, leaving a library of mutants to be transformed. Any stop template remaining after *Dpn*I digestion will produce a non-functional β-lactamase.

2.4.2 Construction of mutant library using golden gate assembly

GeneMorph II random mutagenesis kit was used for epPCR on V_L domain proteins (IGLV1-44-germline, IGLV1-44-patient, IGLV6-57 and IGLV6-57-patient) with an error rate of 6 mutations per 1000 bp. Primers were used that anneal to the Gly/Ser

linker regions up- and down-stream of the V_L sequence and introduce *Bsa*I restriction sites 5' and 3' to the epPCR product (Table 2.8).

Primer	Sequence (5' → 3')
GG-epPCR-Forward	GGGAATGGTCTCGGTGGCTCGAGC
GG-epPCR-Reverse	ACATGCGGTCTCCGCTCCGGATCC

Table 2.10 Primers used in epPCR for golden gate assembly. The *Bsa*I restriction site is highlighted in red and the 4 base overhangs produced from the digestion is shown in green.

The PCR components outlined below were assembled on ice thermocycler and the PCR conditions were performed as per manufacturer's instructions (Table 2.4). Target DNA refers to the DNA sequence to be amplified, not the total amount of plasmid DNA in the reaction.

Target DNA	200 ng
10× Mutazyme II reaction buffer	5 µL
40 mM dNTP mix	1 µL
10 µM Forward primer	0.5 µL
10 µM Reverse primer	0.5 µL
Mutazyme II DNA polymerase	1 µL
Nuclease free water	to 50 µL

The product was purified on a 1.5 % (*w*/*v*) agarose gel (as described in section 2.2.2) and the desired band was excised and purified using QIAquick Gel Extraction Kit, according to the manufacturer's instructions. This epPCR product was cloned into βla-GG_{stop} as described in Section 2.2.7. The three replicate golden gate reaction products were pooled together, purified using QIAquick PCR purification kit and transformed as described in Section 2.4.3.

2.4.3 Library transformation

Products were transformed into TG1 *E. coli* (Table 2.1) by electroporation. 25 µL of cells were transferred to prechilled 0.2 cm gap cuvettes and 1 µL library was added directly before electroporation. Cells were electroporated (2.5 kV field strength,

335 Ω resistance and 15 μF capacitance) and immediately after the cuvette was washed with 1 mL SOC medium and transferred to a 50 mL Falcon tube. Each cuvette was rinsed with a further 1 mL recovery medium and transferred to the Falcon tube. Samples were incubated at 37 °C, 250 rpm for 1h recovery. Ten-fold serial dilutions were plated out onto LB agar plates containing 10 μg/mL tetracycline. The remaining cells were pelleted at 3,000 ×g for 5 min, and the pellet resuspended in 2 mL recovery medium and plated onto 12 × 12-inch LB agar plates containing 10 μg/mL tetracycline. Plates were incubated overnight at 37 °C, following which the number of colonies were counted to estimate the library size (Equation 2.3). The library bioassay plate was scraped into 10 mL LB containing 50 % (*v*/*v*) glycerol. 1 mL glycerol stock was made, and the remaining was purified using PureYield plasmid midiprep system.

$$Library\ size\ =\ number\ of\ colonies\ \times\ dilution\ plate\ \times \left(\frac{dilution\ volume}{volume\ plated}\right) \times\ total\ culture\ volume$$

Equation 2.3 Calculation to estimate library size.

2.5 Directed evolution

2.5.1 Plate preparation

Directed evolution assay plates were prepared by sterilising 12 × 12 inch bioassay plates with 70 % (*v*/*v*) ethanol and leaving to dry under sterile conditions. 500 mL 2.5 % (*w*/*v*) LB, 1.5 % (*w*/*v*) agar was autoclaved for 20 min at 121 °C, 15 psi and left to cool below 50 °C prior to the addition of 10 μg/mL tetracycline, 0.075 % (*w*/*v*) L-arabinose and the required concentration of ampicillin for the screen. The agar was poured into two 12 × 12-inch plates and left to set under sterile conditions.

2.5.2 Library growth

SCS1 supercompetent cells (Table 2.1) were thawed on ice for 10 min and 50 μL cells transferred to a 14 mL transformation tube. 2 μL plasmid DNA (WT or library) were added to the cells and incubated on ice for 30 min before heat shocking at 42 °C for 45 s. Following 5 min incubation on ice 950 μL SOC medium was added to cells and incubated at 37 °C, 200 rpm for 1 h. 3 mL SOC medium was then added to cells with 10 μg/mL tetracycline. Cells were further incubated for 1 h (37 °C, 200 rpm) and β-lactamase expression was induced with 0.075 % (*w*/*v*) L-arabinose for 1h (37 °C, 200 rpm). The culture was spread onto the prepared assay plates (section 2.5.1) and incubated overnight at 37 °C.

2.5.3 Sanger sequencing

Single colonies were picked from the plate and added to separate wells of a sterile 96-well plate containing 100 μL LB containing 10 μg/mL tetracycline. The plate was sealed with hydrophobic and gas permeable plastic films to prevent evaporation and incubated over night at 37 °C, 200 rpm. Glycerol stocks of each clone was prepared by adding 100 μL sterilised 50 % (*v*/*v*) glycerol to each well. 100 μL of each glycerol culture was transferred to separate wells of a 96-well plate and sent for sequencing by GENWIZ, using the primers described in Table 2.6 and the remaining cultures were stored at −80 °C.

2.5.4 Next generation sequencing

The evolved libraries were scraped into 10 mL LB containing 50 % (*v*/*v*) glycerol and DNA was extracted using Qiagen miniprep kit as per manufacturer's instructions. A PCR amplicon of evolved DNA was created using the primers in Table 2.8 that anneal to regions up- and down-stream of the target sequence and PCR conditions described in Section 2.2.1. The product was purified using QIAquick PCR purification kit, eluted in 50 μL 1× TE buffer and prepared for NGS using NEBNext Ultra II DNA library prep kit for Illumina.

2.5.4.1 Adaptor ligation

The PCR product was first prepared for adaptor ligation through end repair, 5' phosphorylation and dA-tailing. The following components were assembled in a nuclease-free tube and mixed thoroughly.

NEBNext Ultra II end prep enzyme mix	3 μL
NEBNext Ultra II end prep reaction buffer	7 μL
DNA	50 μL

The reaction was placed in a thermocycler at 20 °C for 30 mins followed by 65 °C for 30 mins.

After incubation, the following components were added directly to the end prep reaction mixture and mixed thoroughly.

End prep reaction mixture	60 μL
NEBNext adaptor for Illumina	2.5 μL
NEBNext Ultra II ligation master mix	30 μL
NEBNext ligation enhancer	1 μL

The reaction was incubated at 20 °C for 15 mins, following which 3 μL USER enzyme was added, mixed well and incubated for a further 15 mins at 37 °C.

2.5.4.2 Sample purification and size selection

NEBNext sample purification beads were resuspended by vortexing and 25 μL added to the ligation mix from Section 2.5.4.1. The samples were mixed thoroughly and incubated at room temperature for 5 mins. The tube was placed on a magnetic stand for 5 mins until the solution was clear, separating the beads from the supernatant. The supernatant was transferred to a new tube and the beads were discarded. This process was repeated adding 10 μL beads to the supernatant, mixing well and incubating for 5 mins. The beads were separated on the magnetic stand and the supernatant was discarded. The beads were washed by adding 200 μL 80 % (*v*/*v*) ethanol to the beads whilst on the magnetic stand and incubated for 30 s. The supernatant was discarded, and the wash step was repeated. The beads were air dried for 5 mins on the magnetic stand with the lid open to remove traces of ethanol. The target DNA was eluted from the beads by adding 17 μL 1× TE, vortexing the samples and incubating for 2 mins at room temperature. The tube was placed on the magnetic stand and once separated, 15 μL supernatant was transferred to a new tube for PCR amplification.

2.5.4.3 PCR enrichment

The following components were assembled on ice and mixed thoroughly:

Adaptor ligated DNA fragment	15 μL
NEBNext Ultra II Q5 master mix	25 μL
i7 index primer	5 μL
i5 index primer	5 μL

The reaction was transferred to a thermocycler and PCR amplified using the conditions in Table 2.11.

Step	Temperature (°C)	Time (s)
Initial denaturation	98	30
Denaturation	98	10
Annealing/Extension	65	75
Repeat denaturation, annealing and extension (3 cycles)		
Final extension	65	300

Table 2.11 PCR cycling conditions for NGS library preparation.

The PCR product was purified from the reaction mixture as outlined in Section 2.5.4.2, using only one application of sample purification beads (45 μL) and eluting the beads in 33 μL 1× TE. The eluted DNA was quantified using a Qubit fluorometer and libraries were pooled together in equimolar ratios. The pooled library was sequenced using Illumina MiSeq 2 × 250 bp, spiked with 30 % PhiX by GENEWIZ. Information on data analysis can be found in Section 2.8.3.

2.6 Protein purification

2.6.1 IgG purification

IgGs used in this study were synthesised by the Biologics expression team at AztraZeneca and IgG expression vector cloning was performed by Dr James Button and Dr Janet Saunders (AstraZeneca). Sequences were eukaryote codon-optimised and the V_H domain cloned into the IgG V_H IgG1 TM YTE expression vector (pEU1.6) and the V_L domain cloned into the IgG V_L lambda expression vector (pEU4.4). The plasmids were co-transfected into HEK293/EBNA mammalian cells for expression and IgG proteins were purified from the culture medium using Protein A chromatography.

2.6.2 V_L domain purification

A starter culture for prepared for protein expression by inoculating 100 mL sterile LB medium containing 50 μg/mL kanamycin with a single colony from transformed expression cells, BL21 (DE3), with either pET29-IGLV6-57 or pET29-IGLV6-57-patient (Section 2.1.7). The inoculated medium was incubated at 37 °C, 200 rpm for 16 hours. 2 mL of the starter culture was used to inoculate 0.5 L autoinduction medium (Section 2.1.5) containing 50 μg/mL kanamycin. The cultures were

incubated at 20 °C for 24 h with shaking at 200 rpm, before harvesting by centrifugation at 6,000 ×g (JLA 8.1 rotor) for 30 min at 4 °C.

For periplasmic extraction, the pellet was resuspended in 3× volume (i.e., 30 mL per 10 g cells) of ice cold 50 mM Tris pH 8.5 containing 20 % (*w*/*v*) sucrose. The cells were incubated on ice for 1 h with agitation before centrifugation at 4,500 ×g (JLA 16.25 rotor) for 20 min at 4 °C. The supernatant was removed and kept on ice, and the pellet resuspended in the same volume ice cold H_2O. The cells were incubated on ice for a further hour 1 h with agitation before centrifugation at 4,500 ×g (JLA 16.25 rotor) for 20 min at 4 °C. The supernatant was combined with the previous supernatant and dialysed into 50 mM Tris pH 8.5 at 4 °C. The periplasmic extract was loaded onto a Q-Sepharose anion exchange column equilibrated in 50 mM Tris pH 8.5. The protein was eluted with a linear gradient of 0-500 mM NaCl and monitored by absorbance at 280 nm. Fractions containing the V_L domain were analysed by SDS-PAGE (Section 2.7.6) and dialysed into PBS, pH 7.4.

The V_L domain was then loaded onto HiLoadTM 26/60 Superdex 75 prep grade gel filtration column. The protein was eluted from the column with PBS, pH 7.4 at a flow rate of 3 mL/min. The molecular mass of the protein was confirmed by electrospray ionisation mass spectrometry performed by Rachel George at the The Biomolecular Mass Spectrometry Facility (University of Leeds). The protein was snap frozen and stored at −80 °C.

2.7 Biophysical and biochemical methods

2.7.1 High performance size-exclusion chromatography

HP-SEC data were generated by Dr Christopher Lloyd (AstraZeneca) and the Biologics expression team at AstraZeneca. HP-SEC was performed using an Agilent 1,100 series HPLC fitted with a TSK SWXL HPLC guard column and TSK-GEL G3000SW$_{XL}$ HPLC column. 50 μL of IgG at 1 mg/mL in PBS was injected at a flow rate of 1 mL/min using 0.1 M sodium phosphate, 0.1 M sodium sulfate, pH 6.8 as the mobile phase buffer.

2.7.2 Affinity-capture self-interaction nanoparticle spectroscopy

AffiniPure goat anti-human IgG Fcγ fragment specific (IgGα-Fc) and ChromePure goat IgG, whole molecule (IgG$_{Whole}$) were buffer exchanged into 20 mM potassium acetate, pH 4.3 and diluted to 0.4 mg/mL. 9 mL citrate-stabilised 20 nm gold nanoparticles were incubated with 600 μL IgGα-Fc and 400 μL IgG$_{Whole}$ for 2 h at room temperature. Nanoparticles were blocked with 0.1 μm 2,000 MW thiolated PEG at room temperature for 1-2 h. Nanoparticles were concentrated to 800 μL in siliconized Eppendorf tubes and stored at 4 °C. 45 μL of 50 μg/mL antibody samples were mixed with 5 μL nanoparticle solution and incubated at room temperature for 30 min. The mixture was transferred to a 384-well polystyrene UV transparent plate and the absorbance measured on SPECTROstarNano plate reader from 400 nm to 700 nm in 1 nm increments. The maximum absorbance was determined (the plasmon wavelength) and the redshift in plasmon wavelength compared to nanoparticles in the absence of antibodies was then calculated by subtracting one from the other.

2.7.3 Differential scanning fluorimetry

20 μL of 0.52 mg/mL IgG in PBS was added to a white PCR plate. SYPRO Orange protein stain gel (5000× stock concentration) was diluted to 40× in distilled H_2O prior to the addition of 5 μL to each well. The plate was sealed with adhesive sealing film, and samples were heated from 20 – 95 °C in 0.2 °C increments on a CFX96 Real-Time PCR system. The melt curves were obtained by measuring the fluorescence intensities using the FRET channel with excitation from 450 to 490 nm and detection from 660 to 580 nm.

2.7.4 Homogeneous time-resolved fluorescence detection

Test IgGs stocks were diluted in PBS containing 0.5 % (*w*/*v*) BSA to 200 nM, which were then serially diluted in two-fold increments 11 times. The following components were prepared in assay buffer (PBS, 0.5 % (*w*/*v*) BSA, 0.4 M potassium fluoride): 24 nM Dylight650 labelled IgG-WFL, 12.5 μg/mL streptavidin Europium cryptate and 0.8 nM biotinylated Human NGF also containing 12.5 μg/mL streptavidin Europium cryptate in the solution.

5 μL of each serial diluted test IgG was transferred to a white 384 well plate. 2.5 μL Dylight650 IgG-WFL and 2.5 μL NGF + streptavidin Europium cryptate solution was added to each well. Two negative controls were prepared, one that lacked just test IgG and one that lacked both test IgG and antigen (5 μL PBS, 0.5 % (*w*/*v*) BSA, 2.5 μL Dylight650 IgG-WFL and 2.5 μL NGF + streptavidin Europium cryptate or 2.5 μL streptavidin Europium cryptate). The plate was centrifuged for 1 min at 3,000 ×g sealed with adhesive sealing film and incubated at room temperature for 2.5 h. Fluorescence was measured on an EnVision plate reader with the following settings: 100 flashes, delay 70, cycle 2000, Excitation UV2 (TRF) 320 nm, Emission APC 665 (Bandwidth 7.5 nm), Emission Rhodamine 590 (Bandwdith 20 nm), mirror D400/630. The HTRF ratio is calculated by Equation 2.4 and the % DELTA F is calculated by Equation 2.5.

$$HTFR\ ratio = \frac{665\ nm}{590\ nm} \times\ 10{,}000$$

Equation 2.4 Calculation of HTFR ratio

$$\%\Delta F\ =\ \frac{(Sample\ ratio - negative\ control\ ratio)}{Negative\ control\ ratio} \times 100$$

Equation 2.5 Calculation of % DELTA F

2.7.5 Poly (ethylene glycol) (PEG) precipitation assay

PEG precipitation assays were performed by Dr Janet Saunders (AstraZeneca). A 40 % (*w*/*v*) PEG 10,000 solution was prepared in PBS and the pH corrected to 7. In a 96-wll plate, PEG solution, PBS and 20 μL of IgG stock solution were combined to achieve a 0-10 % (*w*/*v*) concentration range of PEG and a final IgG concentration of 0.5 mg/mL. Samples were set up in triplicate, and the plate sealed with adhesive sealing film and incubated at 4 °C for 24 h. Following incubation, the samples were

thoroughly mixed before 2 μL of each sample was transferred to a Lunatic plate for turbidity measurement at 500 nm on a Lunatic plate reader. The turbidity of PBS controls was subtracted from final readings.

2.7.6 Sodium dodecyl Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)

Tris-tricine buffered sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) was used to separate proteins according to their molecular weight. Two glass plates were assembled according to manufacturer's instructions using a 1.5 mm spacer. The resolving and stacking gels were made up of the components in Table 2.12. The resolving gel mixture was poured to within 2 cm of the top of the glass plates and immediately after the stacking gel was poured on top of the resolving gel. A 12-well comb was inserted to create wells for sample loading and gels were left for a minimum of 1 h to set.

Solution component	Resolving gel (mL)	Stacking gel (mL)
30 % (*w*/*v*) Acrylamide: 0.8 % (*w*/*v*) bis-acrylamide	7.5	0.83
3 M Tris-HCl, 0.3 % (*w*/*v*) SDS ,pH 8.45	5	1.55
H_2O	0.44	3.72
Glycerol	2	0
10 % (*w*/*v*) ammonium persulfate	0.05	0.1
Tetramethylethylenediamine (TEMED)	0.005	0.005

Table 2.12 Components for tris-tricine buffered SDS-PAGE gel

Cathode buffer and anode buffer (Section 2.1.4) were added to the inner and outer reservoir of the gel tank respectively prior to sample loading. Protein samples were diluted 1-fold with 2× SDS-PAGE loading buffer (Section 2.1.4and boiled for 5 min and then 15 μL sample was loaded to the gel. To estimate the of molecular weight of resolved protein bands, 5 μL Precision Plus Protein Dual Xtra Prestained protein standard was loaded into one lane on the gel. Gels were electrophoresed at a constant current of 30 mA until the samples entered the resolving gel, where the current was increased to 65 mA until the dye reached the bottom of the gel. The gels were removed from their casts and incubated using Instant Blue stain.

2.7.7 ThioflavinT (ThT) aggregation assay

ThT assays for V_L domains were prepared with 40 μM protein and 20 μM ThT in PBS buffer (pH 7.4) containing 0.5 mM SDS and 0.05 % (*w*/*v*) NaN_3. 100 μL samples of 10 replicates were added to the wells of a 96-well flat bottom assay plate and sealed with adhesive sealing film. Fibril kinetics were monitored in a FLUOstar Omega plate reader at 37 °C with continuous orbital agitation at 600 rpm. The fluorescence of ThT was excited at 444 nm and fluorescence emission was monitored at 480 nm.

2.7.8 Transmission electron microscopy (TEM)

TEM images were collected by Dr Nicolas Guthertz (University of Leeds). Replicate samples from the ThT assay were pooled. Insoluble material was separated by centrifugation (14,000 rpm in a benchtop microcentrifuge, 15 min), and the pellet resuspended in 100 μL acidified water (pH 2, adjusted with hydrochloric acid). 5 μL was added to a glow-discharged carbon-coated copper grid for 1 min. Excess liquid was blotted away and the grid was stained with 5 μL 1 % (*w*/*v*) uranyl acetate for 20 s before being washed with 5 μL water. Micrographs were recorded on a Tecnai T12 transmission electron microscope.

2.7.9 Circular dichroism (CD)

For thermal denaturation experiments an initial spectrum of the sample (20 μm protein in PBS pH 7.4), was obtained at 25 °C. Thermal denaturation experiments were performed by setting up a temperature gradient from 20 to 90 °C in 5 °C steps. Protein samples were equilibrated for 120 s at each temperature before CD spectra were taken. Each spectrum was acquired from 190 nm to 260 nm with a step size of 1 nm and 1 s per point sampling. The thermal melt data were analysed using the software package CDpal[248].

2.8 Bioinformatic methods

2.8.1 *In silico* aggregation

A model of the structure of scFv-WFL (created by mutating PDB 5J7Z[46] in PyMol 2.1.0) was used for this analysis.

2.8.1.1 CamSol

The webserver for CamSol[144] was used to generate a structurally corrected profile at 10 Å patch radius to identify soluble and insoluble amino acids. The webserver can be located at: http://www-vendruscolo.ch.cam.ac.uk/camsolmethod.html

2.8.1.2 Aggrescan

Aggrescan3D 2.0[143] server was used to predict aggregation propensity in dynamic mode with a 10 Å radius and stability calculation option was selected using FoldX to optimise input structure. The webserver can be located at: http://biocomp.chem.uw.edu.pl/A3D2/

2.8.1.3 Spatial aggregation propensity

Spatial aggregation propensity (SAP) calculations were performed using CHARMM[249] simulations and method described by Chennamsetty et al.[149] using a 10 Å radius.

2.8.2 Relative surface accessibility

RSA values were calculated by taking the absolute solvent accessible surface area for the residue in the model of the structure of scFv-WFL (created by mutating PDB 5J7Z[46] in PyMol 2.1.0) and dividing it by the maximum possible area for the amino acid type as described by Miller et al[250].

2.8.3 Next generation sequencing analysis

Output files from MiSeq 2 × bp were demultiplexed by GENEWIZ according to the i7 and i5 index primers used in the PCR (Section 2.5.4.3). FastQ files were aligned to the reference sequence using breseq[251] (version 0.35.4), bowtie2[252] (version 2.4.2) and R (version 3.2.2). The resulting .bam file was converted to a .sam file using samtools[253,254]. Insertion and deletions were filtered out and the remaining aligned sequences were translated in frame using Biopython[255] and scripts prepared by Michael Davies and Romany Mclure (University of Leeds).

Chapter 3

Screening and identifying aggregation hotspots *in vivo*

3.1 Objectives

The preliminary objective of this thesis was to develop a screen to differentiate between aggregation-prone and aggregation-resistant biopharmaceuticals. The screen needs to be able to detect aggregation of therapeutically relevant scaffolds and to be applied to proteins that aggregate via different mechanisms. The screen must be highly sensitive and be able to detect differences due to substitutions of single residues. The assay can then be developed into a directed evolution screen to identify aggregation-prone regions in proteins.

3.2 Tripartite β-lactamase screen for aggregation

As described in Section 1.7.1 the tripartite β-lactamase (TPBLA) sensor has been used as a reporter protein for *in vivo* screening of protein stability[242,247] and screening small molecule inhibitors of amyloid formation[246]. These studies demonstrated the ability to insert large proteins (MBP, 43 kDa) into the linker of β-lactamase[242] and validated the utility of the assay for analysis of proteins that are intrinsically disordered form amorphous aggregates or ordered aggregates such as amyloid[246].

Here, we employ the TPBLA sensor to detect the aggregation of biopharmaceuticals. In this assay, the test protein is inserted into a Gly/Ser linker that separates two domains of TEM-1 β-lactamase (Figure 3.1a). The assumption of this assay is that upon correct folding of the POI, the two domains of β-lactamase are brought into close proximity to associate and form an active enzyme, capable of hydrolysing β-lactam antibiotics (Figure 3.1b). *E. coli* expressing a stable, aggregation-resistant protein can therefore survive in the presence of β-lactam antibiotics (Figure 3.1c). Conversely, if the POI aggregates, the activity of β-lactamase is reduced through co-aggregation and/or cellular degradation of the tripartite construct, causing *E. coli* to lose resistance β-lactam antibiotics (Figure 3.1d). The assay therefore directly links the aggregation-propensity of the test protein to the susceptibility of the bacterium to β-lactam antibiotics.

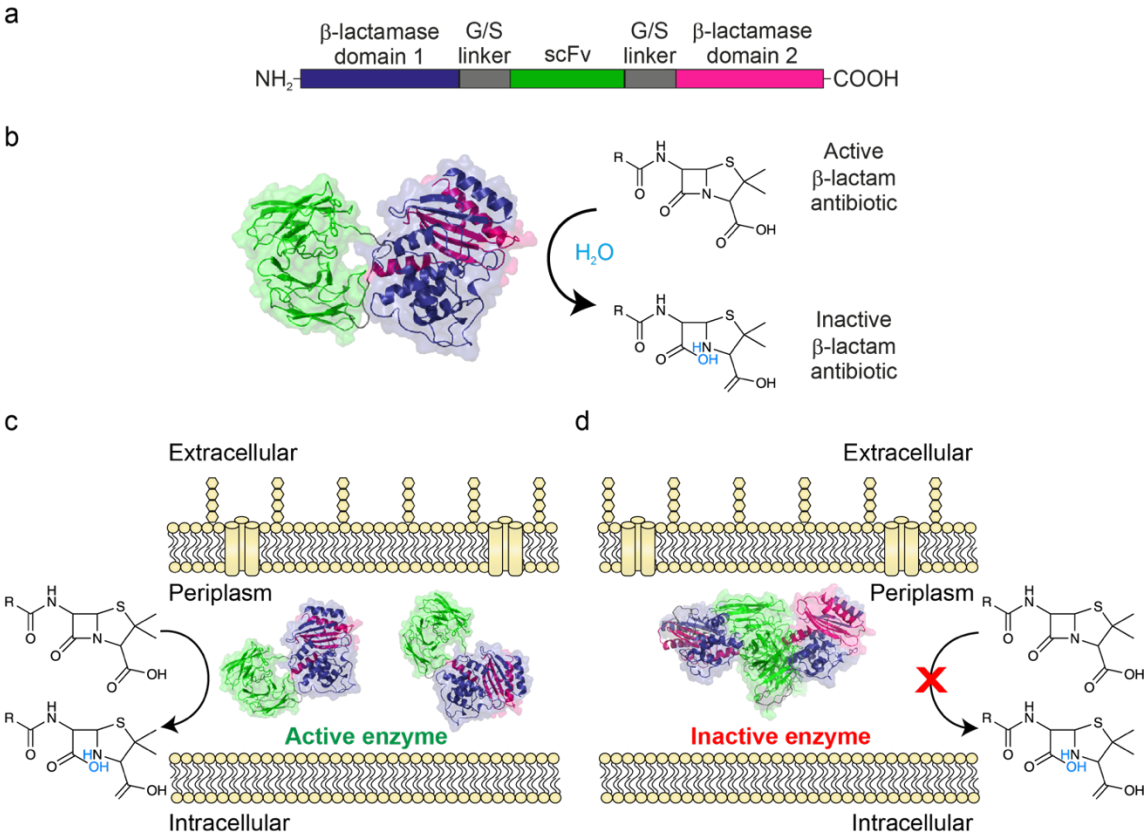


Figure 3.1 Tripartite β-lactamase assay for protein aggregation. a) The POI, such as a scFv, is inserted into a Gly/Ser linker (grey) separating two domains of β-lactamase (purple and pink). b) Correct folding brings together the two domains of β-lactamase, forming an active enzyme capable of hydrolysing β-lactam antibiotics. c) Expression of an aggregation resistant protein in the periplasm of *E. coli* enables bacterial resistance to β-lactam antibiotics. d) If the POI aggregates, β-lactamase is inactive, and bacteria become sensitive to β-lactam antibiotics.

The TPBLA has previously been established using BL21(DE3) *E. coli*[246,256], however for the development of this screen as a directed evolution platform in this thesis, the cell line had to be altered to identify a relationship between phenotype and genotype. BL21 (DE3) are routinely used for protein expression, however, are poor for DNA extraction as their endonuclease I activity may degrade plasmids. An alternative strain of *E. coli,* SCS1, was chosen for their endonuclease (endA) deficiency that improves the quality of miniprep DNA extraction and their recombination (recA) deficiency that enhances plasmid stability and reduced recombination. Any previously published data has been reproduced in this thesis using SCS1 *E. coli.*

Briefly, the assay is performed by culturing SCS1 cells transformed with the construct for screening. Once cells reach an OD_{600} of 0.6 protein expression is induced by the addition of 0.075 % (w/v) arabinose for 1 h at 37 °C, 200 rpm. Cultures are serially

diluted 10-fold before plating out onto a 48-well agar plate containing increasing concentrations of the β-lactam antibiotic, ampicillin. Plates are then incubated at 37 °C for 18 h following which they are scored for the maximal cell dilution at which growth occurs (MCD_{GROWTH}) (Figure 3.2, Section 2.3).

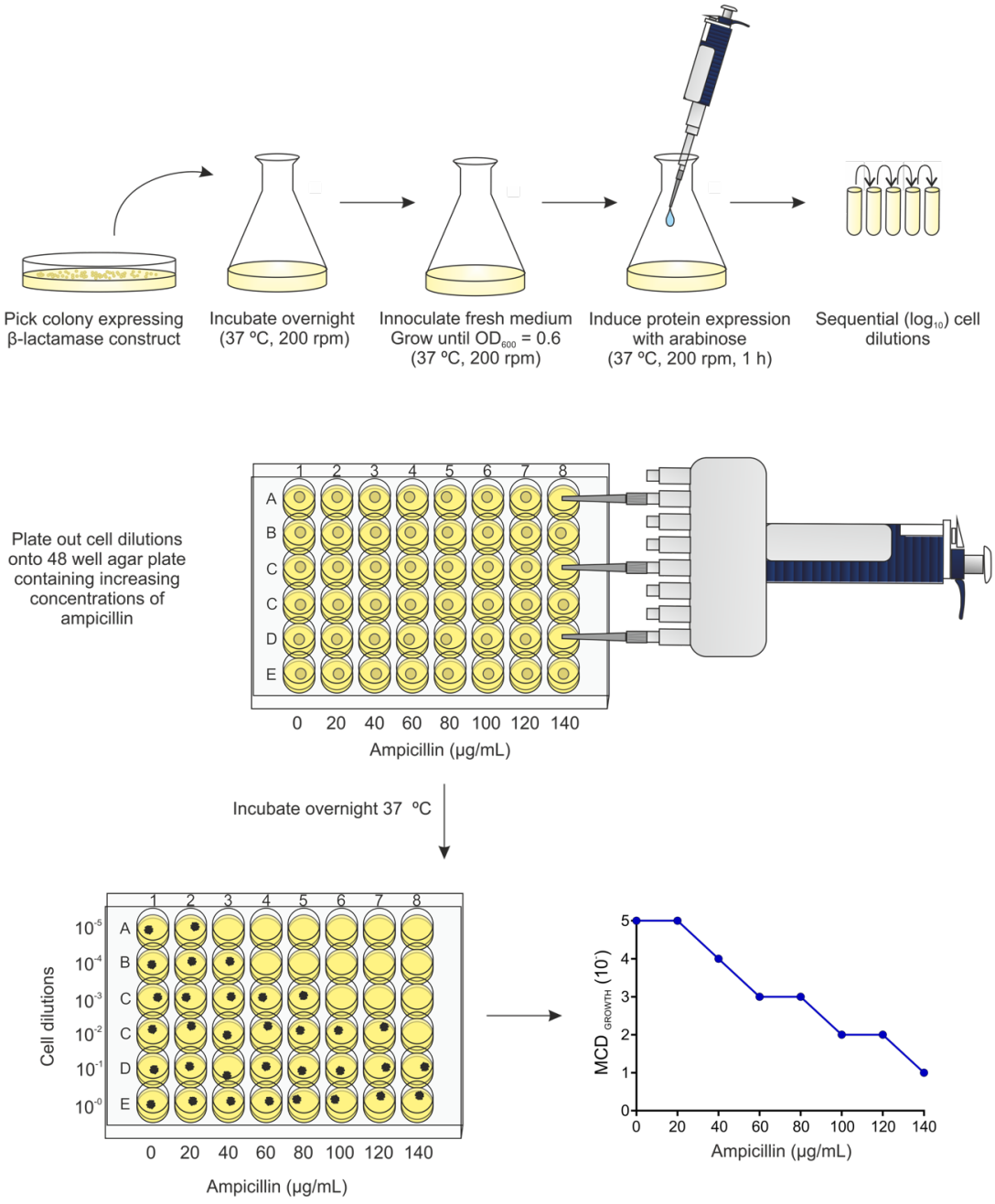


Figure 3.2 Schematic of the *in vivo* growth assay. Colonies are transformed with the β-lactamase fusion construct and are cultured until an OD_{600} of 0.6 is reached. Protein expression is induced by the addition of arabinose. Cultures are serially diluted into 170 mM NaCl and 3 μL pipetted onto each well that contains solid growth medium with increasing concentrations of antibiotic in each column. Plates are incubated at 37 °C overnight. The maximal cell dilution at which growth occurs is scored by visual inspection.

3.3 Screening therapeutically relevant protein scaffolds

3.3.1 WFL and STT: a model pair of mAbs

The results in this thesis will focus on a pair of mAbs with different aggregation propensities: WFL and STT. This pair of model proteins were originally generated by AstraZeneca, the industrial sponsor of this PhD studentship, who provided access to their sequences and biophysical characteristics for this study. MEDI1912 (called IgG-WFL herein) was generated by *in vitro* affinity maturation of MEDI578 that was derived by phage display against nerve growth factor (NGF) for the potential treatment of chronic pain[46]. Whist IgG-WFL had enhanced picomolar affinity for NGF, unlike the parent antibody MEDI578, IgG-WFL displayed poor biophysical characteristics endangering product development.

The mAb displayed significant colloidal instability (protein precipitation, phase separation and low solubility) and was found to interact with column matrices and adsorb to filter membranes which resulted in poor yields during purification[46]. Analysis of the protein using SAP[149] identified a hydrophobic patch on the protein surface (Figure 3.3a) that in combination with hydrogen/deuterium exchange mass spectrometry (HDX-MS) guided the rational design of the protein[46]. Three residues located in the CDRs of the V_H domain were identified as the problematic residues: W35, F36 and L64 (IMGT numbering[257,258]). Rational mutations were engineered to mutate the three residues back to those present in the parent antibody derived from phage display, W35S, F36T and L64T, generating MEDI1912_STT (referred to as IgG-STT herein)[46]. Despite the 99.6 % sequence similarity of the two IgGs (Figure 3.4), the rational engineering of IgG-STT reduced the non-specific interactions with the column matrix in HP-SEC (Figure 3.3b) and the mAb remained monomeric in solution (Figure 3.3c), without compromising the affinity to NGF.

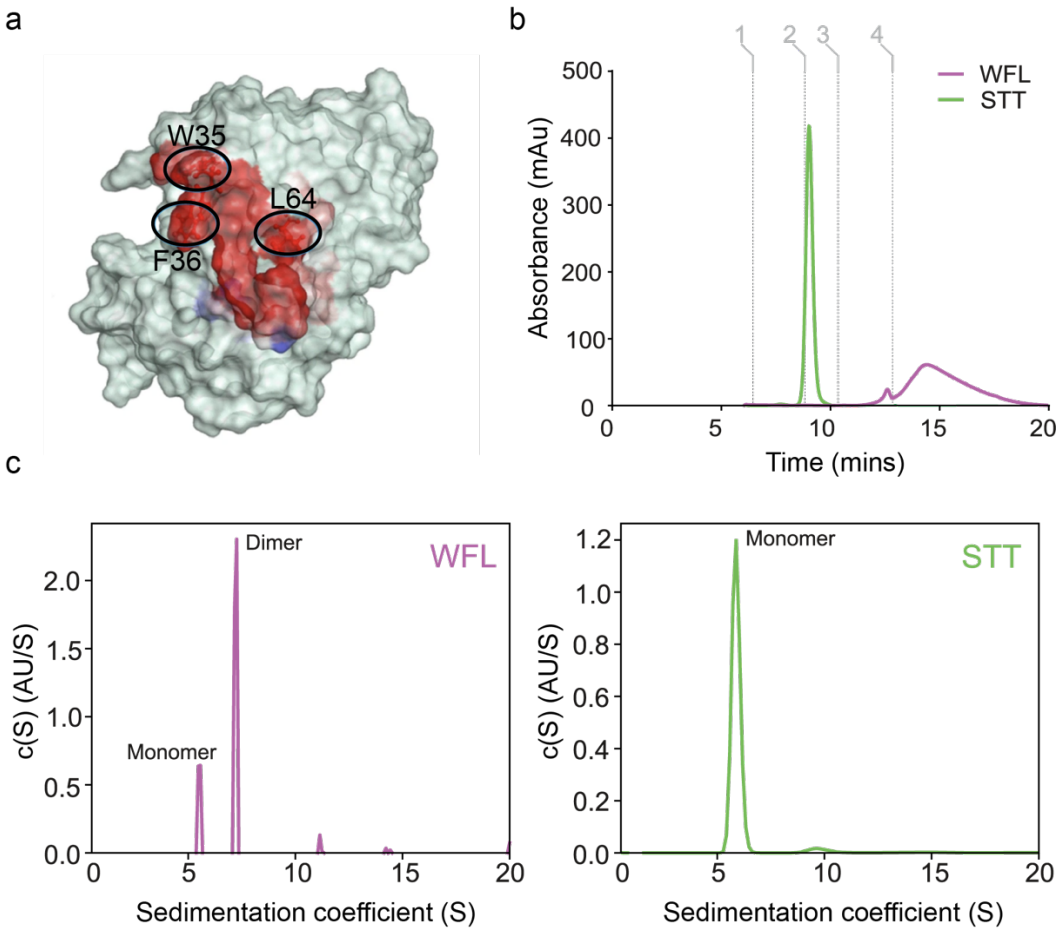


Figure 3.3 Biophysical properties of IgG-WFL and IgG-STT. a) SAP analysis of IgG-WFL revealed a surface exposed hydrophobic patch (red). The three residues W35, F36, L64 highlighted were mutated to STT respectively. b) HP-SEC elution profiles of IgG-WFL (purple) and IgG-STT (green). Grey lines = calibrant proteins: 1, Thyroglobulin (670 kDa); 2, IgG (158 kDa); 3, Ovalbumin (44 kDa) and 4, Vitamin B12 (1.35 kDa). c) AUC of IgG-WFL and IgG-STT. Data from Dobson *et al.*[46].

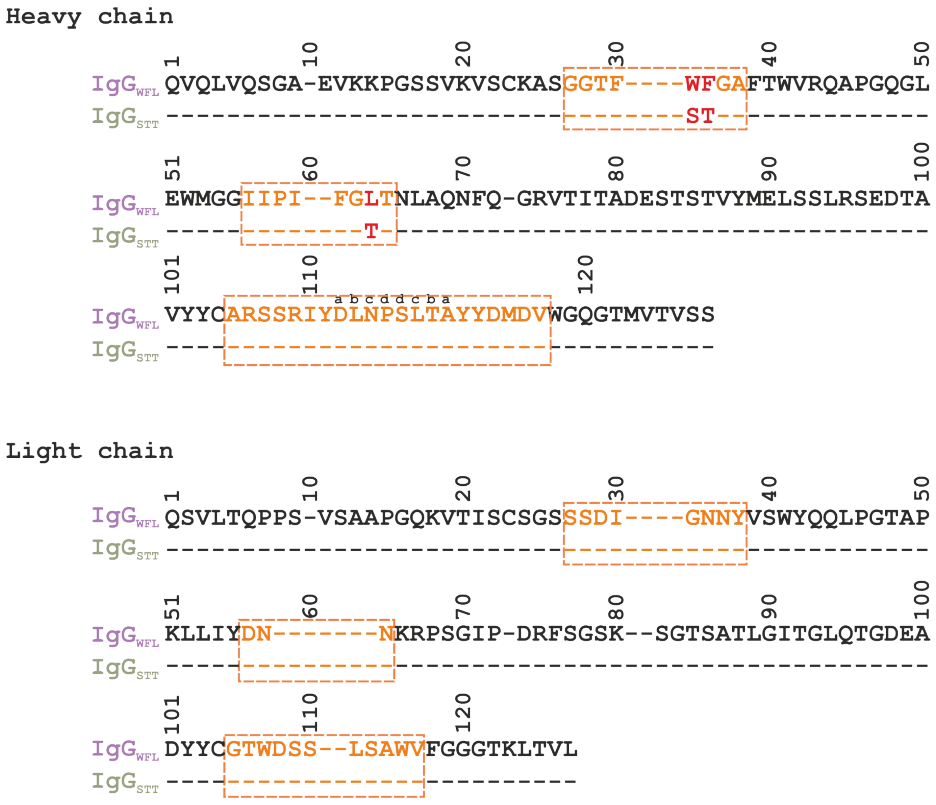


Figure 3.4 Sequence alignment of V_H and V_L domains of IgG-WFL and IgG-STT. Residues 35, 36 and 64 that differ between IgG-WFL and IgG-STT are highlighted in red. CDRs are highlighted in orange. Dash '-' represents conserved residues between the two sequences. Dashes within the IgG-WFL sequence denote IMGT numbering gaps (calculated using ANARCI server[258]).

3.3.2 *In vivo* screening of antibody fragments

Since both IgG-WFL and IgG-STT have been characterised in depth, these two molecules provide a suitable starting point to assess the application of the TPBLA to biopharmaceutical development. However, the quaternary structure and large size of mAbs (150 kDa) precludes their screening by the TPBLA. As the CDRs are often found to be the 'problematic regions' of IgGs, this limitation can be overcome by the use of single-chain Fv regions (Section 1.3.2.1 and Section 1.6.2.1). The variable domains of both IgGs were fused together by a Gly/Ser linker to create scFv-WFL and scFv-STT and were previously cloned into the β-lactamase construct[256], creating βla-scFv-WFL and βla-scFv-STT.

To ensure that the same trend is observed in SCS1 and BL21(DE3) *E. coli* (see Section 3.2) both βla-scFv-WFL and βla-scFv-STT were screened using the TPBLA. *E. coli*

expressing the aggregation prone scFv-WFL displayed a reduced ampicillin resistance compared to the aggregation-resistant scFv-STT (Figure 3.5a). The results also produced the same trend previously observed in BL21(DE3) cells (Figure 3.5b). The reduced total enzymatic activity of scFv-WFL presumably results from the aggregation of the scFv *in vivo* and hence the results from the TPBLA reflect the known differences in the aggregation behaviour of the IgGs previously characterised *in vitro*[46].

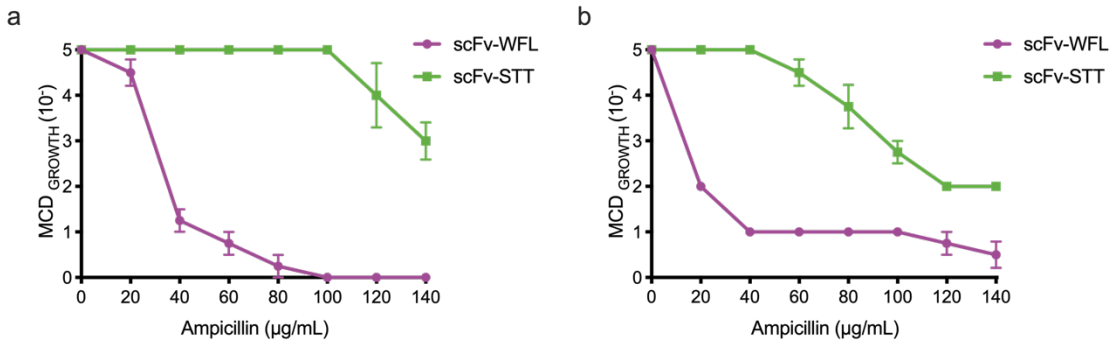


Figure 3.5 *In vivo* growth of scFv-WFL and scFv-STT. Antibiotic survival curve of the maximal cell dilution allowing growth (MCD_{GROWTH}) over a 0-140 µg/mL ampicillin concentration for a) SCS1 and b) BL21 (DE3) *E. coli*. Error bars represent s.e.m. of four independent experiments.

3.3.3 Applicability of the TPBLA to other biopharmaceutical protein scaffolds

To test whether the TPBLA could be used for a variety of applications within the biopharmaceutical sector two other protein pairs were selected with high and low propensities: Dp47d and HEL4 and, G-CSF and G-CSF C3.

3.3.4 Background to test proteins

Two single domain antibodies, Dp47d and HEL4, have previously been studied in the TPBLA to assess the use of the TPBLA to screen for excipients that stabilise/prevent the aggregation of the dAb[233,246]. These proteins were therefore employed again to compare the *in vivo* growth of scores of aggregation-prone therapeutically relevant scaffolds, and the data reproduced in SCS1 cells.

As described in Section 1.6.2.3, Dp47d is an aggregation prone human V_H domain, driven by the hydrophobic residues that reside on the original V_H:V_L interface. Dp47d was subject to directed evolution by phage display for binding to hen-egg lysozyme with elevated temperature conditions. HEL4 was isolated from this approach that contained mutations or insertions mainly to the CDRs of Dp47d (Figure 3.6)[259].

Analysis of the crystal structure of HEL4 revealed that there was a rotation of Trp52 (IMGT numbering) side chain into a cavity increasing the hydrophilicity of the V_H:V_L interface which is thought to be responsible for the aggregation-resistance of HEL4[259].

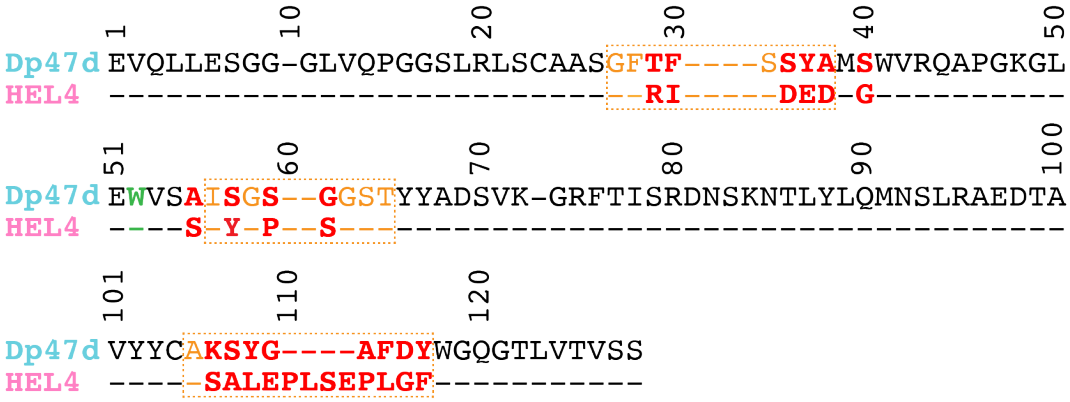


Figure 3.6 Sequence alignment of Dp47d and HEL4. Mutations or insertions are highlighted in red. CDRs are highlighted in orange. Dash '-' represents conserved residues between the two sequences. Dashes within the Dp47d sequence denote IMGT numbering gaps (calculated using ANARCI server[258]).Trp52 responsible for reduced aggregation is highlighted in green.

The final protein that was selected for aggregation analysis in the TPBLA was G-CSF. G-CSF is a 175-residue protein, predominantly involved in the response to infection[260] and was amongst the first cytokines to be identified and transitioned into clinical trials[261]. G-CSF stimulates the production of white blood cells (neutrophils) and is clinically used as a treatment for neutropenia in patients recovering from chemotherapy[262]. Structurally, G-CSF is an all α-helical protein, arranged in a four-helix bundle with an up-up-down-down arrangement (Figure 3.7)[263]. The α-helical architecture of G-CSF differs to the rich β-sheet Ig domains of the scFv previously studied in the TPBLA, allowing the broader applicability of the TPBLA to be assessed.

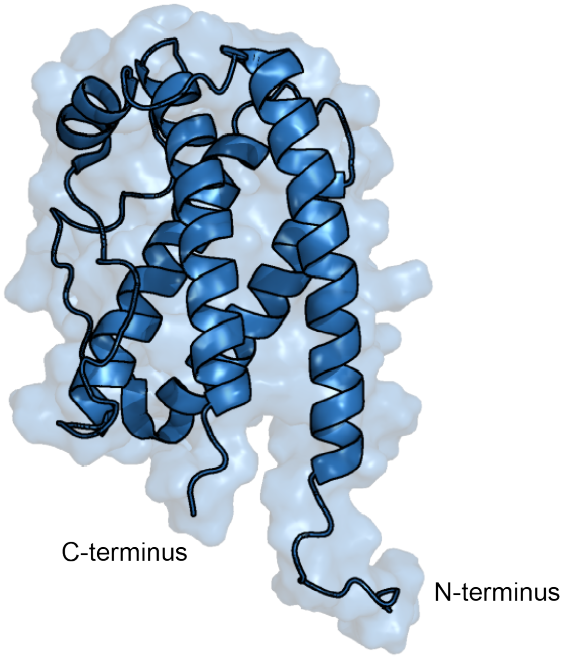


Figure 3.7 Structure of G-CSF.N-terminus and C-terminus of G-CSF are labelled. Figure created using PDB 1GNC[264] in PyMOL (Schrödinger).

Under native conditions (pH 7, 37 °C) G-CSF is aggregation prone[265–267] and has been found to be highly insoluble and has to be refolded from inclusion bodies, requiring a more complicated and expensive manufacturing process[268]. To counter this, Buchanan *et al.* used ribosome display to evolve G-CSF in conjunction with various selection pressures (reducing agent, elevated temperature and HIC matrices) to select for enhanced properties[269]. Outputs from the evolution panning were screened for enhanced soluble expression in the periplasm of *E. coli*. The results identified a variant called G-CSF C3 that contained four substitutions: C18G, W59R, Q71R and F84L (Figure 3.8)[269]. Overall this variant had 1000-fold enhanced expression (compared to WT), was primarily monomeric and retained functional activity[269].

```
                1         10        20        30        40        50
G-CSF      MTPLGPASSLPQSFLLKCLEQVRKIQGDGAALQEKLCATYKLCHPEELVL
G-CSF C3   ----------------G--------------------------------

               51        60        70        80        90       100
G-CSF      LGHSLGIPWAPLSSCPSQALQLAGCLSQLHSGLFLYQGLLQALEGISPEL
G-CSF C3   --------R-----------R------------L----------------

              101       110       120       130       140       150
G-CSF      GPTLDTLQLDVADFATTIWQQMEELGMAPALQPTQGAMPAFASAFQRRAG
G-CSF C3   --------------------------------------------------

              151       160       170
G-CSF      GVLVASHLQSFLEVSYRVLRHLAQP
G-CSF C3   -------------------------
```

Figure 3.8 Sequence alignment of G-CSF and G-CSF C3. Dashes represent the residues conserved between G-CSF and the evolved variant G-CSF C3. Residues that differ between the two proteins are highlighted in red.

3.3.4.1 Cloning of sequences into β-lactamase linker

G-CSF was PCR amplified with the addition of 5' *Xho*I site and a 3' *Bam*HI restriction site and cloned via restriction digestion into the 28-residue Gly/Ser rich linker that had previously been inserted between residues 196 and 197 of TEM-1 β-lactamase[242] (Figure 3.9 and Section 2.2). The newly synthesised construct, βla-GCSF, was then used as a template for Q5 site directed mutagenesis to consecutively introduce C18G, W59R, Q71R and F84L mutations to create the βla-GCSF-C3 variant (Section 2.2.6).

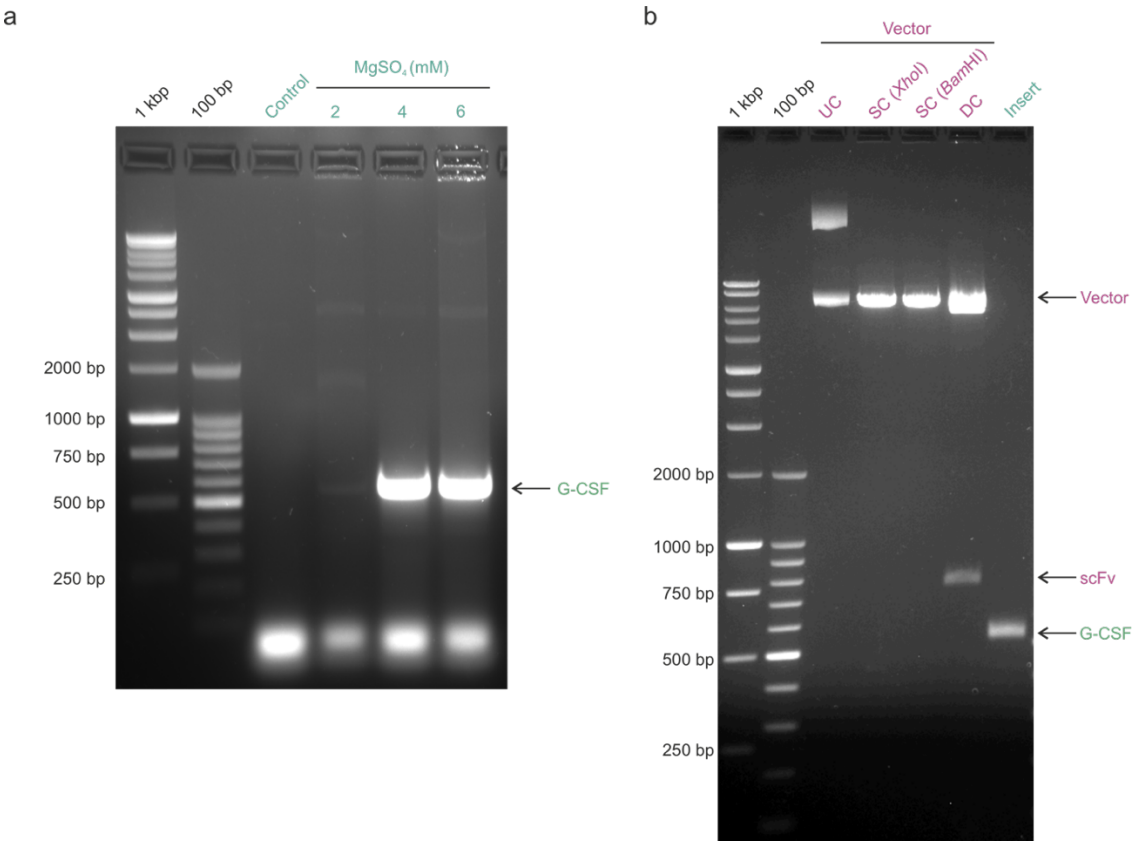


Figure 3.9 Cloning of G-CSF into β-lactamase linker. a) PCR amplification of G-CSF from the pET23a-GCSF plasmid with a range of $MgSO_4$ concentrations to identify optimal PCR conditions. Control reaction contained PCR components with no template DNA. b) Restriction digest reactions for pMB1-βla-scFv-WFL (vector) and G-CSF (insert). Vector digests include uncut (UC), single cut (SC) and double cut (DC) reactions with *Xho*I and *Bam*HI. The PCR product from (a) was digested with the same enzymes. The digested vector and insert were excised from the gel and ligated together.

3.3.4.2 Screening biopharmaceutical scaffolds

Both pairs of proteins were subsequently screened in the TPBLA over a 0-140 μg/mL ampicillin concentration range. The results show that the screen was able to differentiate between the aggregation-prone Dp47d versus the aggregation-resistant HEL4 (Figure 3.10a) and also discriminate between insoluble G-CSF against the evolved soluble variant G-CSF C3 (Figure 3.10b). For each protein the area under the curve can be calculated (Figure 3.10c) to produce a single value that represents the *in vivo* growth score of the *E. coli* expressing the construct (Figure 3.10d). For each of the biopharmaceutical scaffolds screened *in vivo* (scFv, dAb and cytokine) the engineered variant with low aggregation is significantly enhanced relative to its aggregation-prone counterpart.

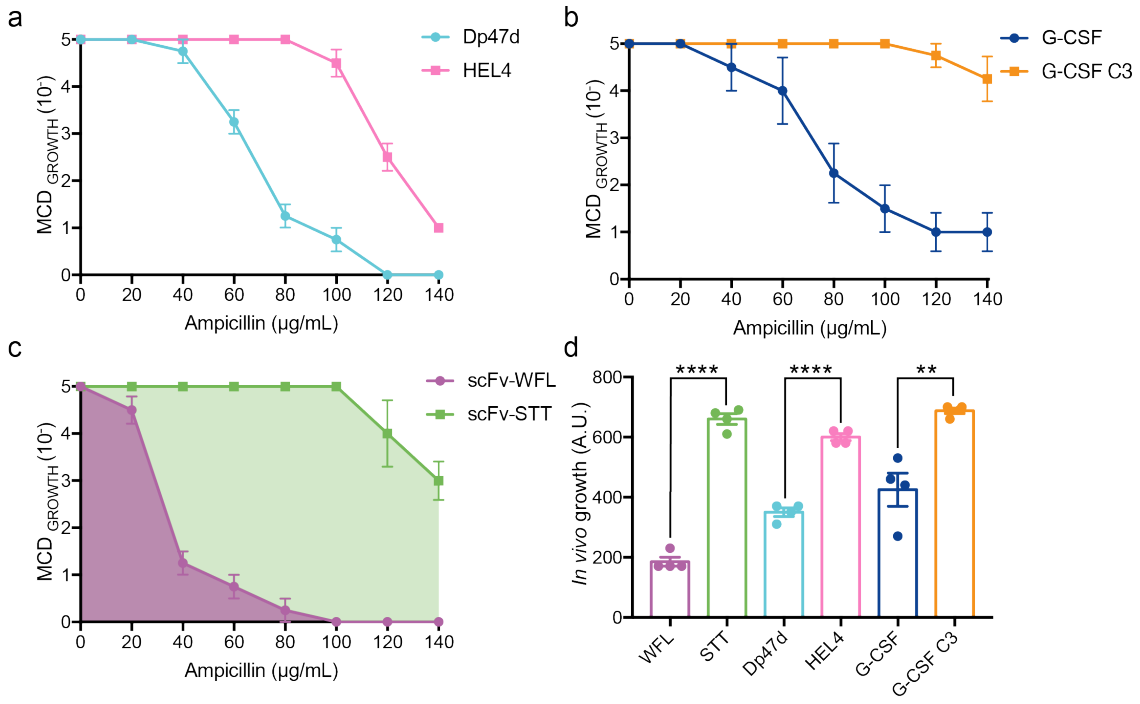


Figure 3.10 TPBLA screen for biopharmaceutical aggregation. MCD_{GROWTH} curves of a) Dp47d and HEL4 dAb and b) G-CSF and G-CSF C3 cytokine over 0-140 µg/mL ampicillin concentration range. c) The area under the curve is calculated (as denoted by the shaded purple (scFv-WFL) and green area (scFv-STT)) to generate d) a single value to compare related sequences. Data are shown for an aggregation prone scFv, dAb and cytokine and their engineered aggregation resistant counterparts. Error bars represent s.e.m. (n = 4 independent experiments). Asterisks denote significance: **** $p < 0.0001$, ** $p < 0.002$ (two-sided t-test).

3.3.5 Assay sensitivity

To employ the TPBLA for directed evolution, the assay needs to be sensitive to small changes in the protein sequence. To investigate this, the effect on the *in vivo* growth score of substituting W35, F36 or L64 (found in WFL) for S, T and T (respectively, found in STT) either individually or in combination was investigated (Figure 3.11). The TPBLA was able to distinguish between the subtle effects that the mutations have on the aggregation propensity of scFv-WFL, with the single point mutation W35S having the most profound effect on rescuing *E. coli* growth. These results demonstrate the sensitivity of the assay to mutations that induce aggregation.

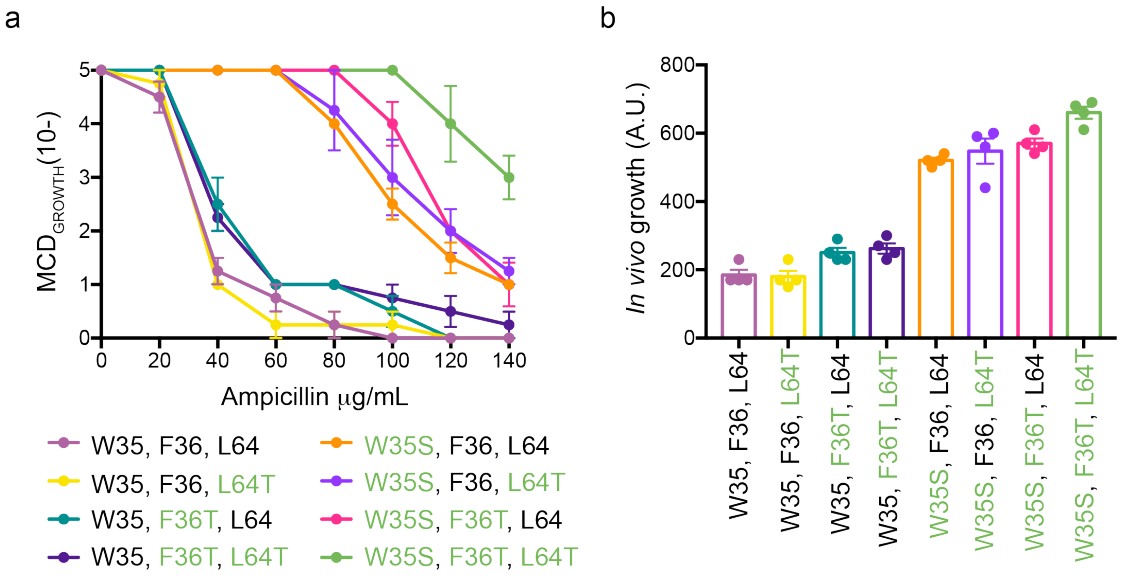


Figure 3.11 Effects of mutations on the *in vivo* growth of scFv-WFL. a) *In vivo* growth curves of βla-scFv variants containing WFL, STT, or one or two acid substitutions between WFL and STT. b) *In vivo* growth values for each construct. Green font indicates mutations away from WFL. Data reproduced from Devine 2016[256] in SCS1 cells. Error bars represent s.e.m (n = 4 independent experiments).

3.3.6 *In vivo* scFv aggregation correlates with IgG1 aggregation

To ensure that the results from the scFv fragments in the TPBLA are a true representation of *in vitro* IgG aggregation, each of the variants studied in section 3.3.5 were reformatted as full-length IgGs. Since it was known that IgG-WFL has non-specific interactions with the column that resulted in a longer retention time than expected for a monomeric IgG (Figure 3.3), the retention times of the six mutant IgGs were quantified and compared to the *in vivo* assay scores (Figure 3.12). Overlaying the IgG retention times for all eight variants with the *in vivo* scFv constructs identifies an excellent correlation between an improvement in bacterial growth and decrease in column retention time. Importantly, this demonstrates how the TPBLA can be utilised in an industrial setting to provide an insight on the aggregation propensity of lead candidates, without the need to express and purify the molecules as IgGs.

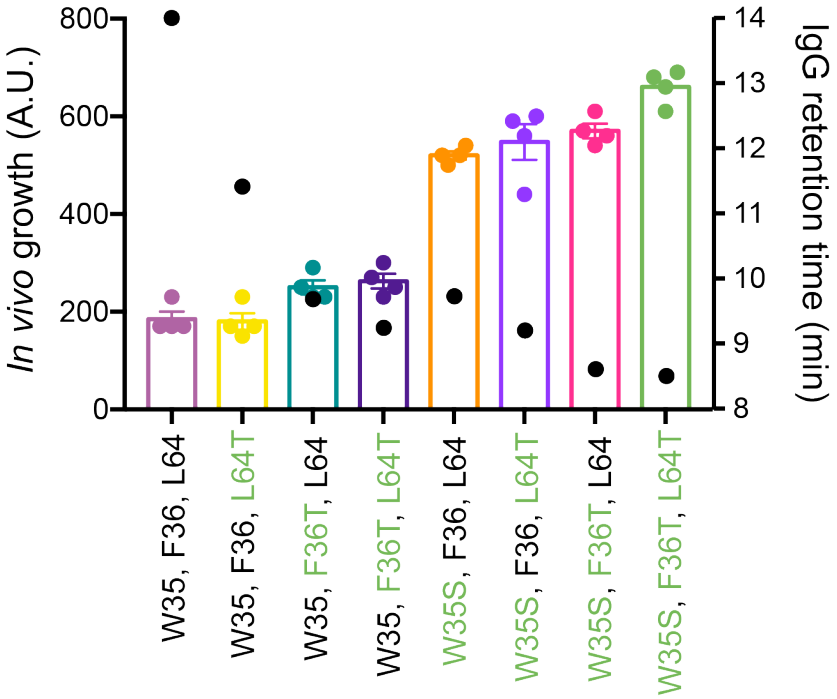


Figure 3.12 Comparison of sequence variants of scFv-WFL *in vivo* and IgG-WFL *in vitro*. Average *in vivo* growth score (bars) for scFv-WFL and scFv-STT with the six combinatorial variants. These data are overlaid with the HP-SEC retention times (black dots) for the same variants reformatted as an IgG1. HP-SEC data provided by Christopher Lloyd (AstraZeneca).

3.3.7 TPBLA is not a measure of protein expression levels

To identify if the results from the TPBLA are due to increased protein expression and therefore increased enzymatic activity of constructs, the expression levels and enzyme activity were quantified to compare the results to those obtained in the TPBLA.

Protein expression levels were measured using western blotting and dot blot analysis, whereby bacterial culture samples of each construct were taken during growth (as in Figure 3.2). Briefly, uninduced samples were taken from cells upon reaching an OD_{600} of 0.6. At this point, expression was induced with 0.075 % (w/v) arabinose for 1 h at 37 °C, 200 rpm following which an induced sample was taken. For all samples the OD_{600} was normalised to quantitatively compare the expression levels. Samples were then either separated by SDS-PAGE and transferred to a nitrocellulose membrane or applied directly to nitrocellulose membrane and expression was detected using a primary anti-β-lactamase antibody and secondary antibody to this conjugated to HRP which is then detected using chemiluminescence (Figure 3.13a and b).

Separation via western blotting (Figure 3.13a) detected multiple fragments of each β-lactamase construct suggesting that aggregation may result in degradation of the

fusion protein *in vivo*. The level of soluble cell protein expression detected by western blotting did not increase relative to the *in vivo* growth score of each construct (Figure 3.13a, left to right increasing *in vivo* growth score) suggesting that the bacterial resistance conferred by the tripartite fusion proteins does not solely report on protein solubility *in vivo*. Overall, protein expression levels differed for each of the constructs, however the expression did not correlate with their aggregation resistance in the TPBLA therefore suggesting that the screen does not simply report on protein expression levels (Figure 3.13c).

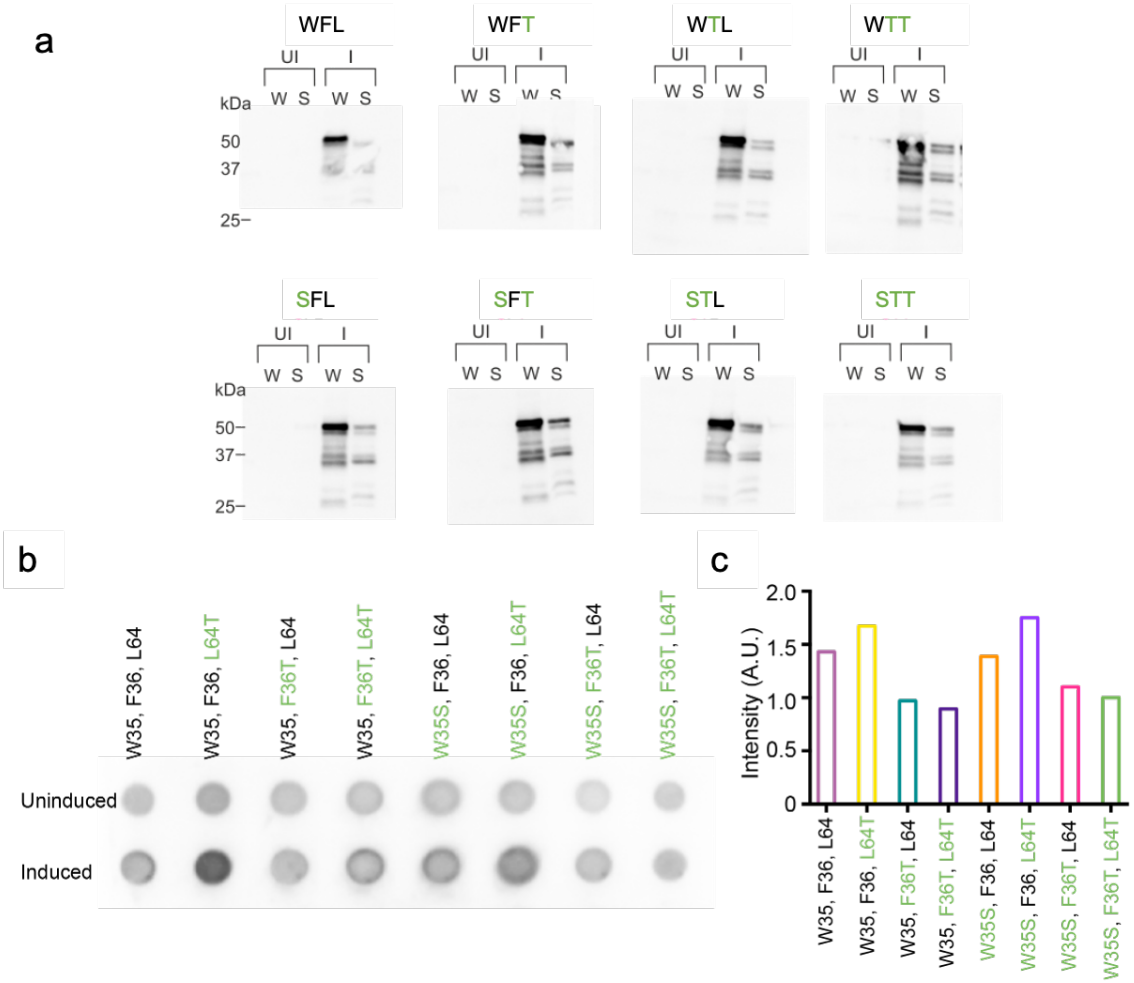


Figure 3.13 Expression levels of β-lactamase constructs. a) Western blot protein expression of Uninduced (UI) and Induced (I) whole (W) and soluble (S) samples and b) Dot blot protein expression levels detected by chemiluminescence signal of an anti-rabbit-HRP antibody to the rabbit anti-β-lactamase antibody. c) The intensity of the whole cell induced samples was quantified using ImageJ[270].

Enzymatic activity of β-lactamase was measured by utilising a chromogenic substrate, nitrocefin. Nitrocefin undergoes a distinct colour change from yellow (λ_{max} = 390 nm) to red (λ_{max} = 486 nm) when the amide bond in the β-lactam ring is hydrolysed. Cultures were grown and samples prepared as described for the dot blot.

Nitrocefin was added to the cultures and the absorbance at 486 nm was monitored over time (Figure 3.14a). The lag observed at the start of the measurements is due to the time taken for nitrocefin to diffuse into the periplasm where β-lactamase is expressed. The apparent initial velocity of the reaction was calculated by fitting a tangent to the line for each construct at the maximum gradient (Figure 3.14b). It was observed that each of the constructs had different levels of β-lactamase activity, however this did not correlate with the results from the TPBLA assay.

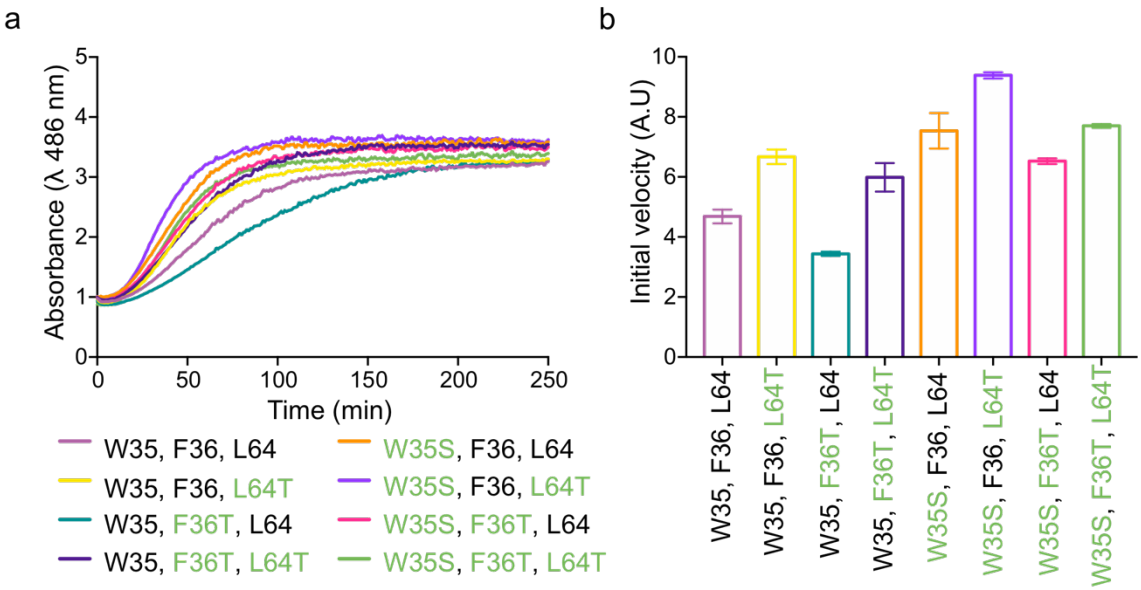


Figure 3.14 Enzyme activity of expressed β-lactamase constructs a) The hydrolysis of nitrocefin was monitored by measuring the increase in absorbance at 486 nm over time. b) The apparent initial velocity of the activity of each β-lactamase construct. Data collected from three technical repeats, error bars represent s.d.

A similar trend was observed between the enzymatic activity and the expression levels of β-lactamase present in *E. coli* (Figure 3.15). These results suggest that the initial expression level determines the activity of β-lactamase and as the activity correlates with protein levels, it suggests the specific activity is the same. However, since the TPBLA is employed under different conditions such as the range of antibiotic concentrations on solid medium and 18 h incubation (that allows for aggregation to occur) a different trend is observed, that may better infer the aggregation propensity of the POI, over measuring expression or enzyme activity.

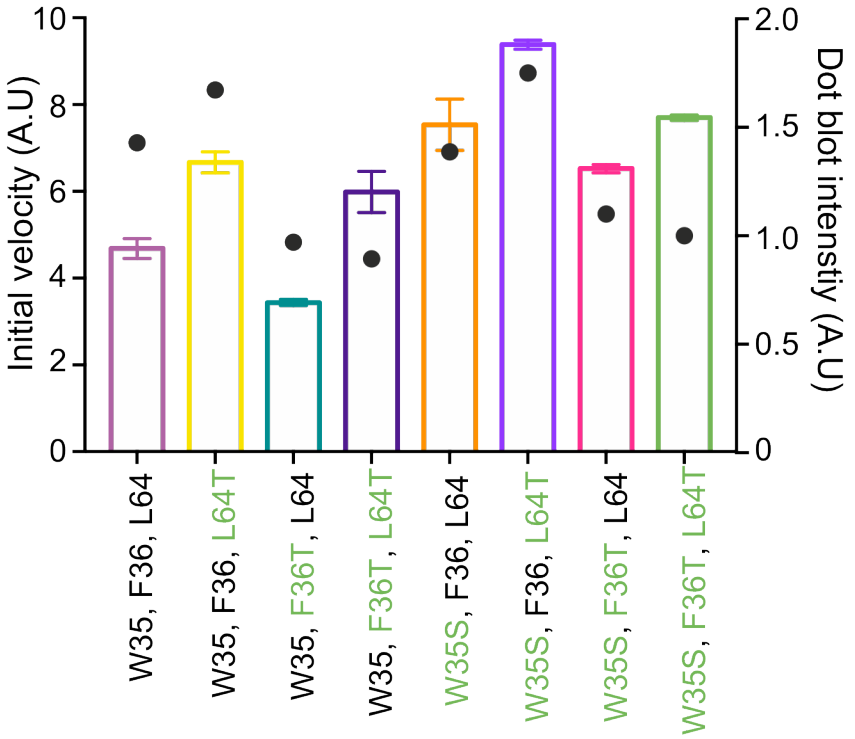


Figure 3.15 β-lactamase activity correlates with protein expression. The maximum gradient of enzyme activity (bars) is plotted on the left y axis and the protein expression determined from dot blot intensity (black dots) is plotted on the right y axis.

3.4 Development of directed evolution platform

As described in Section 1.6.2.3, *in vivo* screening systems have been employed for directed evolution to screen multiple variants by controlling the selection pressure. Having established that the TPBLA has the capability of detecting differences in aggregation that differ by just one amino acid, it was postulated that the TPBLA could be used for directed evolution to identify problematic residues within sequences.

3.4.1 Introducing diversity into scFv-WFL

The first step for directed evolution is to introduce genetic diversity into the gene to create a mutant library. The mutant library of scFv-WFL used in this study was created by Dr Stacey Chin (AstraZeneca). Firstly, genetic diversity was introduced into scFv-WFL by epPCR using primers that flanked the N- and C-terminus of the protein sequence. The PCR product was extracted and purified from the original template by agarose gel electrophoresis and then used as a 'megaprimer' in the second round of PCR. In the megaprimer PCR reaction, the βla-scFv-WFL template used contained two stop codons so that if any background plasmid remains, only the N-terminal half of β-lactamase will be translated. The megaprimer was used in a multi-site directed

mutagenesis reaction to introduce the mutations created during epPCR and to mutate the stop codons back to the WT sequence. The original template was then removed by *Dpn*I digestion, and the purified PCR product was transformed into TG1 cells, yielding a library of 1.3×10^6 colonies.

The DNA sequences of 57 variants in the naïve library were sequenced to evaluate the quality of the library. The analysis revealed that the mutations introduced were located throughout the protein sequence (Figure 3.16) and identified an average of eight amino acid substitutions per scFv.

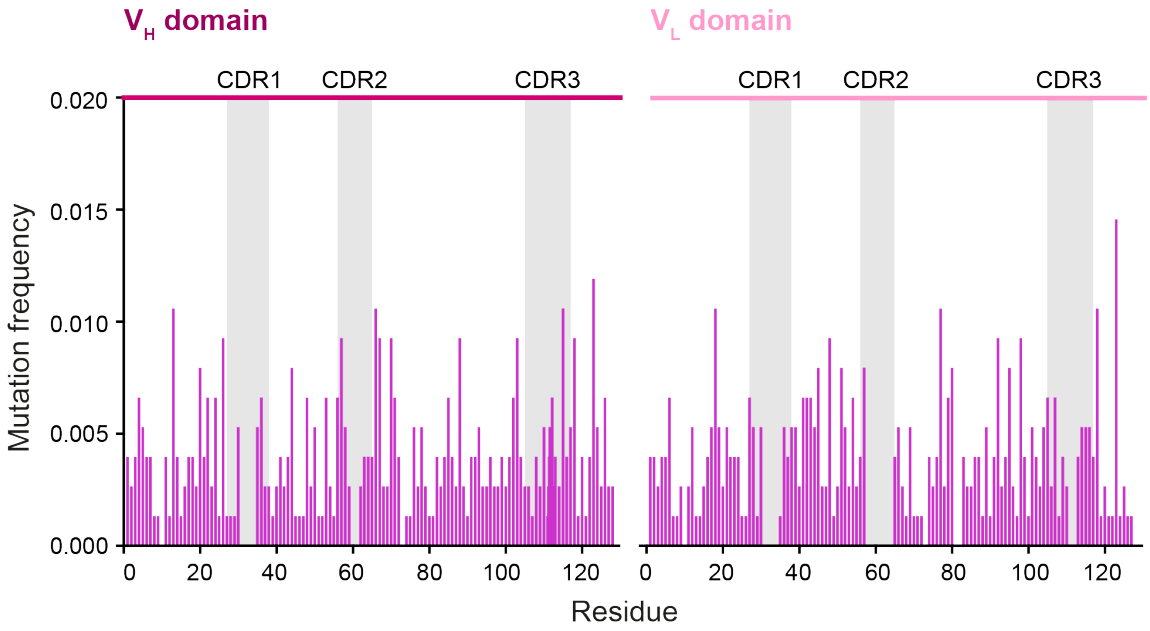


Figure 3.16 Mutational frequency of naïve library. Sequencing data of 57 clones from the scFv-WFL mutant library shows good sequence coverage with no bias at any particular residue. Variable heavy and light chain domains are labelled, and the grey boxes highlight the CDRs. Residues are numbered using IMGT numbering. Regions with no mutations in the CDR domains are due to the gaps introduced from IMGT numbering.

The mutations were further analysed at the DNA level to identify any potential mutational bias within the library. The library DNA sequences were compared to the WT scFv-WFL sequence and the total number of base mutations (e.g., A to C, A to G or A to T) were calculated (Figure 3.17). The results found that for each nucleotide (A, C, T or G) all three possible substitutions were incorporated within the sequences. Overall, this analysis suggests that there should be little bias within the library, making it possible for most amino acid mutations that are available within a single bp mutation from WT to be observed.

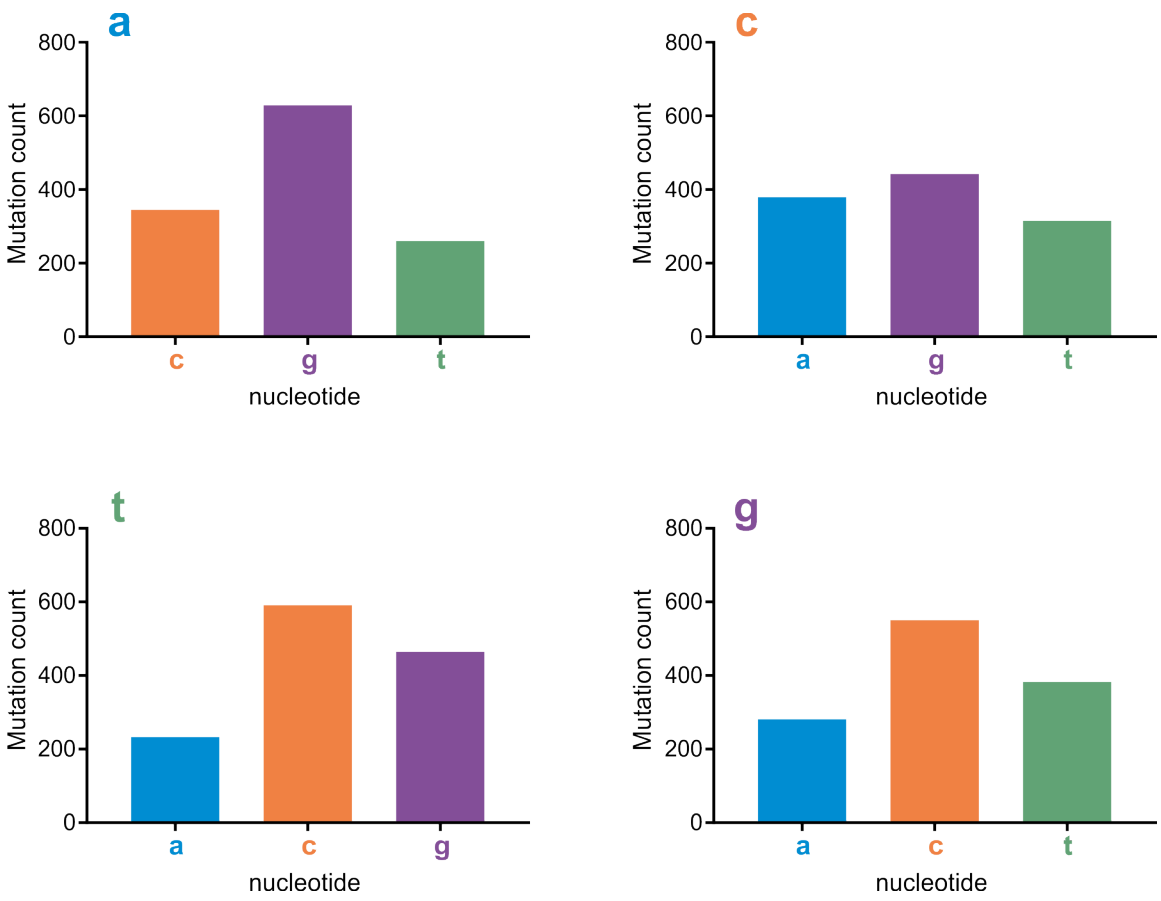


Figure 3.17 Analysis of codon bias within scFv-WFL library. Letter in the top left inset of each graph indicates the WT nucleotide and the bars indicate the mutation count for each substituted base. All three substitutions were incorporated for each of the nucleotide within the scFv-WFL library.

3.4.2 Identification of aggregation hotspots

The principle for screening in this assay is that mutants within the library should be able to grow at higher concentrations of ampicillin than scFv-WFL if they contain mutations that have decreased the aggregation propensity of the protein (Figure 3.18). The plasmid DNA library of variants was transformed into SCS1 *E. coli* cells and following a recovery period SOC medium containing tetracycline was added to the cells and incubated at 37 °C, 200 rpm for 4 h. β-lactamase expression was induced with 0.075 % (*w*/*v*) arabinose for 1 h at 37 °C, 200 rpm. Cells were then plated onto agar containing 80 μg/mL ampicillin, a concentration that severely restricts *E. coli* growth when expressing scFv-WFL (refer to Figure 3.5). The colonies are then sent for sequencing to analyse the mutations that endow enhanced antibiotic resistance.

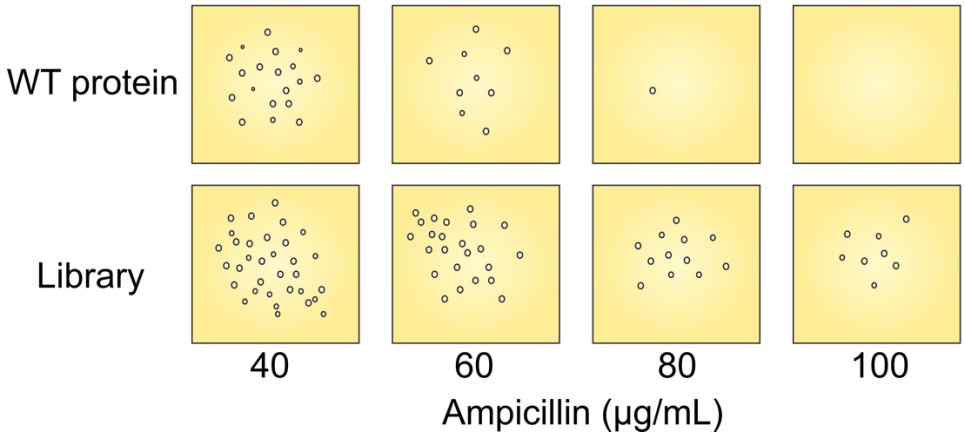


Figure 3.18 Principe of directed evolution screening. Library mutants with enhanced properties, such as improved stability or reduced aggregation propensity, can grow at higher antibiotic concentrations in comparison to the WT protein.

The number of times each residue is found to be substituted in the screened library can be used to create mutational frequency histograms. This identifies those residues most frequently mutated, generating a protein-specific profile of frustration that can be used to begin to unpick the molecular mechanism(s) of aggregation.

To assess the ability of this approach to identify the problematic sequences within IgG-WFL, the mutational frequency profile across the V_H and V_L domains of the library was constructed from the sequences of 315 variants screened at 80 μg/mL ampicillin (Figure 3.19). Twelve 'hotspot' residues (nine in V_H and three in V_L) were identified that occurred at a frequency of two standard deviations above the average mutational frequency (0.0126, Figure 3.19 and Figure 3.20). Seven of these hotspot residues (W35, I59, F30, F36, I57, F62 and I56), which are all hydrophobic or aromatic, formed two clusters centred on CDR 1 and 2. The remaining hotspot residues in V_H (I110 and L112c) which are also hydrophobic, form a third cluster around CDR3. By contrast, two of the three hotspots in the V_L domain target a charged (K18) or a hydrophilic (N57) residue and two of the three (K18 and I71) reside in the framework region, suggesting either these residues are neutral mutations or may reflect an additional stabilisation mechanism.

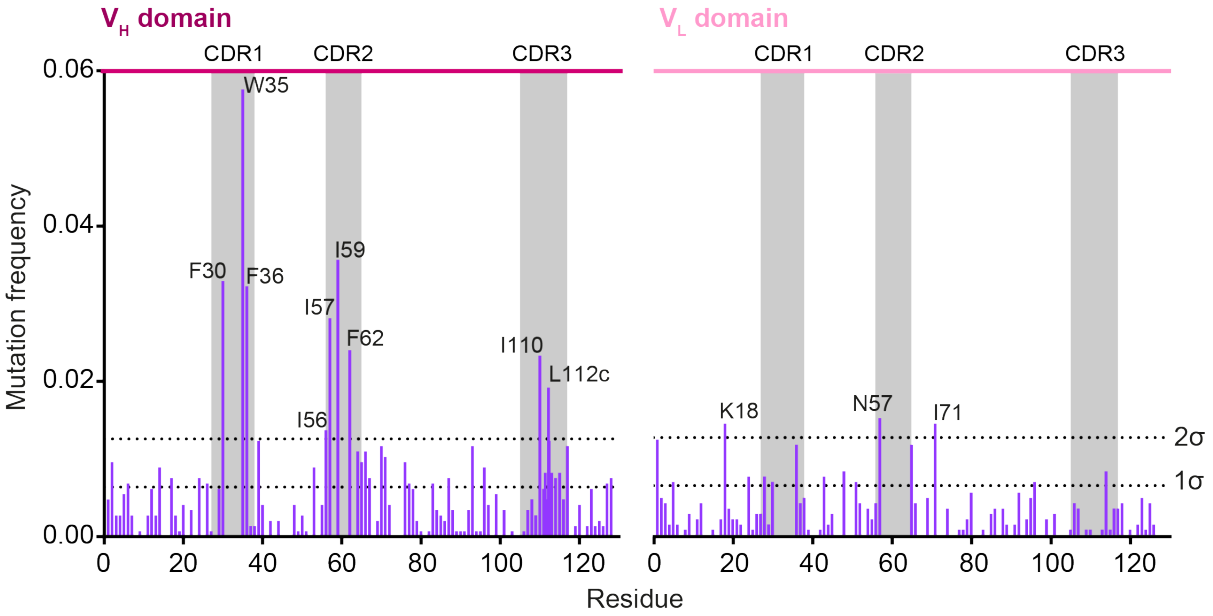


Figure 3.19 Mutation frequency profile of evolved scFv-WFL. Directed evolution reveals twelve hotspot residues with a mutational frequency greater than two standard deviations from the average value (2σ). Variable heavy and light chain domains are labelled, and the grey boxes highlight the CDRs. Residues are numbered using IMGT numbering. The location of the hotspot residues on the 3D structure of scFv-WFL are shown in Figure 3.20.

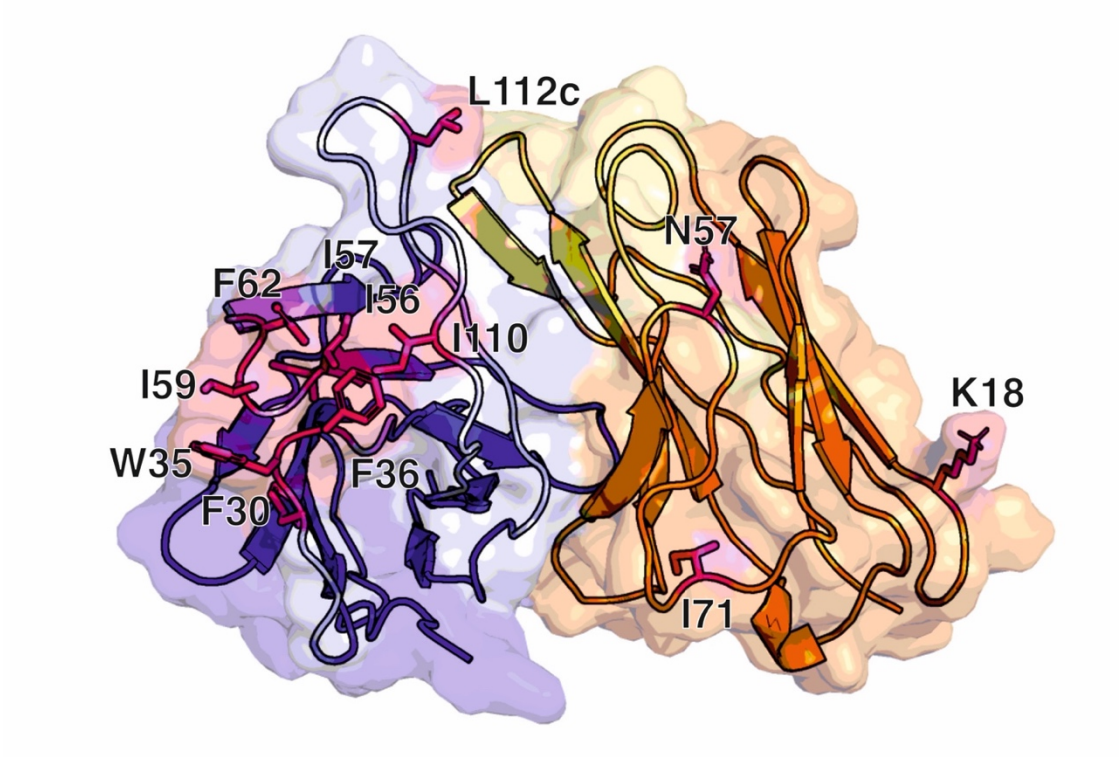


Figure 3.20 Mutational hotspots of scFv-WFL. Structural location of the mutational hotspots identified from evolution (identified in Figure 3.19). V_H domain is shown in blue, V_L domain is shown in orange, CDRs are shown in light blue and yellow for the V_H and V_L respectively. Hotspot residues are labelled and shown in pink.

The chemical identity of the most frequently selected residue and whether a particular amino acid residue is enriched relative to the other residues accessible via a single-base pair change was also assessed (Table 3.1). In general, these hydrophobic and aromatic residues (most with relatively high solvent exposure) located in CDRs 1-3 were substituted with more hydrophilic residues. The hotspot residues in the V_L domain which were initially charged (K18), hydrophilic (N57) or hydrophobic (I71) were substituted with polar or other charged amino acids.

Residue	RSA[a]	Most frequently observed aa substitution	Mutation frequency of most often observed mutation	Amino acid substitutions observed[b,c]	Available residues with single DNA base change[c]
F30	0.07	S	0.81	SL**P**V	IVLFCSY
W35	0.63	R	0.93	RG	LCGSR
F36	0.66	S	0.60	SL(V**P**)(I**T**)	IVLFCSY
I56	0.09	V	0.50	VT(LF)	IVLFMTSN
I57	0.01	T	0.66	TNV**A**	IVLFMTSN
I59	0.44	T	0.60	TNVF	IVLFMTSN
F62	0.45	S	0.60	SLY	IVLFCSY
I110	0.19	T	0.82	TV(LFM)	IVLFMTSN
L112c	0.65	P	1.00	P	IVLFPHR
$K18_{VL}$	0.81	E	0.52	ERNQ	ITEQNKR
$N57_{VL}$	0.18	D	0.73	DS**G**	ITSYHDNK
$I71_{VL}$	0.24	T	0.62	TVN	IVLFMTSN

Table 3.1 Summary of the twelve most frequently substituted residues after directed evolution of scFv-WFL. [a]RSA = relative surface area (0 = completely buried residue, 1 = maximally solvent exposed residue). [b]Residues are listed in decreasing mutational frequency with brackets indicating residues with equal frequency of mutation. Substitutions shown in bold are due to two base-pair changes in the DNA codon. [c]Amino acids are listed in decreasing hydrophobicity (left to right) using the Kyte-Doolittle scale[271].

3.4.3 Comparison of mutational hotspots to *in silico* predictions

Computational algorithms have been developed to predict insoluble or aggregation hotspots in protein sequences and have been valuable for reducing the extensive *in vitro* screening of a large range of mutant variants[18]. No single algorithm has been shown to be superior to others and each focus on a different underlying factor such as stability, solubility or APRs as discussed in Section1.5.1.

Since computational predictions have been successful at detecting APRs, the overlap between three commonly employed *in silico* methods (CamSol, SAP and Aggrescan3D) and the hotspot residues identified from the TPBLA evolution platform were examined.

Briefly, CamSol measures the hydrophobicity, electrostatic charges, and interplay spatial patterning to identify predicted insoluble residues. SAP is a rational and simulation-based technique that is based on amino acid hydrophobicity, surface exposure and the contributions of other amino acids within a preassigned radius. Aggrescan3D energetically minimises the structure using FoldX, following which the aggregation is predicted by calculating the intrinsic aggregation propensity, solvent accessibility, and the effective distance between adjacent residues.

Comparison of the location of the hotspots on a structural model of scFv-WFL identified using the TPBLA with CamSol, Aggrescan3D and SAP are shown in Figure 3.21. Globally, the mutations predicted from each computational method and the *in vivo* platform appear to overlap, with the main problematic region forming a large patch on the surface of the V_H domain of the scFv. This patch correlates with the hydrophobic interface for aggregation previously identified for IgG-WFL by Dobson *et al*[46] (Figure 3.3).

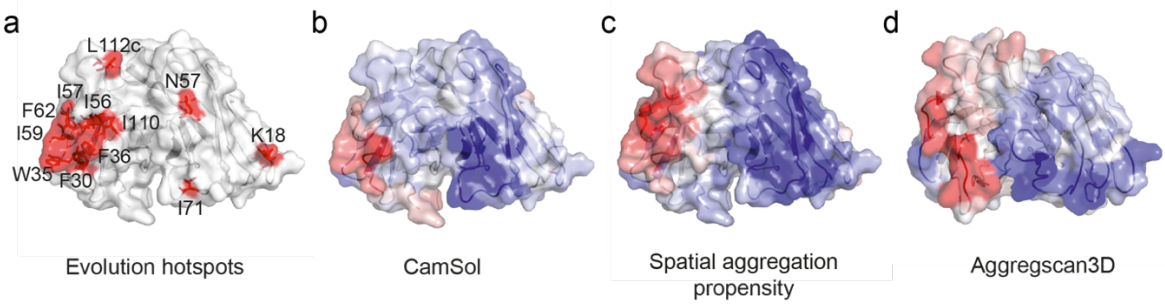


Figure 3.21 Structural location of aggregation prone residues. Comparison of a) evolution hotspots (residues >2 σ), b) structurally corrected CamSol[144], c) SAP[149] and d) Aggrescan3D[141] for scFv-WFL. Insoluble/aggregation prone residues are shown in red and soluble/non-aggregation prone residues are shown in blue on a surface model of the protein (created from PDB 5JZ7[46]).

The hotspot locations identified by the TPBLA and *in silico* methods were further compared by viewing their position on the primary sequence of scFv-WFL (Figure 3.22 and Table 3.2). Cut offs were applied to each algorithm: for CamSol <−1 indicates insoluble and >1 soluble; for SAP >0.5 is aggregation prone <−0.5 is non-aggregation prone; for Aggrescan3D >1 is aggregation-prone and <−1 is non-aggregation prone

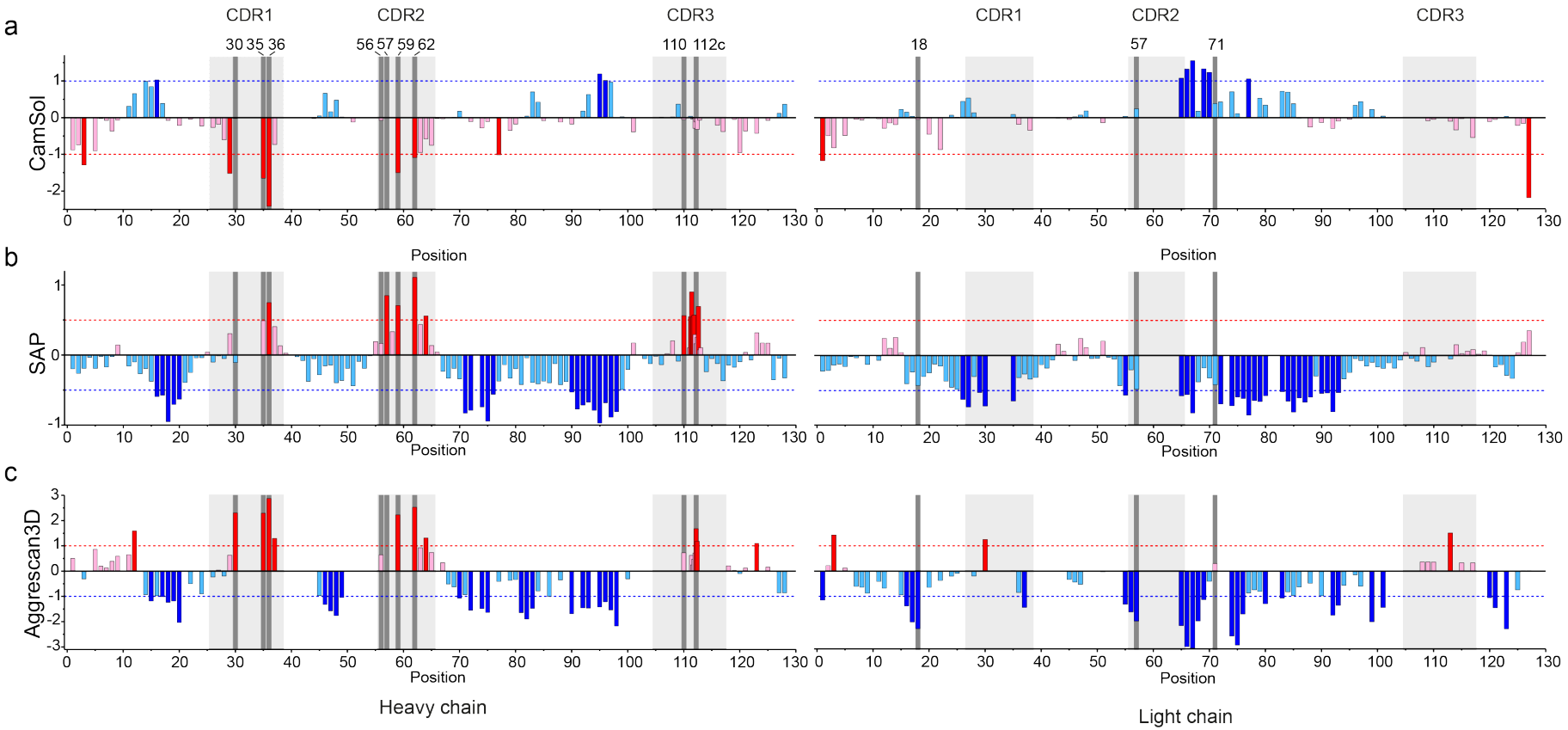


Figure 3.22 Comparison of *in silico* predictors of aggregation with the evolved mutational hotspots for scFv-WFL. The individual residue scores for a) CamSol (<−1 indicates insoluble and >1 soluble), b) SAP (>0.5 is aggregation prone <−0.5 is non-aggregation prone) and c) Aggrescan3D (>1 is aggregation-prone and <−1 is non-aggregation prone). Red bars highlight insoluble/aggregation prone residues and blue bars represent soluble/non-aggregating residues. In each plot the significance values are highlighted by dotted lines. Dark grey vertical bars denote evolution hotspot residues and light grey boxes highlight CDRs. Residues are numbered according to IMGT.

It can be seen that every algorithm detected at least one residue in CDR1 and CDR2 (that forms the large patch shown in Figure 3.21), but the identity of the residues varied between algorithms. SAP, and to a lesser extent Aggrescan3D, also identified the third TPBLA hotspot-cluster in CDR3 (Figure 3.22). In contrast to this agreement, each *in silico* method highlighted additional residues in the V_H not identified by the TPBLA and no *in silico* method detected the residues identified by directed evolution in the V_L domain.

In total, the three algorithms flagged 26 residues as potential positions that could be the cause of the aggregation of scFv-WFL (Table 3.2). This included eight of the twelve hotspots identified by the TPBLA, however, only three of these residues are highlighted by all three computational algorithms (Figure 3.23).

Residue	Computational tool		
	CamSol	SAP	Aggrescan3D
F30	-0.203	-0.107	2.30
W35	-1.65	0.490	2.23
F36	-2.41	0.745	2.87
I56	-0.074	0.165	0.641
I57	0	0.192	0
I59	-1.50	0.71	0.64
F62	-1.087	1.10	2.52
I110	-0.053	0.56	0.722
L112c	-0.116	0.161	1.67
K18	-0.046	-0.429	-2.28
N57	-1.98	-0.480	0.236
I71	0.382	-0.419	0.299

Table 3.2 Analysis of hotspot residues using CamSol, SAP and Aggrescan3D. The numerical score is colour coded based on arbitrary cut-off values for each algorithm in line with Figure 3.22. Structurally corrected CamSol, +1 indicates soluble and −1 indicates insoluble; SAP (using a 10 Å radius) where values >0.5 and <−0.5 are significant; and Aggrescan3D, where values > 1 and <−1 are significant.

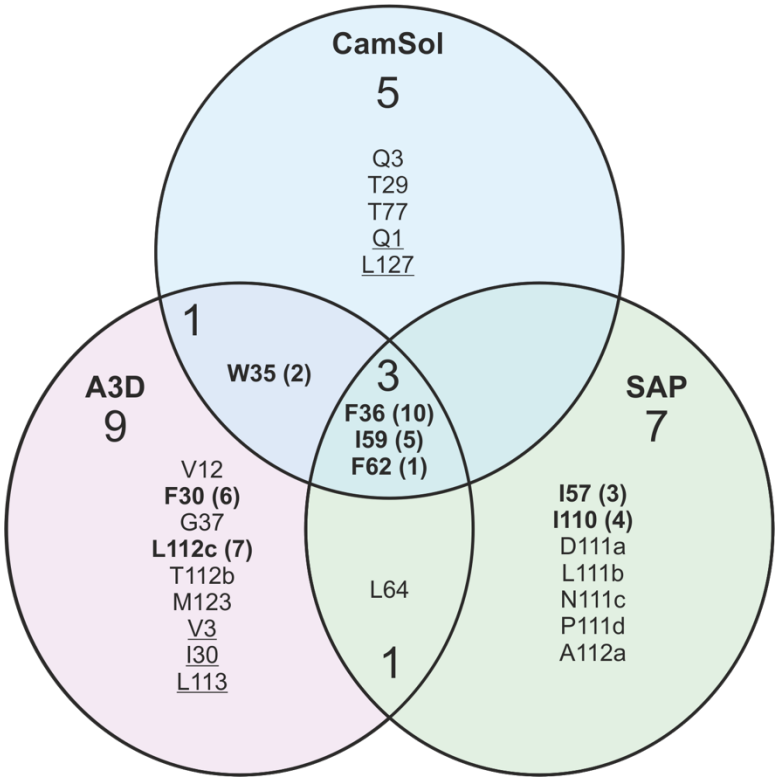


Figure 3.23 Comparison of aggregation prone/insoluble residues identified by Camsol, Aggrescan3D and SAP. Residues were identified as aggregation prone/insoluble from CamSol (output score < -1), Aggrescan3D (> 1) or SAP (> 0.5). V_L domain residues are underlined. Residues in bold are those identified as hotspots using the TPBLA. The number in brackets corresponds to the rank in aggregation propensity based on the *in vivo* growth score when introduced as single substitutions (see next section).

3.4.4 Screening hotspot mutations

While the amino-acid changes for the residues in the mutational hotspots identified in section 3.4.2 have been postulated to reduce the aggregation of scFv-WFL, the relative importance of each substitution remains unclear. To examine the relative significance of these substitutions and how they relate to *in silico* predictions, single-residue substitution variants of βla-scFv-WFL were created by site-directed mutagenesis to introduce the most common amino acid substitution for each of the twelve aggregation hotspot residues (Table 3.1). The βla-scFv-WFL hotspot variants were then subsequently screened using the TPBLA to determine a rank order of the hotspot residue mutations (Figure 3.24).

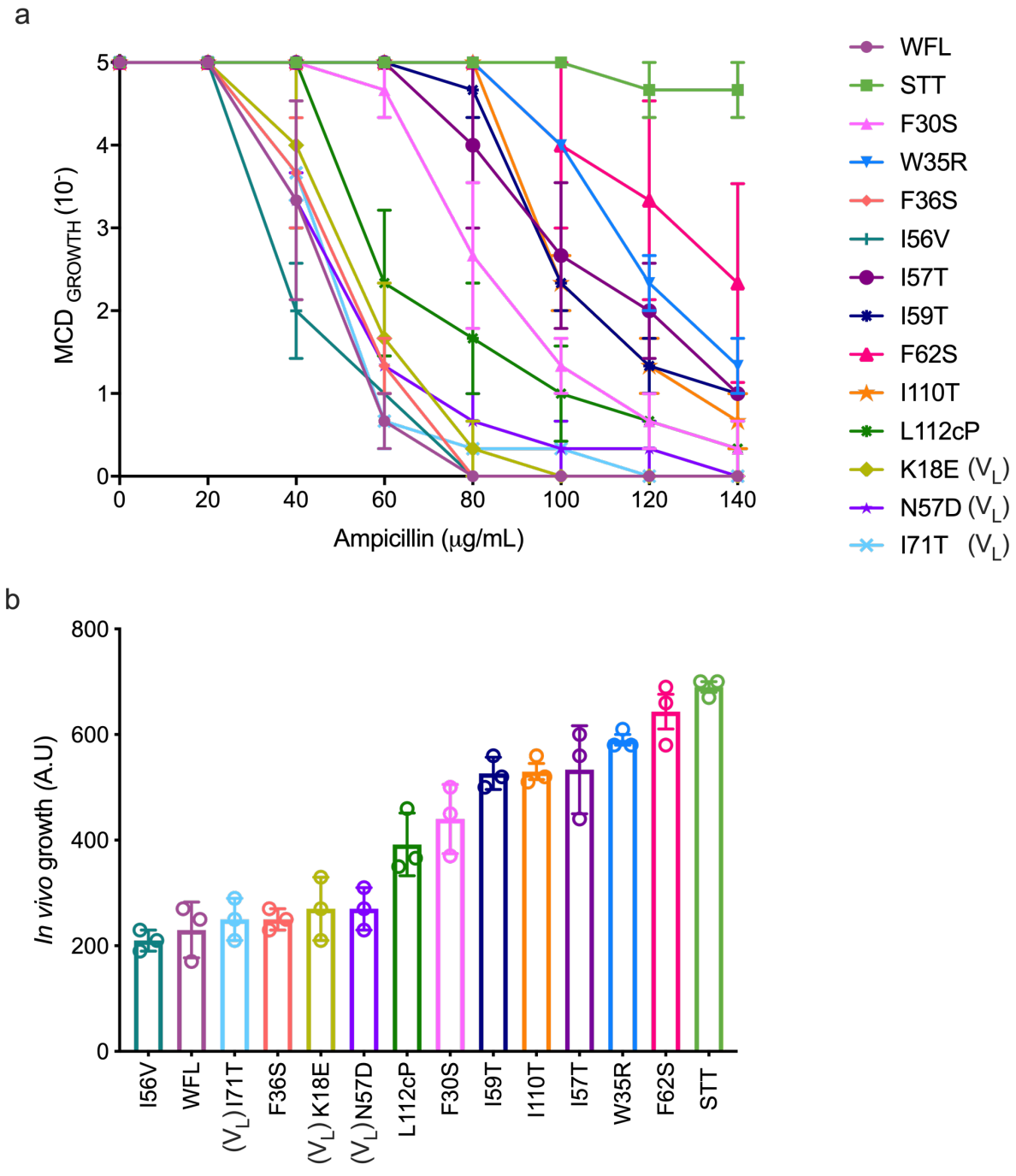


Figure 3.24 *In vivo* growth of the twelve hotspot residues in scFv-WFL. The most frequently observed mutation at each hotspot was introduced into scFv-WFL as identified in Table 3.1. a) Growth survival curves for each variant. b) *In vivo* growth scores for each variant as calculated from the data in a. n = 3, error bars represent s.e.m.

These single substitutions showed two main types of behaviour. The substitutions (in increasing *in vivo* growth score) L112cP, F30S, I59T, I110T, I57T, W35R and F62S yielded significantly enhanced growth scores relative to scFv-WFL. By contrast,

residues I56V, I71T (V_L), F36S, K18E (V_L) and N76D show little effect in isolation. The small enhancement in antibiotic survival observed for the latter group in isolation, suggests that these are evolutionary neutral or that these sites may act synergistically with others.

Although no single substitution was found to match the *in vivo* growth score for scFv-STT (690 ± 8 A.U.), F62S was identified as the highest scoring point mutation, achieving 91% of this growth enhancement (643 ± 27 A.U.). Interestingly F62 was not previously identified as a problematic residue for the rational re-design of scFv-WFL (see section 3.3.1). In light of this observation, it is also notable that though W35R (mutated to S in STT) had the second greatest growth score (590 ± 14 A.U.), F36 (mutated to T in S**T**T) had little effect in isolation (scores for WFL and F36S were 172 ± 53 and 250 ± 9 A.U., respectively).

Ranking these variants by *in vivo* growth score does not improve the correlation with the residues predicted to be problematic by *in silico* methods. For example, only three residues (F36, I59 and F62) are flagged by all three computational methods, yet these vary considerably in the effect (ranked 10th, 5th, and 1st, respectively (Figure 3.23)). It should be noted, however, that the TPBLA cannot identify the optimal residue (and hence largest growth score) for each hotspot, due to the limitations of epPCR for library synthesis (Section 1.6.2.3).

One way to overcome this caveat could be to implement the computational algorithms following the identification of mutational hotspots to perform site saturation mutagenesis *in silico* to identify the optimal residue to introduce at this position. To investigate this possibility, residue F62 (that upon mutation to F62S has the highest *in vivo* growth) was mutated to each of the 19 amino acids *in silico*. These variants were screened using CamSol, Aggrescan3D and FoldX to predict the change in solubility, aggregation, and stability relative to WT scFv-WFL. This *in silico* screening identified that F62S introduced through evolution is predicted to enhance the protein solubility (Figure 3.25a), but has no effect on the aggregation propensity as determined by Aggrescan3D (Figure 3.25b), and may cause destabilisation of the protein (Figure 3.25c). The majority of the 19 mutations introduced enhanced the solubility as determined by CamSol, but in doing so, most mutations reduced the stability of the protein, suggesting that this position has a fine balance between stability and solubility[272]. The site saturation *in silico* screening would suggest that introducing residues D, N, K or R may have the greatest overall enhancement in solubility and reducing aggregation (Figure 3.25). Interestingly, these four residues were not available from one DNA base change and hence not observed during evolution (see Table 3.1).

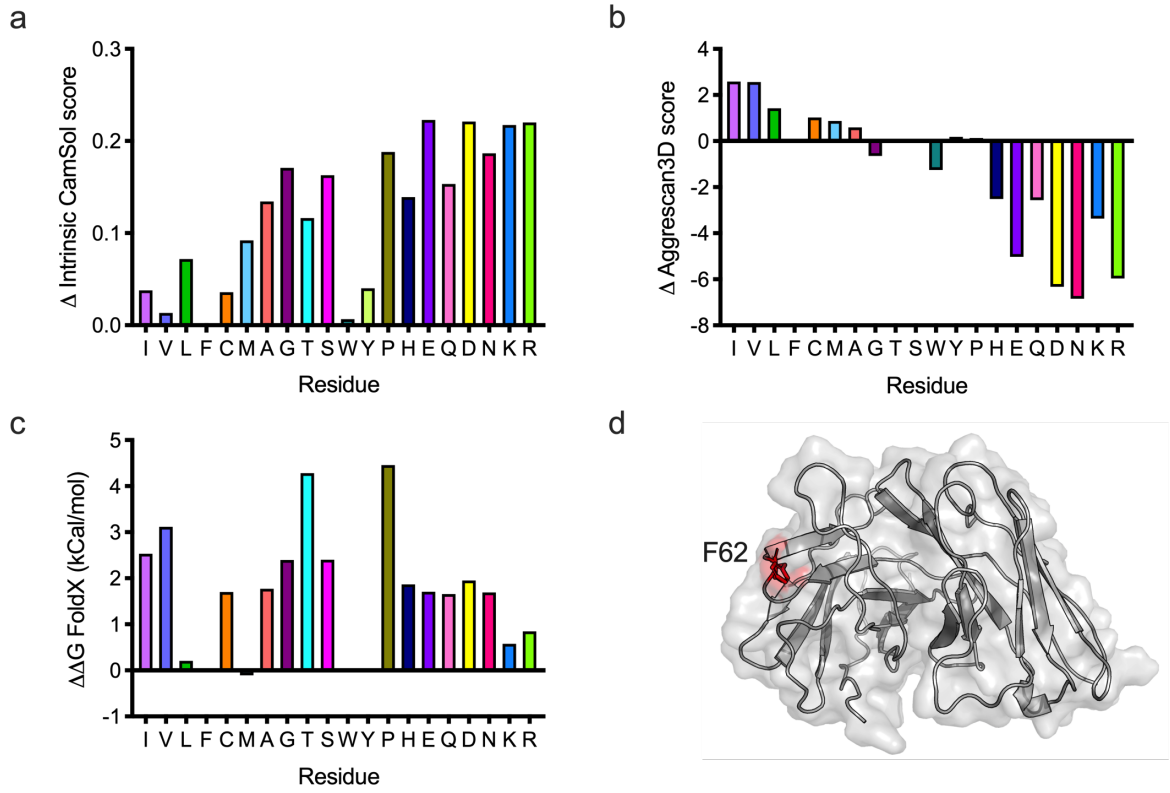


Figure 3.25 *In silico* site saturation mutagenesis of F62. Residue F62 in scFv-WFL was mutated to every other 19 amino acids *in silico* and was screened by a panel of computational techniques. a) CamSol prediction of the overall intrinsic solubility of the protein (positive ΔCamSol score = improved solubility relative to WT). b) Aggrescan3D prediction of aggregation propensity of the protein (negative ΔAggrescan3D = reduced aggregation). c) FoldX stability prediction (positive ΔΔG = reduced stability). d) Structural location of F62 on scFv-WFL model (made using PDB 5JZ7[46]).

3.5 Discussion

Studying protein aggregation is a time consuming and costly challenge during the development of biopharmaceuticals, hence there is a drive to identify problematic sequences as early as possible during development. Identification of aggregating sequences can be a challenging due to the APRs being buried in the protein structure and unfolding may be required for aggregation to occur. Aggregation may only be detected during the later stages of development due to the manufacturing stresses the protein endures that increase the risk of protein misfolding and aggregation, compromising the quality and safety of a drug product.

Despite the range of techniques employed during lead isolation and optimisation (discussed in Section 1.5) biologic developability remains a significant obstacle to the delivery of drug candidates to the clinic[98]. It is often necessary to purify a large number of highly avid and specific candidates to identify which to take forward for further development. The work presented in this chapter presented shows how the TPBLA, that directly links aggregation-propensity to a phenotypic read out of antibiotic resistance, could be employed following binding selection to filter aggregation-prone candidates.

The TPBLA has many advantages over current methods employed to investigate protein aggregation. Firstly, it removes the need to purify often difficult to handle proteins prior to analysis, facilitating the screening large numbers of variants. In contrast to other *in vivo* systems for studying protein aggregation[226,273,274], the fusion proteins are expressed in the oxidative periplasm of *E. coli*, allowing the correct formation of the disulfide bonds found in IgGs and their derivatives. Most importantly, no perturbant such as increased temperature, pH or chemical denaturant is used to accelerate aggregation.

The work highlights the broad applicability of the TPBLA to distinguish between aggregation-prone and aggregation-resistant sequences by its application to three pairs of different protein scaffolds, with different mechanisms of aggregation. The TPBLA requires no structural or functional knowledge and is therefore useful for early stages of development when there is no prior knowledge of the POI.

As the biopharmaceutical sector is dominated by mAbs[41], this chapter (and subsequent work in this thesis), further focused on how the TPBLA can applied for the use of IgGs. To ensure the assay was sensitive to study mutational variants that differ by just small changes in sequence, single- and double- mutations of scFv-WFL through to scFv-STT were investigated. The results showed that W35S is largely responsible for the aggregation resistance of scFv-STT. Importantly, the results of

each scFv variant produced a striking correlation with the aggregation of the full-length IgG, variant as determined by the retention time from HP-SEC.

In addition to screening for aggregation-prone or aggregation-resistant sequences the TPBLA was developed as an evolution platform to identify aggregation prone residues within scFv-WFL. The decision to perform the evolution platform on solid agar instead of in liquid culture was reached from the results obtained from measuring the expression levels and enzyme activity in liquid culture. This identified that the expression levels of the construct correlated with the enzymatic activity of β-lactamase, but not with the trend observed from the TPBLA, and hence liquid screening may not provide a suitable indication of IgG aggregation.

The library created by epPCR was estimated to have 1.3×10^6 mutants, this is however expected to be lower as it is unlikely that every colony contained a unique clone. The validity of the library used is therefore unknown without sequencing every colony. A further limitation to library synthesis by epPCR is that only a small sample of mutations to each residue are accessible with one or two base pair mutations to the codon whilst maintaining a low mutation rate. To overcome both of these limitations a deep mutational scanning approach could be employed whereby a library is generated with site-saturation mutagenesis throughout the sequence in combination with next generation sequencing[275–277].

The results from screening a library of mutants at a higher concentration of ampicillin that scFv-WFL could grow at identified 315 colonies. The sequencing of these colonies revealed mutational hotspots within the protein sequence, that are presumably the residues that contribute to the aggregation of in scFv-WFL. Unsurprisingly, given their importance in determining epitope binding affinity, the majority of hotspot residues identified are located in, or close to, the CDRs. The crystal structure of the IgG-WFL parental antibody (MEDI578), in complex with its ligand[46] shows that the CDR3 is most important for binding, with 13 residues out of a total 22 making contacts to NGF, whereas only 6 of the 16 in CDR1 and CDR2 have any direct interaction with NGF. Since the largest mutational hotspots occur within CDR1 and CDR2 it is possible that the binding affinity for NGF may be retained, which is investigated in Chapter 4.

In addition to the presence and location of the hotspots, analysis of the substitutions made may also be informative to begin to understand what characteristic drives selection. Here, for example, ten of the twelve hotspot residues were hydrophobic/aromatic in nature, and all were substituted with more hydrophilic residues, consistent with the mechanism for aggregation being driven by the hydrophobic patch on the protein surface[46]. In the future, this directed evolution

platform may allow for the rational design of inherently manufacturable IgGs, by identifying target residues to mutate to reduce the aggregation potential.

Although *in silico* tools are currently used to guide rational design, this chapter shows that these algorithms yield different predictions, making the identification of the key residues to target by rational engineering difficult using a multi-algorithm approach, highlighting the power of using evolution to find solutions to the problem of aggregation.

To counter the disagreement of the computational algorithms, machine learning could be employed to develop a novel *in silico* screening method to identify a novel metric for developability. For this, a large dataset of negative and positive selections would need to be collected that comprises of a diverse range of IgG1 sequences. Furthermore, to develop a valuable dataset for machine learning, it is key to have a suitable platform that screens the property for which the algorithm is to be developed for. The work in this chapter demonstrates that the TPBLA would be successful for this approach.

In summary, the data presented in this chapter demonstrates the broad utility of the TPBLA for screening protein aggregation early during biopharmaceutical development. The strong correlation between *in vivo* aggregation propensity as an scFv and the *in vitro* aggregation propensity as an IgG validates the use of the TPBLA for screening a large range of mutational variants before candidate selection for IgG expression and purification. Moreover, development of a directed evolution platform allows the residues that cause aggregation to be pinpointed (and substituted), which could be used to guide rational design experiments or could be used to improve or design computational algorithms.

Chapter 4
Evolution of aggregation resistant antibodies

4.1 Objectives

In Chapter 3, a mutant library was utilised to identify evolved sequences of scFv-WFL that could grow under an increased selection pressure. This determined a sequence-aggregation profile of scFv-WFL, that highlighted the residues predicted to be aggregation prone. The next aim of this thesis is to investigate the behaviour of the evolved scFv-WFL sequences *in vivo* and to characterise the extent to which the evolved variants have improved the biophysical properties of IgG-WFL *in vitro*. This chapter also aims to investigate the general applicability of the platform for the evolution of aggregation resistant antibodies, through evolution of a second IgG with a different mechanism of aggregation.

4.2 Directed evolution of biopharmaceuticals

The first law of directed evolution "you get what you screen for" is an important consideration for establishing a screen for selection that reflects the desired result from the experiment[278]. Previously, as discussed in Section 1.6.2.3, most directed evolution screens for biopharmaceutical development have focused on evolving enhanced solubility[228], increased thermal stability[279] or heat- or acid- induced aggregation[229,231]. These approaches usually employ arbitrary methods to destabilise the protein, that may not reflect the inherent partial or full unfolding of the protein[280]. Moreover, as no singular property drives biopharmaceutical aggregation (Section 1.4), the lack of a suitable screen has prevented the optimisation of biopharmaceuticals for resistance to innate aggregation by directed evolution until now.

Irrespective of the evolution method employed, identified variants need to be purified for *in vitro* analysis to ensure the enhanced property has indeed been selected for. One advantage of the *in vivo* platform presented in Chapter 3, is that the sequences identified from the initial directed evolution experiment can be further assayed using the TPBLA in the 48-well plate format. The results from this secondary screen allow a greater insight into the aggregation resistance of the clones identified from the numerical score calculated for each variant in contrast to the simple live/dead screen. Ultimately this may be useful at screening hits from evolution and reducing the number of candidates to take forward for purification and *in vitro* biophysical analysis as an IgG.

4.3 Screening evolved sequences *in vivo*

Screening of the mutant library in Chapter 3 identified 315 colonies that could grow under the 80 μg/mL ampicillin selection pressure. 185 of these variants were selected at random and subjected to a full *in vivo* growth assay (Section 2.3) to verify that selection at 80 μg/mL ampicillin did engender greater antibiotic resistance for each individual clone. To allow the sensitivity of the evolved variants to be compared, the area under the curve was calculated to produce a rank order of the mutant scFv-WFL clones. This approach allows both the best variant to be identified, which in an industrial setting could be taken forward for further development and, by comparing aggregation-phenotypes across the rank, to begin to assess the relationship between sequence and self-association.

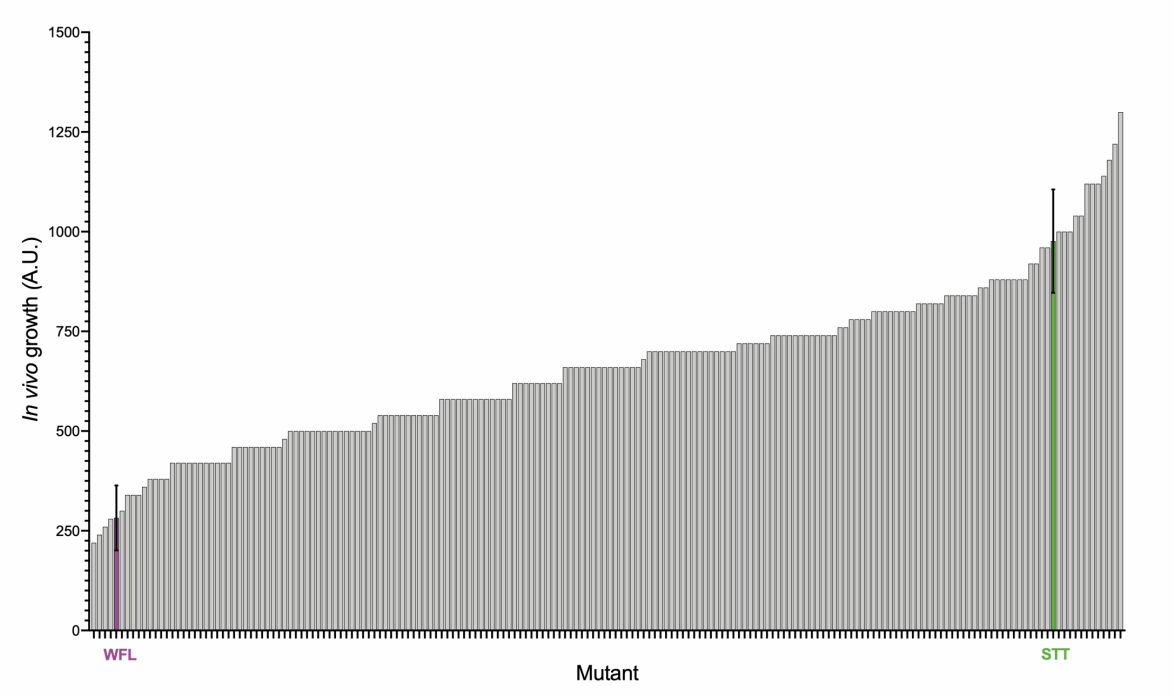


Figure 4.1 Full TPBLA assay of scFv-WFL variants. Ranked *in vivo* growth score of 185 evolved variants (grey bars). βla-scFv-WFL (purple) and βla-scFv-STT (green) error bars represent s.d. n = 16 biological repeats.

The *in vivo* growth assay showed that 181 of the 185 variants had enhanced growth relative to βla-scFv-WFL, with twelve having superior growth to the rationally engineered aggregation-resistant STT. Analysis of the sequences with enhanced growth over scFv-STT is highlighted in Table 4.1. Of note, every sequence contains a mutation of at least one of the twelve hotspot residues in Chapter 3, further validating that indeed these residues are important for the improved aggregation resistance of scFv-WFL mutants.

Interestingly, clone 132 contains only hotspot mutations (F36L (V_H), I56T (V_H), I110T (V_H), K18N (V_L)). As identified in Chapter 3, mutations to residues I56, F36 and K18 did not significantly enhance the *in vivo* growth of scFv-WFL (although this this may be dependent on the residue substitution), and I110T mutation enhanced the antibiotic resistance of scFv-WFL from 172.5 ± 53.2 to 530 ± 15.3 A.U. Remarkably, as identified here, the combination of mutations to these four residues enhanced the *in vivo* growth to 1120 A.U., suggesting possible synergy between these residues. Accordingly, F36S and I110T are present together along with two other mutations, to produce the best screened clone identified by the TPBLA suggesting that these mutations may indeed have an additive effect.

Clone ID	*In vivo* growth (A.U.)	Mutations
130	1000	T29A (V_H), I57T (V_H), L64H (V_H)
147	1000	T29A (V_H), W35R (V_H), T125A (V_L)
163	1000	V12A (V_H), I57T (V_H), K51R (V_L), N57S (V_L), S69P (V_L)
96	1040	S26G (V_H), W35G (V_H)
105	1040	Q1R (V_H), S26G (V_H), T29A (V_H), W35R (V_H), F62S (V_H), Q6P (V_L), N65S (V_L)
16	1120	F36S (V_H), F62S (V_H), V87A (V_H), F118I (V_L)
66	1120	V2A (V_H), K14R (V_H), F36S (V_H), Q48R (V_H), L64P (V_H), I78F (V_H)
134	1120	F30S (V_H), G55S (V_H), L64P (V_H)
132	1140	F36L (V_H), I56T (V_H), I110T (V_H), K18N (V_L)
114	1180	F30S (V_H), I57N (V_H), S2G (V_L), T92A (V_L)
140	1220	V5A (V_H), K20A (V_H), F30S (V_H), G49A (V_H), F62S (V_H), L111bS (V_H), N111cI (V_H), V11A (V_L), N36D (V_L), S114G (V_L)
139	1300	S26G (V_H), F36S (V_H), I110T (V_H), S86T (V_L)

Table 4.1 Sequence analysis of evolved mutants that outperformed scFv-STT in the TPBLA. Residues in purple are those identified in WFL evolution hotspots. Residues underlined are W35, F36 or L64 from WFL.

4.3.1 *In silico* screening in comparison to TPBLA

It has previously been suggested *in silico* screening could be employed during lead candidate selection, using algorithms such as CamSol or Aggrescan, to rapidly screen mutant library sequences to rank protein variants according to their solubility or aggregation propensity[140,144–146]. Therefore, the intrinsic solubility (CamSol) and the amino-acid aggregation propensity value sequence average (a^3vSA, Aggrescan) of each variant screened using the TPBLA was calculated and compared to their *in vivo* growth score (Figure 4.2). The analysis identified a poor correlation between the value obtained from TPBLA with both CamSol (Figure 4.2a, $R^2 = 0.08$) and Aggrescan (Figure 4.2b, $R^2 = 0.04$). This poor correlation is potentially due to the nature of the TPBLA to screen for multiple limiting factors, such as solubility, aggregation, stability, and expression, which may not affect the calculation of the *in silico* CamSol or Aggrescan score. In contrast to this, both CamSol and Aggrescan showed a higher correlation relative to that observed for each tool with the TPBLA (Figure 4.2c, $R^2 = 0.325$).

Despite the lack of correlation between the TPBLA and the *in silico* methods, it was observed that many of the evolved sequences have enhanced solubility and reduced aggregation. As determined by CamSol, 180 sequences had improved solubility over scFv-WFL, and 48 of these variants displayed a higher intrinsic solubility score than scFv-STT (higher score = increased solubility). Likewise, 58 sequences had an a^3vSA value less than or equal to scFv-STT (negative value = reduced aggregation). Interestingly, the 4 clones that had *in vivo* growth lower than scFv-WFL in the TPBLA had enhanced solubility and reduced aggregation *in silico* and thus their poor *in vivo* growth may be due to an overriding destabilising factor. Taken together, this analysis identifies that TPBLA may be a better alternative to *in silico* screening for lead candidate selection due to its ability to screen for multiple factors and hence therefore may be a better representation of the *in vitro* biophysical properties of IgGs.

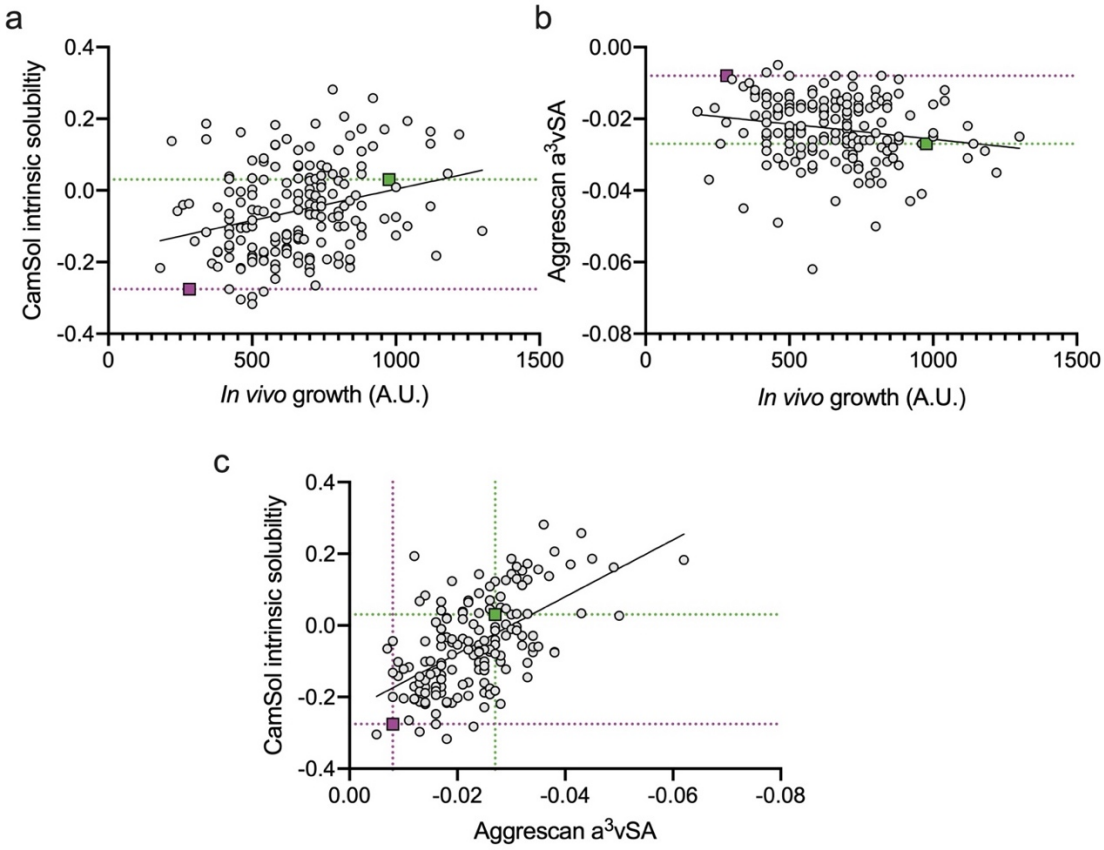


Figure 4.2 *In silico* screening of 185 scFv-WFL variants. Protein sequences were screened and compared to TPBLA values. a) CamSol intrinsic solubility scores (higher value = enhanced solubility) plotted versus *in vivo* growth score. Linear regression R^2 = 0.08. b) Aggrescan amino acid aggregation propensity value sequence average (a^3vSA) (lower value = lower aggregation propensity) plotted versus *in vivo* growth score. Linear regression R^2 = 0.04. c) CamSol intrinsic solubility plotted versus Aggrescan amino acid aggregation propensity. Linear regression R^2 = 0.325. For all graphs, purple square denotes scFv-WFL value, green square represents scFv-STT, and grey dots indicate mutational variants. The dashed lines highlight the *in silico* value for scFv-WFL (purple) and scFv-STT (green) for each computational tool.

4.3.2 Identification of variants for biophysical analysis

To further investigate whether the *in vivo* growth score for the evolved scFv variants also correlates with reduced aggregation propensity within an IgG1 scaffold, ten variants that spanned the rank order were selected to convert to IgG1 molecules. To identify a range of sequences to take forward, the standard deviation of the replicate error of βla-scFv-STT (n = 16, s.d. = 130) was utilised. Starting with the best performing clone, 139, further variants were selected sequentially across the rank by

selecting the next variant with an *in vivo* growth score separated by one standard deviation (130 A.U.) from this value.

If for this value there was a range of sequences that have the same *in vivo* growth score, the variant with the fewest substitutions (relative to WFL) was selected. This identified 8 variants: 11, 176, 59, 72, 126, 130, 16 and 139. Two further sequences (37 and 128) were selected to increase the number of sequences to study that retain the original WFL residues (W35, F36 and L64) yet had improved *in vivo* growth score. The identity of each of the substitutions for these variants are shown in Table 4.2.

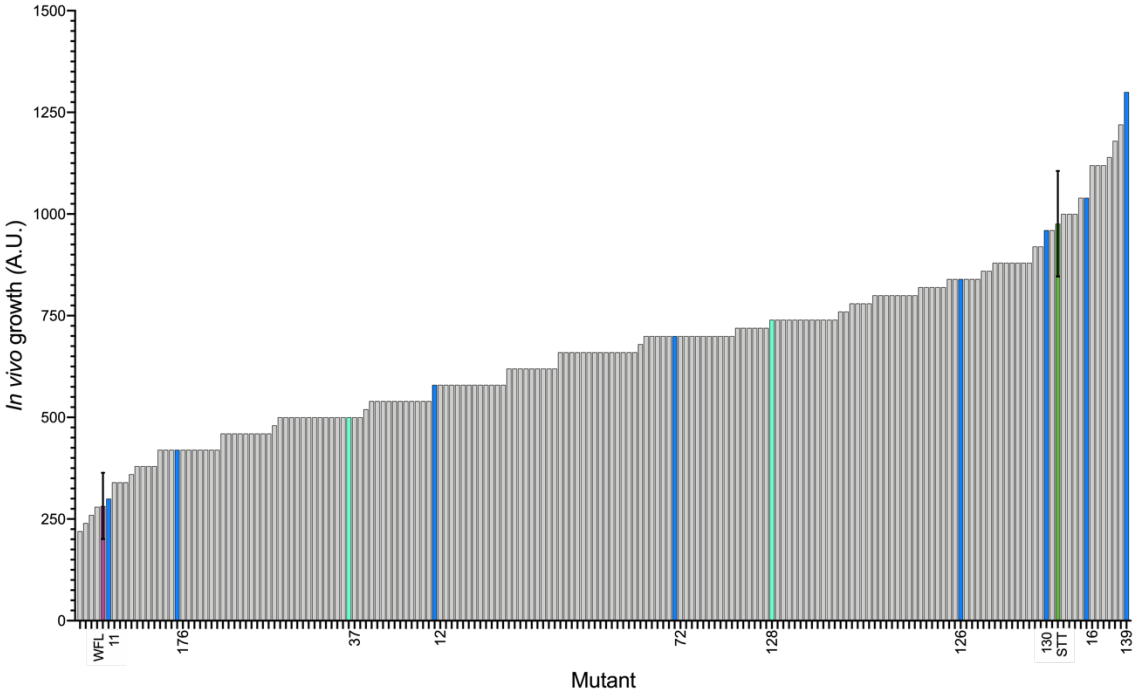


Figure 4.3 Selection of sequences to take forward. Eight vaiants (blue bars) were selected based on their *in vivo* growth seperated by the standard deviation of scFv-STT (130 A.U.). Two further variants (cyan bars) were selected that did not contain mutations to W35, F36 and L64.

Variant	Mutations
11	Q6R (V_H), T112bA (V_H), S22P (V_L), I91V (V_L)
176	Q6R (V_H)
37	Q6P (V_H), I59T (V_H), S26P (V_L)
59	W35R (V_H), V42A (V_H), V76A (V_H), I54T (V_L), D74G (V_L)
128	Y112D (V_H), K18E (V_L), L45P (V_L), T48A (V_L)
72	F30S (V_H), I56T (V_H), E97D (V_H), D116G (V_H), N36D (V_L)
126	I56F (V_H), F62S (V_H)
130	T29A (V_H), I57T (V_H), L64H (V_H)
16	F36S (V_H), F62S (V_H), V87A (V_H), F118I (V_L)
139	S25G (V_H), F36S (V_H), I110T (V_H), S86T (V_L)

Table 4.2 Identity of amino acid substitutions within the ten evolved IgGs. Sequences are ordered in increasing *in vivo* growth score. Residues W35, F36 and L64 from scFv-WFL are underlined. Residues in purple are those identified in WFL evolution hotspots.

The location of each of these mutations (Figure 4.4) are spread across the whole protein sequence. Although most mutations occur within the CDRs of the V_H domain, mutations are also introduced within the framework of the V_H domain. Seven of the ten sequences also contain at least one mutation within the V_L domain.

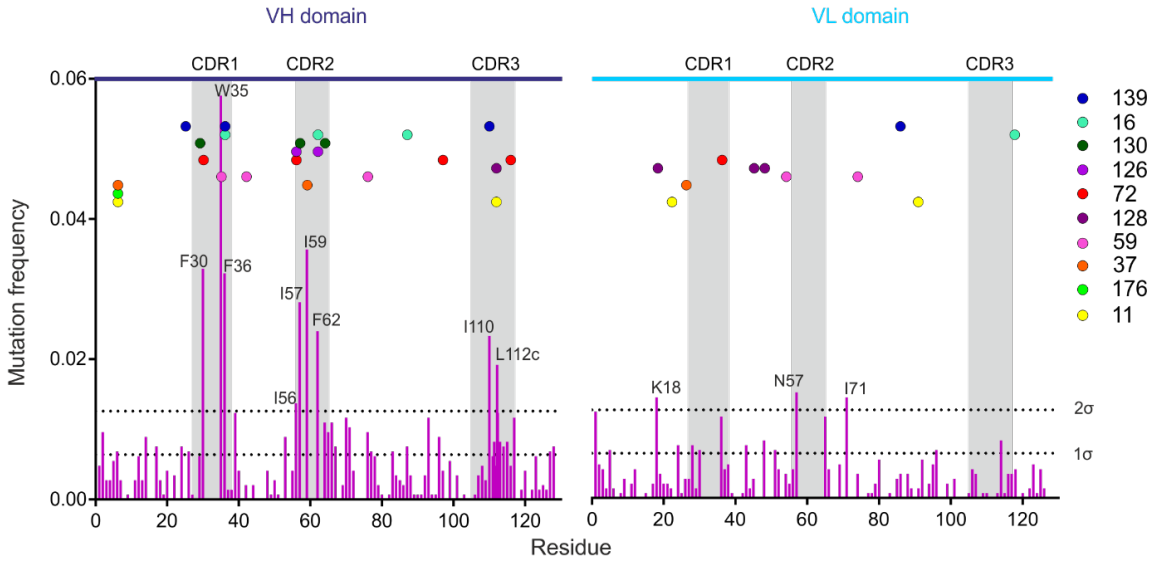


Figure 4.4 Location of substituted residues. Circles show the location of the amino acid substitutions of each of the ten evolved variants relative to the mutational frequency graph of the scFv-WFL evolved library.

A full TPBLA was performed on each variant in a single experiment (Figure 4.5) to ensure reproducibility of the trend obtained previously (Figure 4.1) before cloning into expression vectors for the purification as IgG1 molecules.

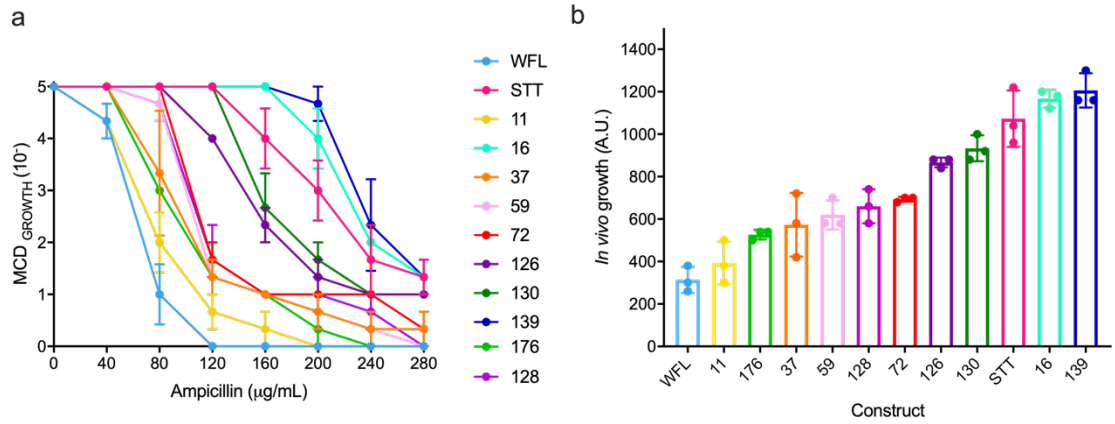


Figure 4.5 *In vivo* growth of selected scFv variants. Ten variants were selected for *in vitro* characterisation based on their *in vivo* growth. a) MCD growth curves of ten evolved scFv-WFL mutants and b) *In vivo* growth scores calculated from the area under the curve of values from a. Error bars represent s.e.m. (n = 3).

4.4 *In vitro* characterisation of evolved proteins

Introducing substitutions into a protein sequence can have an effect on a wide range of biophysical characteristics, such as the functional properties, stability and aggregation propensity[281]. To demonstrate that the *in vivo* growth score for these directed evolution variants correlate with improved biophysical properties, *in vitro* characterisation of each IgG was next investigated. The V_H and V_L domains were codon optimised for eukaryotic expression and cloned into human TM-YTE IgG1 heavy and light chain expression vectors[282] for expression in HEK293 mammalian cells, from which the IgG was purified from culture medium using Protein A chromatography.

As described in Section 1.5 a plethora of techniques are currently employed to study the biophysical properties of lead candidates *in vitro*[98,153]. Here, the molecules were studied for the effect that the mutations have had on the aggregation of IgG-WFL, along with the stability of the protein.

4.4.1 Aggregation

In Chapter 3, the non-specific interactions of IgG-WFL and IgG-STT was determined using HP-SEC. Briefly, IgG-WFL displayed an asymmetric elution profile with a longer retention time than expected (based on monomer mass), due to non-specific interactions with the column matrix. Consequently, the retention time for the evolved IgG variants were compared to assess the non-specific interactions (Table 4.3). The retention times highlight that IgG-WFL has the longest retention time (15 mins) and the evolved variants displayed shorter retention times, presumably due to a reduction of non-specific interactions. The area of the elution profile of the monomer peak can also be used to determine the percent monomeric species in the sample. Interestingly, there was no simple correlation between *in vivo* growth score and monomeric population. Despite this, the sequences that behaved similarly or better *in vivo* to the aggregation resistant scFv-STT (130, 16 and 139), had high levels of monomeric species and retention times expected for the monomeric IgG1 molecular weight (Table 4.3).

IgG variant	Retention time (mins)	% Monomer
WFL	15	ND
11	12.1	85.7
176	10.4	ND
37	12.3	ND
59	12.3	87.7
128	10.1	52.7
72	10.3	33.3
126	9.7	48.5
130	9.9	90.5
STT	9.8	98
16	8.5	96.9
139	9.3	98

Table 4.3 Retention time and percent monomeric species of evolved IgGs determined by HP-SEC. The variants are listed in ascending *in vivo* growth. The monomer peak retention time for each variant is listed. The area of the monomer peak was calculated to determine the % monomer species. ND = not determined from the elution profile.

It was previously shown in Chapter 3 that the retention times of single and double substitutions to IgG-WFL correlated with the enhanced bacterial survival for each of the scFv constructs in the TPBLA. Therefore, the retention times for each evolved IgG was overlaid with its *in vivo* growth score as an scFv to see if a similar trend is observed from directed evolution (Figure 4.6). Once again, it was found that the scFvs with low *in vivo* growth scores (high inferred aggregation propensity) had higher retention times suggesting non-specific interactions with the column matrix. Whereas the evolved variants with high *in vivo* growth (low inferred aggregation) had retention times expected for the monomer mass, producing an inverse correlation between the β-lactamase-scFv *in vivo* growth score and the IgG retention time, analogous to that identified in Chapter 3.

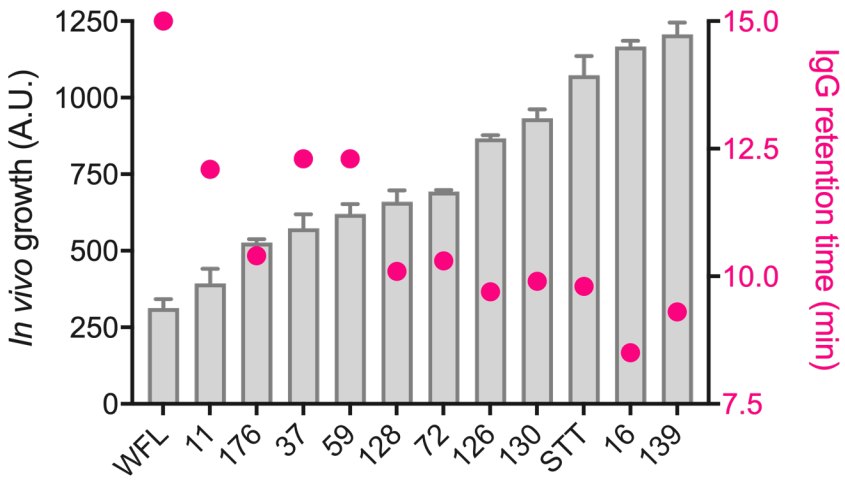


Figure 4.6 HP-SEC retention times of evolved IgGs. HP-SEC retention time (pink dots, right *y* axis, longer times indicate greater interaction with column matrix). These data correlate inversely with *in vivo* growth score (grey bars represent mean values, error bars represent s.e.m. n = 3 technical repeats).

To measure the self-association of the evolved mAbs, affinity-capture self-interaction nanoparticle spectroscopy (AC-SINS) was used[155,168,169]. This method utilises gold nanoparticles that are coated with 'capture' antibodies that bind to the Fc region of human antibodies. When a solution of the test antibody is added they rapidly become immobilised through conjugation to the gold nanoparticles via the capture anti-Fc antibodies. If the test IgG has an increased propensity to self-associate, this causes the antibodies to attract one another, resulting in clustering of the gold nanoparticles. This can then be detected by spectroscopy, through measuring the red-shift of the plasmon wavelength (wavelength of maximum absorbance), as clustering results in shorter interatomic distances and longer absorption wavelengths (Figure 4.7).

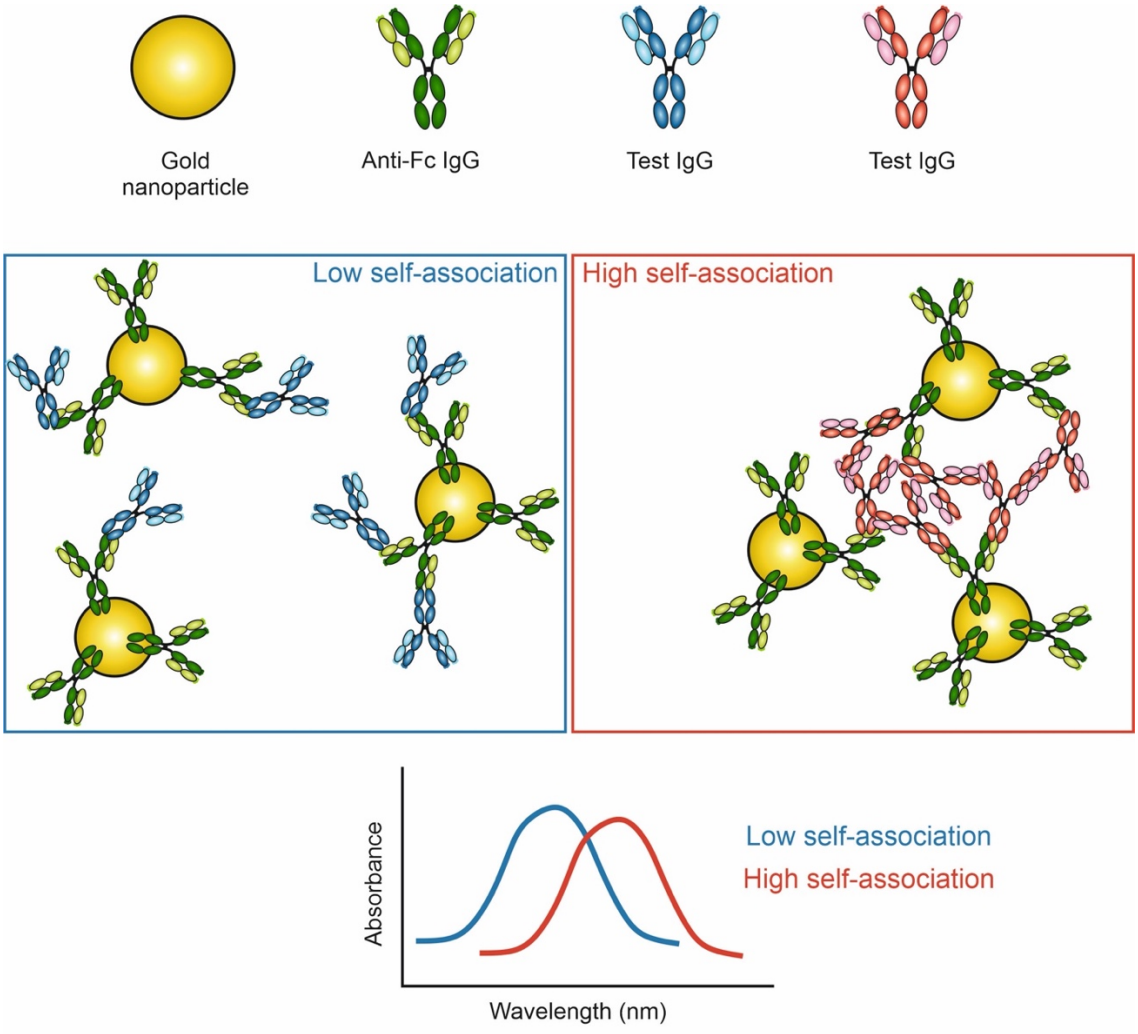


Figure 4.7 Schematic of AC-SINS. Gold nanoparticles are coated with 'capture' antibodies (green), that bind to the test IgG (blue = IgG with low self-association and red = IgG with high self-association) via binding to the Fc region. Test IgGs with low self-association remain dispersed in solution (blue). Test IgGs with high-self association will attract and reduce the interparticle separation distances. The interparticle separation distances (which is reduced upon mAb self-association) is detected via measurements of plasmon wavelength (wavelength at maximum absorbance). Red shifted wavelengths correlate to attractive self-interactions.

The results from AC-SINS identified three groups of behaviour within the antibodies: those that highly self-associated (WFL, 11, 176, 37 and 59), an intermediate performing group (128 and 72) and then those with reduced self-association (126, 130, STT, 16 and 139). Again, overlaying these results with the *in vivo* growth scores from the TPBLA identified an excellent correlation between the magnitude of red-shift in AC-SINS and the enzymatic activity in the TPBLA (Figure 4.8).

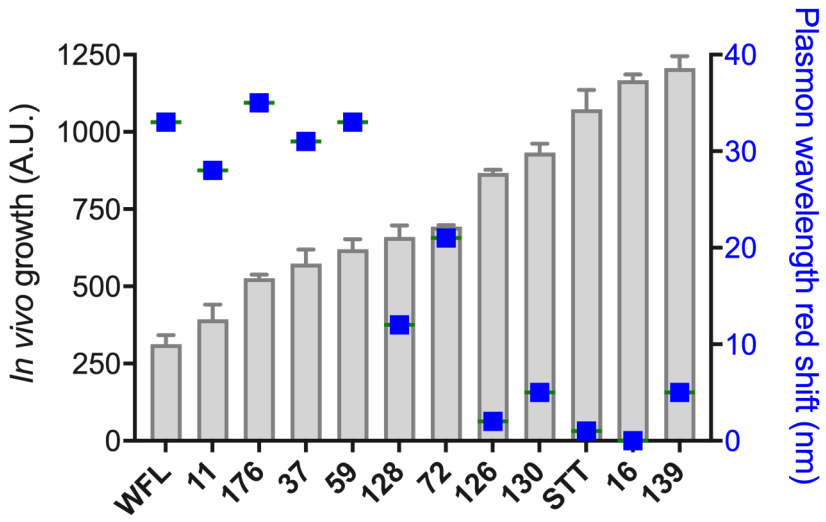


Figure 4.8 AC-SINS of evolved IgGs. AC-SINS (blue squares, right *y* axis, larger plasmon shifts correlate with greater self-association. n = 3 technical repeats). These data correlate inversely with *in vivo* growth score (grey bars represent mean values, error bars represent s.e.m. n = 3 technical repeats).

4.4.2 Stability

Proteins tolerate narrow ranges of stability and aggregation[283]. As some residues are important for both properties, changing the identity of the amino-acid at these positions can simultaneously affect each characteristic[281]. Since the TPBLA was initially developed to identify thermodynamically stabilising mutations *in vivo*[242], the evolved IgGs from this study were also examined for any changes in thermostability (Figure 4.9 and Table 4.4).

Differential scanning fluorimetry (DSF) was employed to measure the temperature at which the C_H2 (T_m1) and the Fab (T_m2) domains of the IgG unfold[165,166,284]. The protein is denatured by increasing the temperature of the solution in the presence of a fluorescent dye, SYPRO Orange. An interaction with a hydrophobic surface increases the quantum yield of the dye[284]. As the protein unfolds, buried hydrophobic regions of the protein become exposed resulting in an increase in the fluoresce signal, from which the transition temperatures can be calculated[284]. The higher the T_m, the higher the thermal stability and therefore the protein is less likely to spontaneously unfold which may cause an immunogenic response.

The first unfolding transition (T_m1) was found to be similar for all IgGs (~56 °C), suggesting that the mutations do not have any allosteric effects on protein stability (Figure 4.9 and Table 4.4). The second unfolding transition (T_m2) revealed that the sequences with high *in vivo* growth score, maintained or slightly improved the T_m in comparison to WT (73 °C), (Figure 4.9 and Table 4.4). Variants with lower *in vivo*

growth score (11, 176 and 37) had the stability of the Fab domain reduced by ~10 °C in comparison to IgG-WFL. Additionally, for variants 59 and 72, no T_m2 was detected by the fluorescence emission, suggesting that the Fab unfolds at the same time as the C_H2, and therefore the mutations have greatly impacted the stability of these two proteins. Both IgG-59 and IgG-72 contained the greatest number of mutations of all the evolved variants studied (five each, Table 4.2). This enhanced sequence variation could be the basis for this change in stability, highlighting the need to keep mutation rates low. Overall, no correlations can be drawn between *in vivo* growth score and stability, presumably because self-association is responsible for the poor biophysical behaviour of IgG-WFL, and its aggregation is not driven by thermodynamic stability.

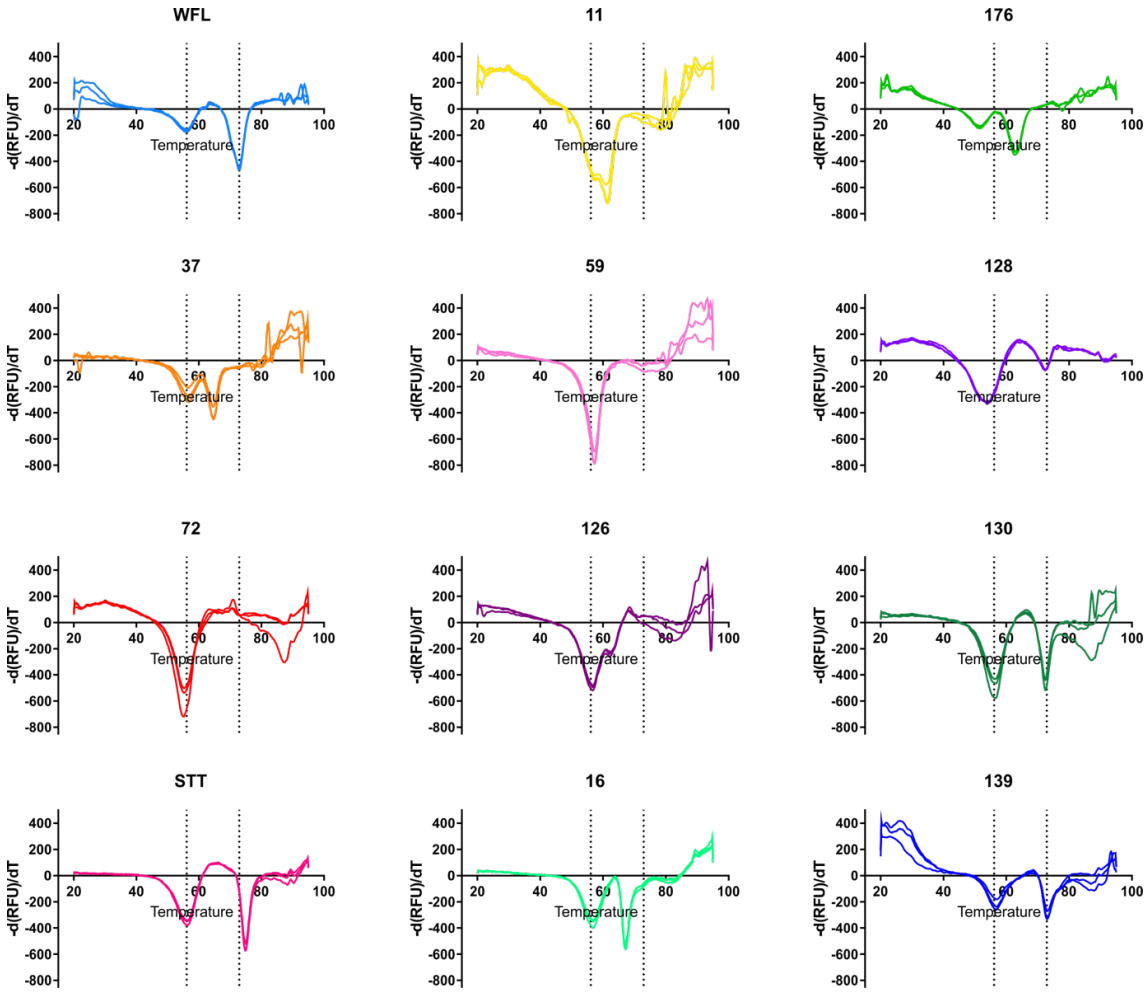


Figure 4.9 Thermal stability of evolved IgGs. Differential scanning fluorimetry (DSF) measurements of IgG-WFL, IgG-STT and the ten evolved variants selected for study. Graphs are arranged by increasing *in vivo* growth assay score from top left to bottom right. Triplicate biological repeat data are presented as first derivatives of fluorescence units (RFU) versus temperature (°C). The T_m values of IgG-WFL (56 °C and 73 °C) are shown on all plots as dotted lines to enable comparison of the data to WT. The T_m values for each are also listed in Table 4.4. Data provided by Janet Saunders (AstraZeneca).

IgG	T_m1 (°C)	T_m2 (°C)
WFL	56.1 ± 0.1	72.9 ± 0.1
11	57.4 ± 0.0	61.3 ± 0.1
176	51.5 ± 0.2	62.8 ± 0.0
37	56.7 ± 0.2	64.6 ± 0.0
59	57.3 ± 0.1	*
128	54.1 ± 0.1	72.4 ± 0.0
72	55.2 ± 0.0	*
126	56.7 ± 0.1	62.1 ± 0.1
130	56.4 ± 0.2	72.4 ± 0.0
STT	56.3 ± 0.1	74.8 ± 0.0
16	56.8 ± 0.0	67.2 ± 0.0
139	56.8 ± 0.0	73 ± 0.0

Table 4.4 Melting temperatures of evolved IgGs. Thermal stabilities of IgG-WFL, IgG-STT and their variants showing transition mid-point temperatures (T_m1 and T_m2) from first and second peaks of first derivative DSF measurements. Errors represent s.d. (n = 3 biological repeats). * = a single transition temperature detected. Thermal unfolding profiles for each variant are shown in Figure 4.9. Data provided by Janet Saunders (AstraZeneca).

4.4.3 Binding affinity

Protein engineering and evolution for enhanced expression, increased stability or reduced aggregation can often impede on the functional activity of the protein[285]. The application of the TPBLA in this context only includes a single selection pressure for reduced aggregation and thus there is the potential for loss of activity, akin to activity/stability trade-offs[279].

To assess this possibility, a homogeneous time resolved fluorescence (HTRF)[286,287] epitope competition assay was employed to establish the relative affinity of each IgG to the cognate antigen, NGF (Figure 4.10). The assay determines the affinity by measuring the reduction in binding of NGF to DyLight650-labelled IgG-WFL in the presence of increasing concentrations of test IgG. The binding of DyLight650-IgG-WFL to NGF is detected by FRET between streptavidin Europium cryptate (which binds the biotinylated NGF) and the DyLight650 label. The decrease in FRET signal

can then be used to calculate an inhibitory concentration (IC_{50}) value for each test IgG.

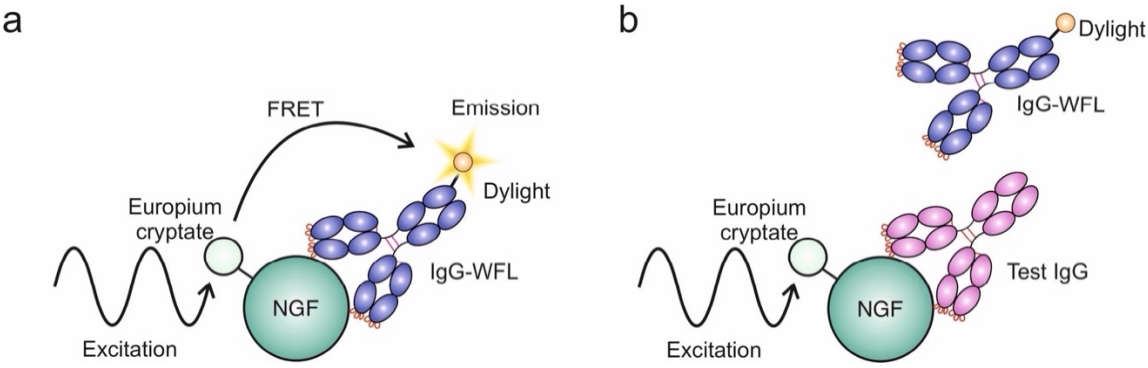


Figure 4.10 Schematic representation of the homogenous time-resolved fluorescence assay. a) Europium cryptate binds to biotinylated NGF (green). Binding of IgG-WFL (purple) is detected through excitation of europium cryptate and FRET to the Dylight conjugated IgG-WFL. b) The relative affinity of IgG variants (pink) is established by measuring the inhibition of NGF binding to Dylight IgG-WFL.

The results showed that all of the evolved IgGs maintained their affinity for NGF (Figure 4.11 and Table 4.5), with all constructs having higher affinity to NGF than MEDI578 (the parent antibody prior to affinity maturation to IgG-WFL[46]). It was observed that four out of the ten evolved IgGs (11, 176, 37 and 59) had improved affinity for NGF over IgG-WFL, whereas the remaining six had slightly lower affinities to IgG-WFL (363 ± 2.4 pM), with IgG-126 having the largest reduction (5.33 nM) to the affinity of IgG-WFL to NGF (Table 4.5).

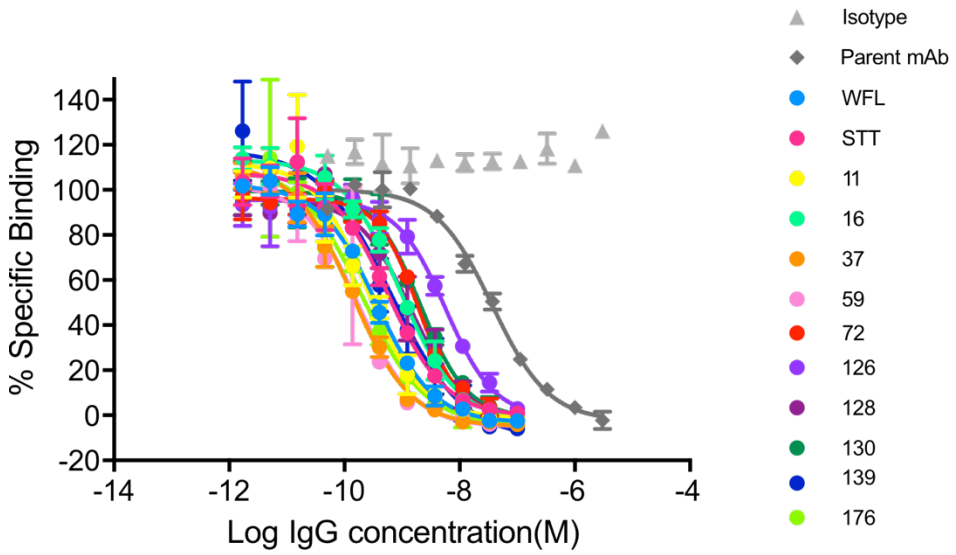


Figure 4.11 Binding affinity of evolved IgGs to NGF. Data used to calculate the IC_{50} values for binding of the ten evolved variants in an IgG1 format to NGF using a HTRF assay. IC_{50} values calculated are shown in Table 4.5. IgG MEDI578 is the parent antibody subjected to affinity panning which generated IgG-WFL. 'Isotype' is a negative control antibody is non-specific to NGF. Data represent mean values, error bars represent s.d. n = 3 technical repeats.

Evolved construct (IgG)	IC_{50} (pM)
11	251 ± 4.8
176	201 ± 6.4
37	136 ± 3.1
59	177 ± 5.9
128	1413 ± 4.2
72	1911 ± 2.0
126	5693 ± 4.5
130	2063 ± 3.2
16	935 ± 2.6
139	649 ± 6.9
WFL	363 ± 2.4
STT	573 ± 4.0
MEDI578	36760 ± 2.8

Table 4.5 IC_{50} values of evolved IgGs binding to NGF. Values measured using an epitope competition assay (Figure 4.11) (n = 3 technical repeats, error = s.e.m). IgG MEDI578 is the parent antibody subjected to affinity panning which generated IgG-WFL.

Importantly, all IgGs retained their functional activity whilst simultaneously reducing the aggregation propensity to various extents. No correlation is observed between the IC_{50} and the *in vivo* growth score (Figure 4.12). In Chapter 3, it was postulated that the IgGs would retain their affinity for NGF, as the majority of the evolution hotspots reside in the V_H CDRs 1 and 2, and not in CDR3 which is important for function. From the evolved sequences analysed in this study, only four clones contained mutations in the V_H CDR3 (Table 4.2, sequences 11, 128, 72 and 139, CDR3 = residues 105-117 IMGT numbering), however these mutations had minimal effect on their affinity for NGF.

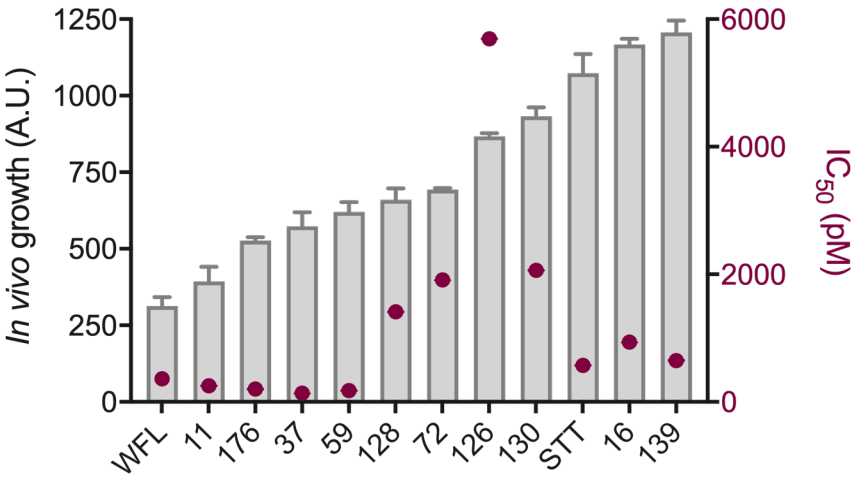


Figure 4.12 Comparison of *in vivo* growth scores and IC_{50} values. *In vivo* growth scores of βla-scFv-WFL variants (bars), error bars indicate s.e.m (n = 4 biologically independent experiments) overlaid with IC_{50} values (maroon circles). Error bars for the IC_{50} values are smaller than the symbol (n = 3 technical repeats).

4.5 Directed evolution of an IgG with a different mechanism of aggregation

To assess the general applicability of the TPBLA for the evolution of aggregation resistant antibodies, a second test IgG scaffold was selected. IgG-Li33 was originally isolated as a Fab using phage display to screen against the glycoprotein LINGO-1[188]. Inhibiting LINGO-1 aims to restore the repair of damaged myelin as a potential therapeutic for multiple sclerosis[288]. The antibody isolated had high affinity and high bioactivity, however it was found to exhibit glycosylation-dependent aggregation and poor solubility[188]. It was found that switching the IgG framework dramatically enhanced the solubility (i.e., 0.9, >50 and >30 mg/mL when expressed as IgG1, IgG2 and IgG4, respectively)[188]. This suggests that, in contrast to IgG-WFL, aggregation of IgG-Li33 is mediated via CDR-framework interactions.

4.5.1 Comparison of mutation frequency profiles for evolved antibody fragments

In order to understand whether the mutation frequency profile for scFv-WFL was specific for this Fv sequence or reflected innate frustration of the Ig-fold itself, directed evolution was performed on scFv-Li33. Using an identical procedure to that described for WFL in Chapter 3, a library of Li33 variants inserted between domains 1 and 2 of β-lactamase was created (estimated to be 1×10^6 mutants) and the colonies

selected at 140 μg/mL ampicillin (see inset of Figure 4.15 for the *in vivo* growth curve of Li33).

The resultant mutational frequency profile from 140 DNA sequences contrasts markedly with that for WFL (Figure 4.13). Only three residues, F30 and I36 in V_H (most commonly substituted with S or T, respectively) and L53 in V_L (most commonly substituted with P), exhibited substitution rates greater than two standard deviations higher than the mean. The diversity of the profiles for Li33 and WFL is remarkable, given the similarity of their framework regions (66.5 % similarity and 48.2 % identity (Figure 4.14)). The mutation hotspots identified for each IgG is therefore unlikely to be due to the innate frustration of the Ig-fold but instead indicative of their different aggregation mechanisms: aberrant CDR-CDR (IgG-WFL) and CDR-framework (IgG-Li33) interactions.

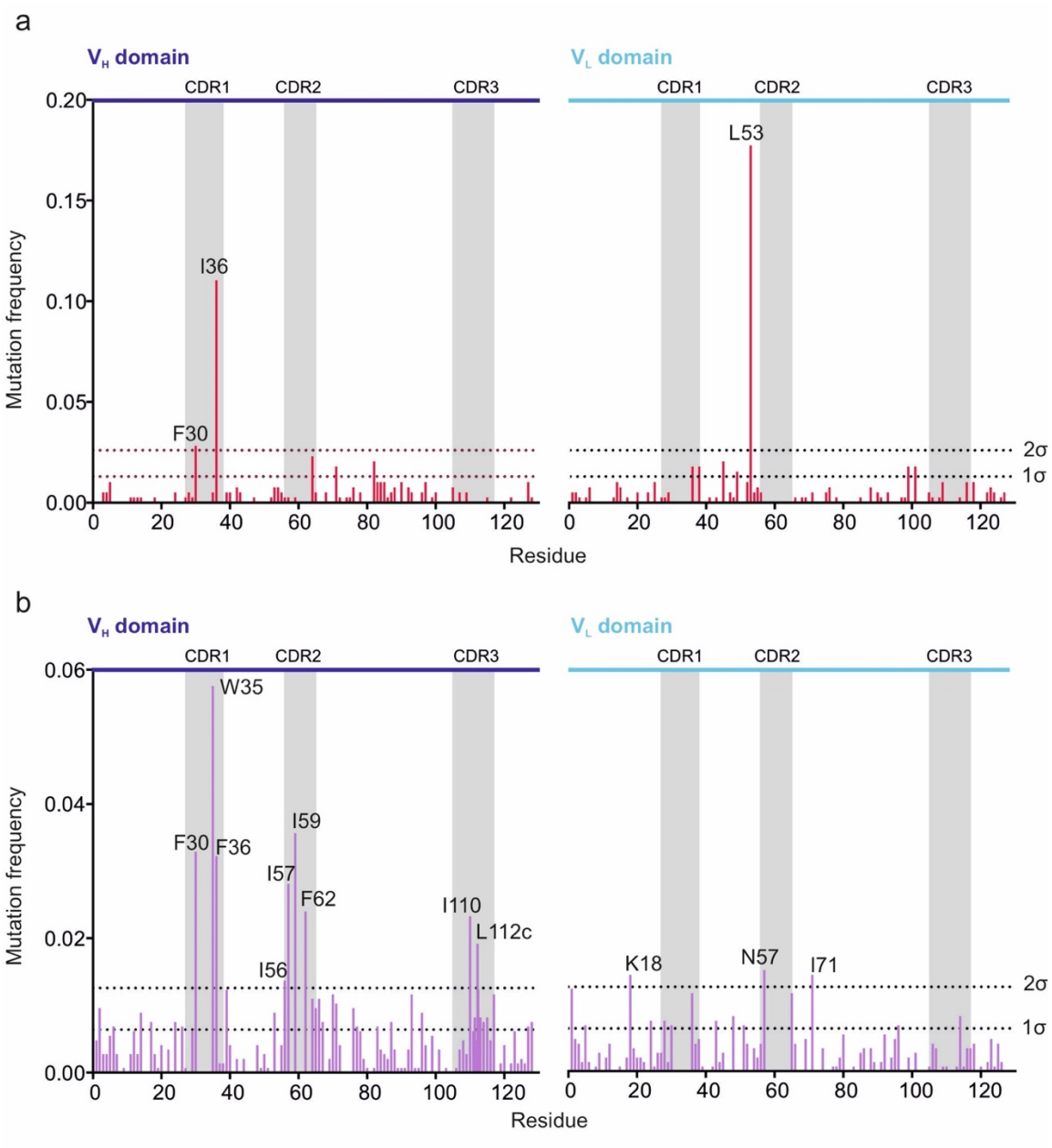


Figure 4.13 Mutation frequency profiles of evolved antibody fragments. a) Mutational frequency of the screened scFv-Li33 library identifies only three residues with a mutational frequency >2σ. b) The mutation frequency profile of scFv-WFL (Chapter 3). Variable heavy and light chain domains are labelled, and the grey boxes highlight the CDRs. Residues are numbered using IMGT numbering.

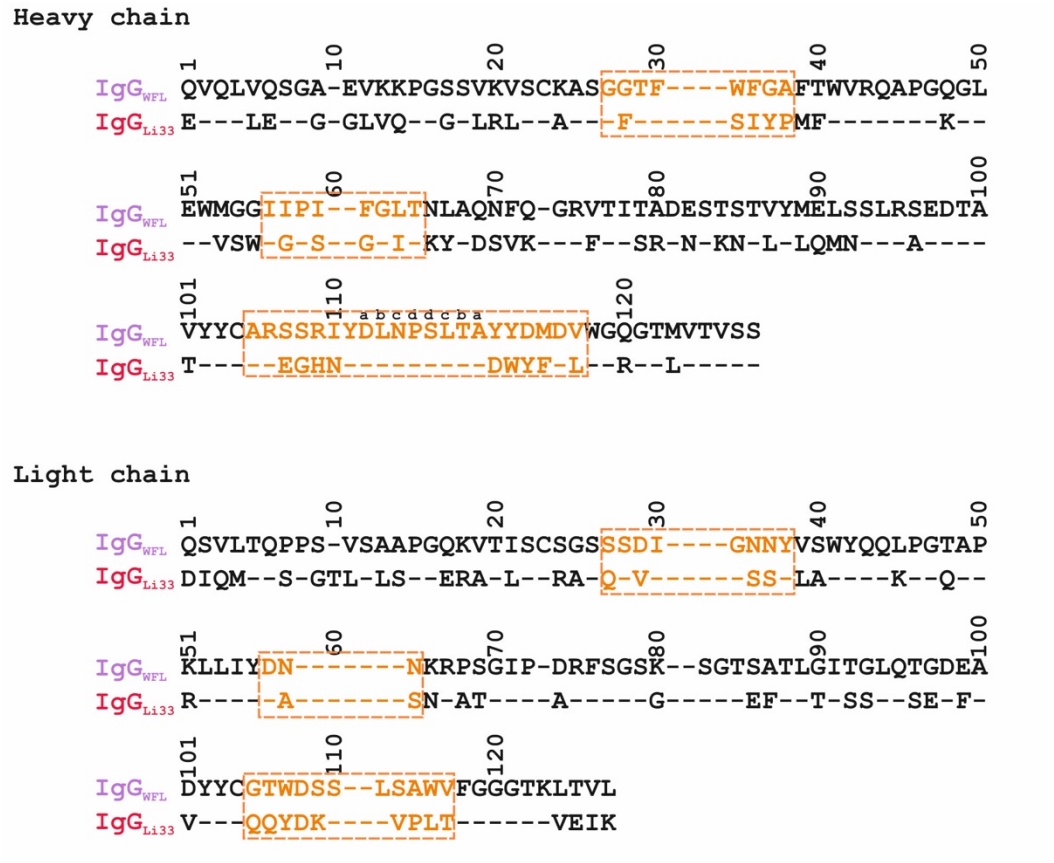


Figure 4.14 Sequence alignment of IgG-WFL and IgG-Li33. CDRs are highlighted in orange. Dash '-' represents conserved residues between the two sequences. Dashes within the IgG-WFL sequence denote IMGT numbering gaps (calculated using ANARCI server[258]).

4.5.2 Screening evolved Li33 proteins

Out of the 140 colonies that grew under the 140 μg/mL selection pressure, 66 were randomly selected and subjected to a full *in vivo* growth assay (Figure 4.15). This identified that 64 of sequences had enhanced *in vivo* growth relative to WT βla-scFv-Li33. However, it is unknown what evolutionary pressure the scFvs have been selected for since the solubility of IgG-Li33 depends critically on the IgG scaffold, yet the TPBLA uses scFv sequences and so cannot detect the aberrant CDR-framework interactions that occur in the IgG format. To investigate this, sequences were selected for *in vitro* analysis in an IgG1 scaffold as this was the scaffold with the lowest solubility for IgG-Li33. Three variants were selected for this analysis, Y88D (V_H) the variant with a single point mutation with the highest *in vivo* growth score identified during screening (Figure 4.15), along with I36T (V_H) and L53P (V_L), two of the mutation hotspots identified (Figure 4.13) that had enhanced growth measured in the TPBLA (Figure 4.15).

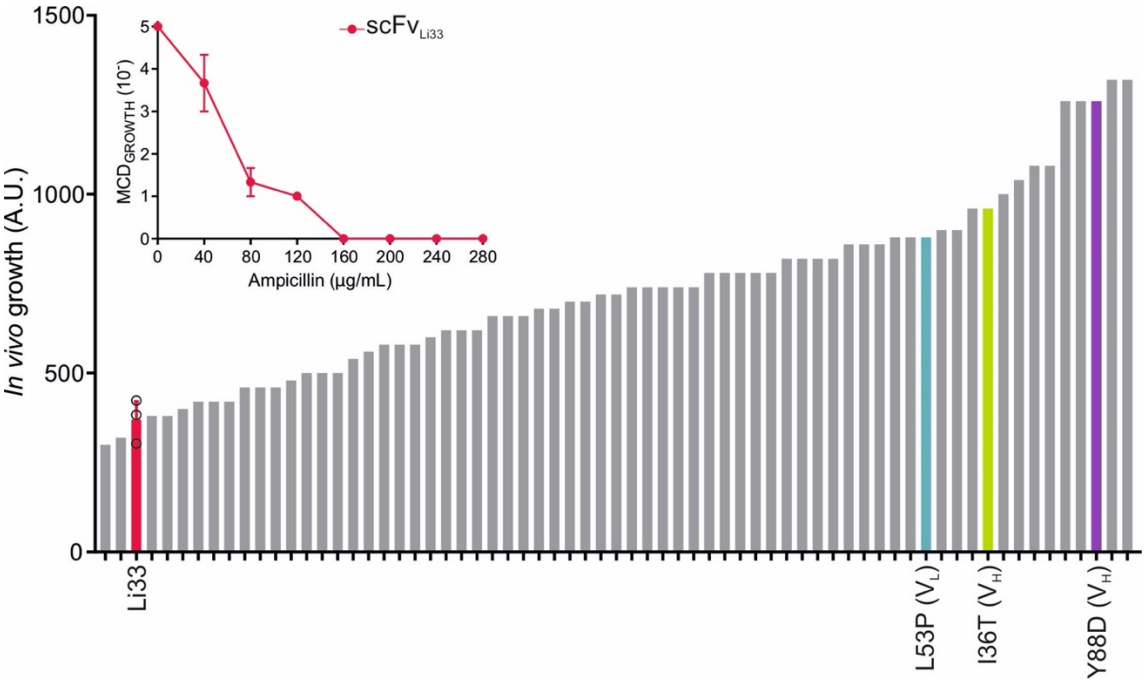


Figure 4.15 Full TPBLA assay of evolved scFv-Li33 variants. *In vivo* growth scores of 66 evolved scFv-Li33 variants. Wild-type scFv-Li33 together with the frequently mutated 'hotspot' residues L53P (V_L), I36T (V_H) and the highest scoring single point mutation, Y88D (V_H) are highlighted. The full *in vivo* growth curve for wild-type βla-scFv-Li33 is shown inset. Error bars = s.e.m (n = 3 independent experiments).

Since the evolved sequences were selected in an scFv format, it was postulated that the variants had enhanced stability, as was previously observed for the soluble globular protein Im7 using this assay[242]. To test this theory, the thermal stability of the IgGs was measured using DSF (Figure 4.16 and Table 4.1). Surprisingly, no significant changes in thermal stability were detected between the wild-type and evolved IgG-Li33 variants.

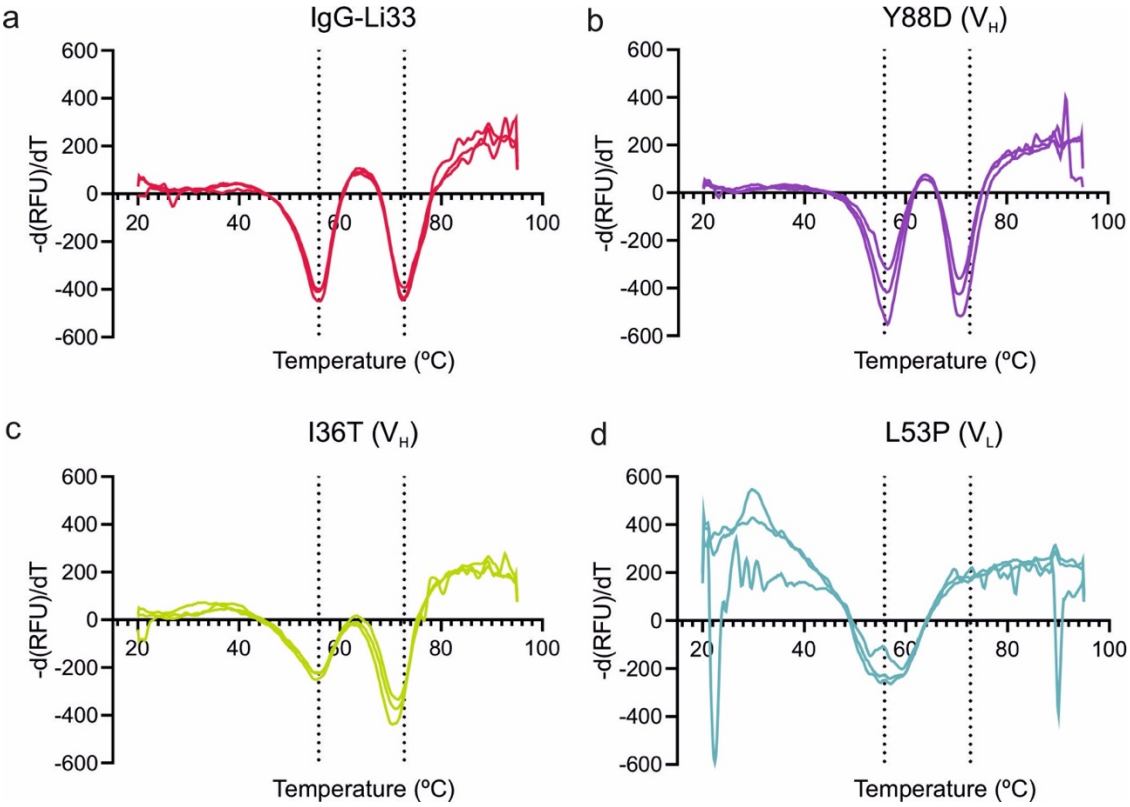


Figure 4.16 Thermal stability of IgG-Li33 and three evolved variants. DSF measurements for a) WT IgG-Li33, b) IgG-Y88D (V_H), c) IgG-I36T (V_H) and d) IgG-L53P (V_L). Data are presented as first derivatives of relative fluorescence units (RFU) versus temperature (°C). The thermal unfolding transitions of IgG-Li33 (56 and 73 °C) are shown on all plots as dotted lines to enable comparison of the data. The T_m values for each IgG are also listed in Table 4.6.

IgG	T_m1 (°C)	T_m2 (°C)
Li33	55.8 ± 0.4	72.7 ± 0.8
L53P	55.6 ± 1.1	*
I36T	55.8 ± 0.4	71.0 ± 0.7
Y88D	56.2 ± 0.2	71.2 ± 0.9

Table 4.6 Melting temperatures of IgG-Li33 and three evolved variants. Thermal stabilities of IgG-Li33 variants showing transition mid-point temperatures (T_m1 and T_m2) from first and second peaks of first derivative DSF measurements. Errors represent s.d. (n = 3 biological repeats). * = a single transition temperature detected. Thermal unfolding profiles for each variant are shown in Figure 4.16.

To determine if the evolved IgGs had improved self-association, similar to the evolved IgG-WFL antibodies (section 4.4.1), the self-association was measured by AC-SINS (Figure 4.17). This identified that IgG-Li33 had high levels of self-association, as did Y88D, the single mutant with the highest *in vivo* growth in the TPBLA. A reduction in self-association was observed for both I36T and L53P over wild-type Li33, suggesting that the substitutions I36T and L53P may have different effects on the mechanism of protein aggregation to Y88D.

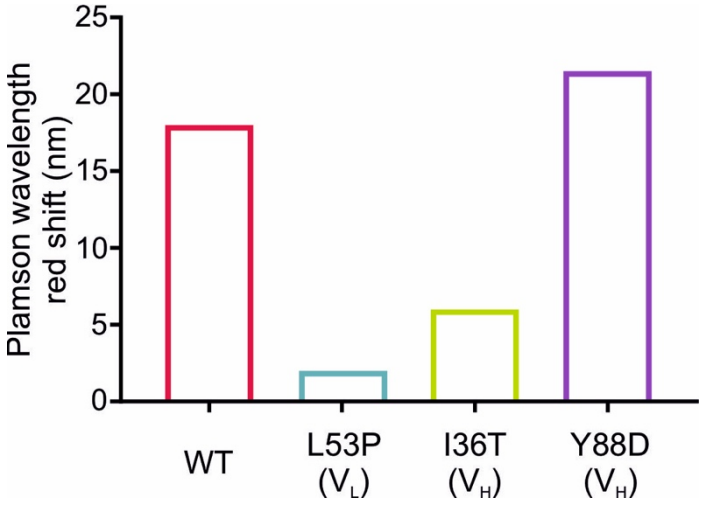


Figure 4.17 AC-SINS of IgG-Li33 and evolved variants. Self-association of IgGs was measured using AC-SINS where a larger plasmon shift correlates with greater self-association.

Since IgG-Li33 had poor solubility in an IgG1 scaffold, the evolved variants were subjected to a polyethylene glycol (PEG) precipitation assay (Figure 4.18) to investigate the effect of evolution on their solubility. The results found that indeed, each of the three evolved variants had improved solubility relative to IgG-Li33. These results indicate that the TPBLA can resolve the limiting factor of a protein irrespective of the mechanism of aggregation.

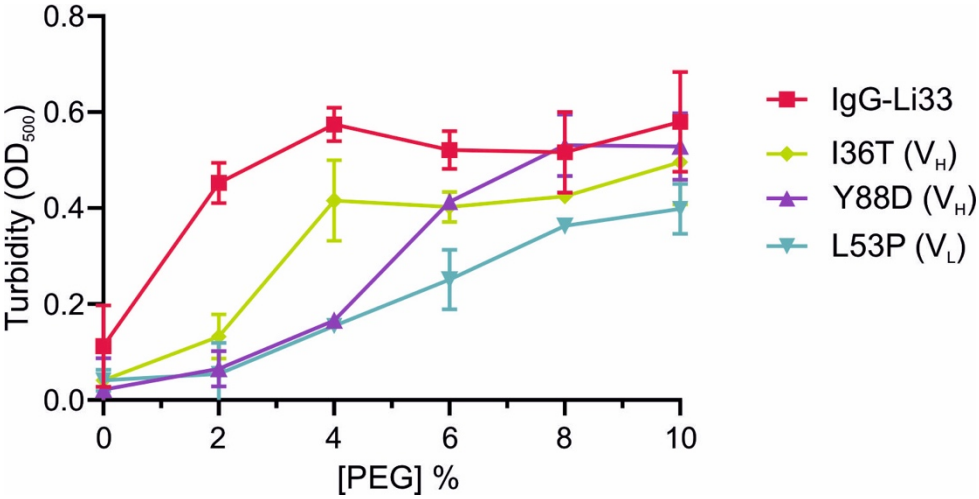


Figure 4.18 PEG precipitation of IgG-Li33. Precipitation was quantified by turbidity at 500 nm. Error bars represent s.d. (n = 3). Data generated by Janet Saunders (AstraZeneca).

4.6 Discussion

Having established that the TPBLA could be used for candidate selection to screen and identify aggregation hotspots in Chapter 3, this chapter investigated the capability of the assay as a screen for directed evolution to search for novel sequences able to ameliorate the poor developability of IgGs.

From the 315 evolved scFv-WFL variants that were identified in Chapter 3, 185 were randomly selected and their aggregation resistance measured *in vivo*. In an industrial setting, the number of variants to assess could be reduced by using a higher stringency selection (higher antibiotic concentration) in the initial directed evolution selection. For this proof-of-concept study, screening a larger number of sequences allowed for both the best variant to be identified and to select a rank of evolved antibody sequences for further study (Figure 4.3 and Table 4.2).

In addition to isolating candidate variants with enhanced biophysical properties, directed evolution coupled to a screen that successfully measures innate aggregation allows assessment of current *in silico* methods and, if necessary, provides experimental data to further enhance their predictive power. The screening of the protein sequences by both CamSol and Aggrescan identified that the majority of the 185 sequences had improved solubility and reduced aggregation propensity. However, the results between the *in vivo* growth score and the predicted *in silico* scores produced a low R^2 correlation for both algorithms (Figure 4.2). The TPBLA may be more sensitive to mutations than *in silico* screening and therefore screening large datasets of mutational variants in the TPBLA could further aid the development of *in silico* tools. This may also suggest that the TPBLA is measuring something more complex than *in silico* methods, such as the aggregation of partially unfolded states, or a convolution of several properties (e.g., stability and solubility).

Analysis of sequences that outperformed scFv-STT in the TPBLA identified that all sequences contained at least one of the residue hotspots identified in Chapter 3, emphasising the importance of these residues in reducing the aggregation of scFv-WFL. Although nine of the twelve sequences contained at least one mutation to residues to W35, F36 and L64 (residues mutated from WFL to STT) the rationally engineered IgG-STT was not isolated during screening, highlighting the advantages of natural selection over rational approaches.

Expression and purification of ten evolved sequences in an IgG format allowed the relationship between *in vivo* growth and *in vitro* aggregation propensity to be further explored. As with the rationally engineered variants in Chapter 3, an excellent indirect correlation was observed between the IgG retention time and the scFv *in vivo* growth

score (Figure 4.6), in which antibodies with high *in vivo* growth scores exhibited shorter retention times due to reduced interactions with the column matrix. The ten evolved IgGs were also screened using AC-SINS as an orthogonal method to investigate the self-association of each IgG. Similarly, this method also identified a strong correlation between the self-association of the evolved IgGs and the *in vivo* growth score (Figure 4.8).

Although the mechanism of aggregation of IgG-WFL is not thought to be driven by thermal instability[46], the introduction of mutations into a protein can modify the thermal stability, and this assay has previously been employed to select for increased thermostability[242]. Therefore, the T_m for the evolved antibodies were measured to observe if the proteins maintain or have enhanced stability. Both IgG-WFL and IgG-STT have similar T_m values, along with the majority of the IgGs from evolution (Table 4.4). Hence, no correlation was found between the *in vivo* growth score and thermal stability, however the mutations introduced into two clones (59 and 72) significantly reduced the T_m of the Fab fragment.

One caveat of directed evolution experiments is that there is often only one selection pressure and hence only one property is evolved. In this case, only reduced aggregation was selected for which has the potential to reduce the functional activity of the protein. However in this study, the binding affinity to NGF was maintained for all of the evolved variants, to levels still higher than the parent from phage display (MEDI578[46]) whilst concurrently reducing the self-association of the molecules.

A second industry-derived sequence was selected for evolution by the TPBLA that aggregated via a different mechanism to IgG-WFL to understand whether hotspots identified were specific for each scFv, or simply reflected innate frustration of the Ig-fold. The evolution of IgG-Li33 identified a different mutational profile to IgG-WFL, suggesting that the TPBLA is able to identify (and resolve) specific problematic residues between proteins with identical topologies and highly similar sequences. Utilising the TPBLA in combination with a variety of sequences will therefore be a valuable tool for understanding the molecular determinants of aggregation associated with bioprocessing.

In summary, the work in this chapter demonstrated the power of the TPBLA in combination with directed evolution to identify and rectify problematic sequences. The evolved scFvs identified in Chapter 3, were found to have enhanced *in vivo* growth properties that correlated with reduced self-association as an IgG as determined by HP-SEC and AC-SINS, without any compromise for the affinity to NGF. Furthermore, an additional IgG was examined, IgG-Li33, that through evolution generated sequences with reduced self-association and improved solubility. The

TPBLA therefore has broad utility to evolve sequences with reduced aggregation for enhanced bioprocessing.

Chapter 5
Evolution of disease-causing antibody domains

5.1 Objectives

The work described in Chapters 3 and 4 demonstrated that the TPBLA can be used to identify aggregation hotspots in the V_H and V_L domains of IgGs for protein therapeutics. The evolved sequences from the studies correlated with enhanced biophysical properties *in vitro* for antibodies with different mechanisms of aggregation, and thus revealed the general utility of this assay. This chapter aims to explore this utility further, by investigating the use of the TPBLA for the application of antibody V_L domain aggregation in human disease. The objective of this work was to assess whether the TPBLA was able to differentiate between germline and patient derived V_L sequences, and to identify sequence hotspots in disease relevant proteins, furthering our understanding of light chain amyloidosis.

5.2 Light chain amyloidosis

Amyloid formation is a complex process, characterised by the self-assembly of proteins into highly ordered fibrils with a cross-β structure. As discussed in Chapter 1 the accumulation and deposition of amyloid fibrils is associated with more than 50 human diseases, known as amyloidosis[289,290].

Systemic amyloidosis comprises a variety of diseases such as light chain, dialysis-related and transthyretin amyloidosis that are characterised by deposition of fibrils in multiple tissues and organs, such as the heart, kidneys and digestive tract[289]. How amyloid fibrils damage tissue is not fully understood, but organ dysfunction may result from disruption of tissue architecture by amyloid deposits or may be due to the cytotoxicity of non-native conformers[291].

AL is the most common form of systemic amyloidosis, affecting ten patients per million per year[292]. The amyloidogenic protein in AL amyloidosis is an Ig light chain (LC), or a fragment of this, resulting from abnormal monoclonal plasma B cell proliferation and is therefore often found in patients with multiple myeloma[293]. Free LCs secreted without an associated antibody heavy chain are removed from the blood by the kidneys, however LCs can escape this protein quality control and are secreted into the blood stream[293]. In AL patients, the protein misfolds and aggregates forming amyloid fibres that are deposited in the tissues with the heart, and kidneys often being the most severely affected[294,295]

As described in Chapter 1, LCs comprise an N-terminal variable domain (V_L) attached to a C-terminal constant domain. The repertoire of LC sequences is highly diverse due to the random assembly of different gene segments known as variable (V), diversity (D) and joining (J) genes in a process known as V(D)J recombination[44,296–298]. The diversity of antibodies is then further enhanced by the process of somatic hypermutation, that introduces mutations into the V region to enhance antigen binding[44,293,297,299]. Each patient's clonal plasma cells therefore secrete a single, unique LC sequence, and thus the protein sequence implicated in AL is generally unique to the patient[294].

To understand how the mutations introduced during this diversification impact stability and/or aggregation, most research compares patient proteins to their germline sequence. There are 40 kappa and 33 lambda germline genes available to form a V_L, however several germlines genes are highly associated with amyloidosis (most prominently κI, λI, λII, λIII and λVI)[294,300,301].

In AL, amyloid fibres contain full-length LCs or solely the V_L domain, however the ubiquitous presence of the V_L indicates that this domain may be the essential unit for fibril assembly[302–307]. It has been shown *in vitro* that full-length LCs are less amyloidogenic relative to the isolated V_L domain[308–310]. The LCs from AL patients are less kinetically stable than germline sequences and therefore their amyloidogenicity may be initiated by endoproteolysis of kinetically unstable LCs[305,308]. A decrease in thermal stability of the V_L domain can drive amyloidogenicity[311–316], but other factors such as LC dimerization[309,317] and conformational dynamics[300,310,318] also need to be taken into account.

The current treatment is to prevent production of LCs with chemotherapy targeting the underlying B cell clone, however age and cardiac involvement often render patients too high risk for chemotherapy[294]. Daratumumab was recently approved by the FDA for the treatment of AL[319]. This IgG binds to CD38 which is highly expressed in the surface of myeloma cells, triggering the patient's immune system to attack the cancer cell via, complement-depended cytotoxicity and antibody dependent cellular phagocytosis[320]. Despite this success, there is still an urgent need for earlier diagnosis of the disease. Since it is generally not clear which mutations induce misfolding and amyloid aggregation in patients, identifying how and why some, but not all, LCs aggregate as amyloid fibrils may lead to earlier diagnosis and new therapeutic strategies[301].

5.2.1 Test proteins for this study

Two LC fibril structures have recently been determined by cryoelectron microscopy (cryo-EM) after isolation from patient cardiac tissues[306,307]. Both fibril structures show complete structural rearrangement from the native Ig fold, providing insight into the molecular mechanism of LC misfolding and amyloid formation (Figure 5.1 and Figure 5.2).

Radamaker *et al.* determined the structure of an AL patient fibril that had ten mutations relative to the IGLV1-44 germline[306] (Figure 5.1a). The fibril consisted of one protofilament that encompassed a 91-residue segment of the V_L domain. The protofilament fold roughly resembles the shape of a 'ram head' and encloses three cavities, two hydrophobic and one hydrophilic, with the CDRs of the V_L located on the fibril surface (Figure 5.1c). A substantial structural rearrangement occurs from the native state to fibril structure around the disulfide bond. It is thought that misfolding induces a 180° rotation of residue segments 16-23 and 86-93 relative to each other around the disulfide bond (C22 and C89).

Out of the ten mutations present in the patient sequence, only one was predicted to be unfavourable to the native state (I76V, Figure 5.1b). This substitution impacts a buried residue, potentially destabilising the protein. Relative to the fibril structure several mutations have a beneficial effect. Three mutations increase the fibril surface charge (S31R, N35K and N53D), one removes a charge from the non-polar cavity (D94A), and one inserts a basic residue into the polar cavity (S25R).

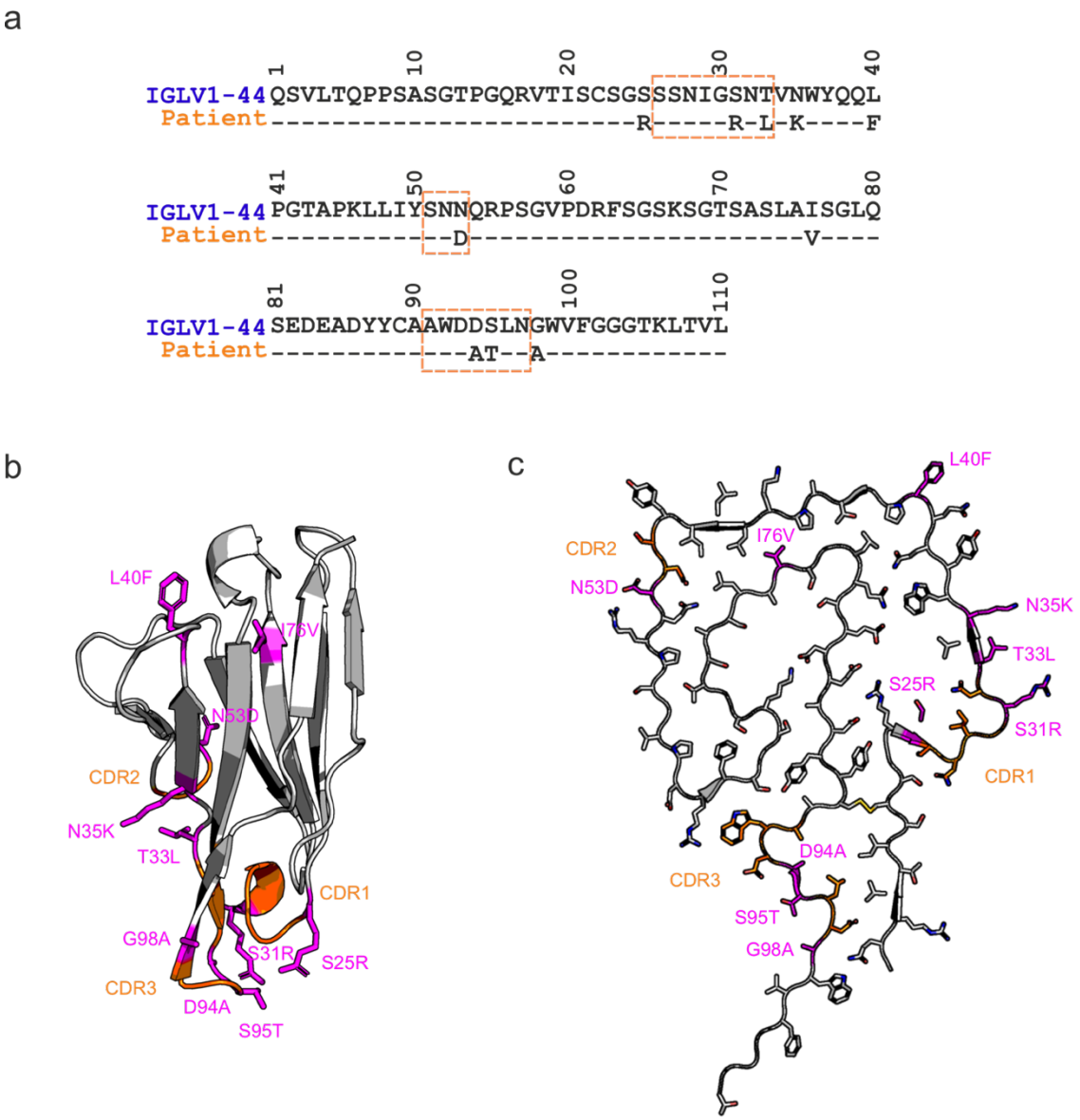


Figure 5.1 IGLV1-44 patient sequence and structure. a) Sequence alignment of V_L domain from the germline IGLV1-44 and patient sequence. CDRs are highlighted in orange. Dash '-' represents conserved residues between the two sequences. b) Native structure of V_L domain (model created using ABodyBuilder[321]). c) Fibril structure of V_L domain (PDB 6IC3[306]). Disulfide bridge is shown in yellow. In both b) and c) CDRs are coloured in orange and mutations found in the patient are labelled in magenta.

Swuec *et al.* also recently reported the fibril structure from a patient with AL containing thirteen substitutions from the IGLV6-57 germline[307] (Figure 5.2a). This fibril was composed of an asymmetric protofilament that encompassed two polypeptide stretches from 77 residues of the V_L domain. The fibril structure has a different fold to that reported by Radamaker *et al.*[306], with the inner polypeptide segment displaying a 'snail-shell' trace (residues 1-37) surrounded by a 'C-shaped' segment. (residues 66-105)[307] (Figure 5.2c). The two segments that make up the fibril core are linked together by a disulfide bridge.

The patient IGLV6-57 sequence contained the mutation R24G that has been reported to be present in 25 % of IGLV6-57 mutant sequences[322,323]. This mutation alone has been found to decrease the stability of the germline by 1.7 kcal/mol and is significantly more fibrillogenic[324]. Eight of the thirteen mutations occur in the CDRs of the patient V_L domain. In the fibril structure, CDR1 (containing mutations R24G, G27A and N32H) and CDR3 (containing mutations S96G, S97N and V100W) contribute the structured fibril core, suggesting that the mutations introduced not only decrease stability and/or increase aggregation of the native state, but also may stabilise the molecular interactions for fibril assembly.

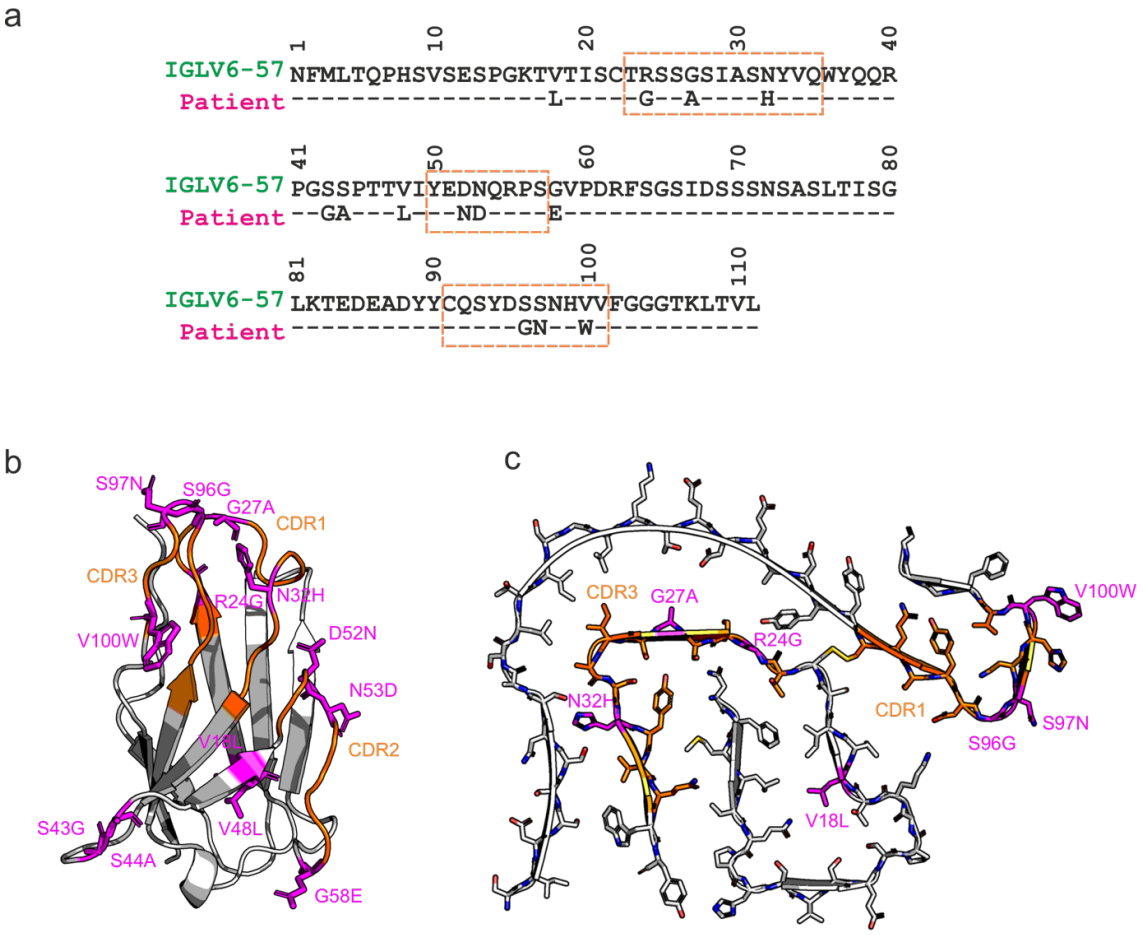


Figure 5.2 IGLV6-57 patient sequence and structure. a) Sequence alignment of V_L domain from the germline IGLV6-57 and patient sequence. CDRs are highlighted in orange. Dash '-' represents conserved residues between the two sequences. b) Native structure of V_L domain (model created using ABodyBuilder[321]). c) Fibril structure of V_L domain (PDB 6HUD[307]). Disulfide bond linking the two peptides is shown in yellow. In both b) and c) CDRs are coloured in orange and mutations found in the patient are labelled in magenta.

5.3 Aggregation screening of germline and patient antibody domains

5.3.1 *In vivo* screening of antibody domains

The V_L domains of both germline and patients were cloned into β-lactamase using Golden Gate assembly (see methods). The pairs of V_L domains were then screened in the TPBLA over a 0-280 μg/mL ampicillin concentration range (Figure 5.3). In both cases, the patient V_L had a lower resistance to ampicillin and therefore have a higher inferred aggregation propensity compared to their germline counterparts. However, overall, only small differences were observed in the *in vivo* growth score of all constructs (Figure 5.3d). Both germline and patient sequences have a high sequence similarity (IGLV1-44, 90 % identity and IGLV6-57, 88 % identity) and appear to aggregate, yet the molecular mechanism of aggregation for each is unknown based on this coarse measurement.

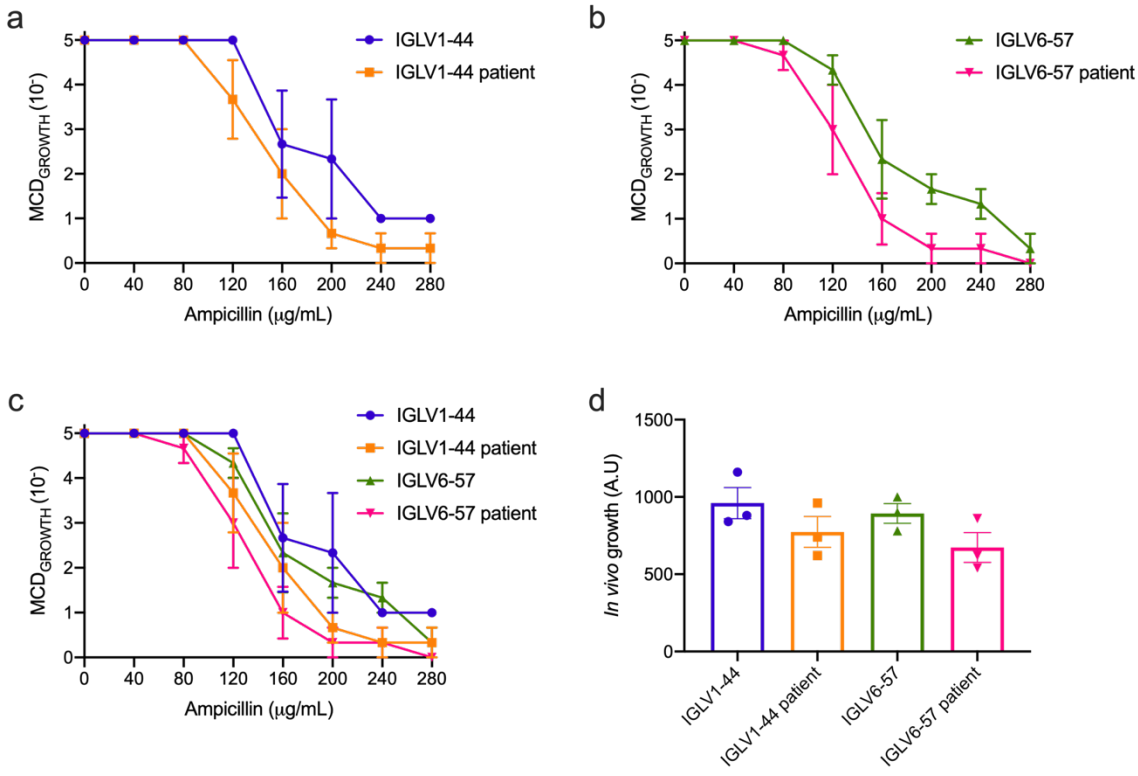


Figure 5.3 *In vivo* growth of germline and patient V_L domains. MCD_{GROWTH} curves of a) IGLV1-44 germline and patient V_L domains and b) IGLV6-57 germline and patient V_L domains over 0-280 μg/mL ampicillin concentration range. c) Both germline and patient *in vivo* growth curves displayed together on one graph. d) The area under the curve is calculated to generate *in vivo* growth values for each construct. Error bars represent s.e.m (n=3).

5.3.2 *In vitro* aggregation of antibody domains

One set of V_L germline and patient samples were selected for *in vitro* purification and aggregation assessment to validate the results observed in the TPBLA, that both germline and patient samples are aggregation prone.

The IGLV6-57 germline and patient V_L domain were targeted to the periplasm of *E. coli* BL21(DE3) by the addition of a pelB signal sequence to the N-terminus of the V_L domain for protein expression. The purification involved a two-step chromatographic process in which the periplasmic extract was taken forward for an initial anion exchange chromatography step before size exclusion chromatography (Figure 5.4). At each step of the expression and purification process, the presence of the ~12 kDa constructs were confirmed by SDS-PAGE analysis. The molecular weight and purity was confirmed using ESI-MS analysis (Appendix 7.4).

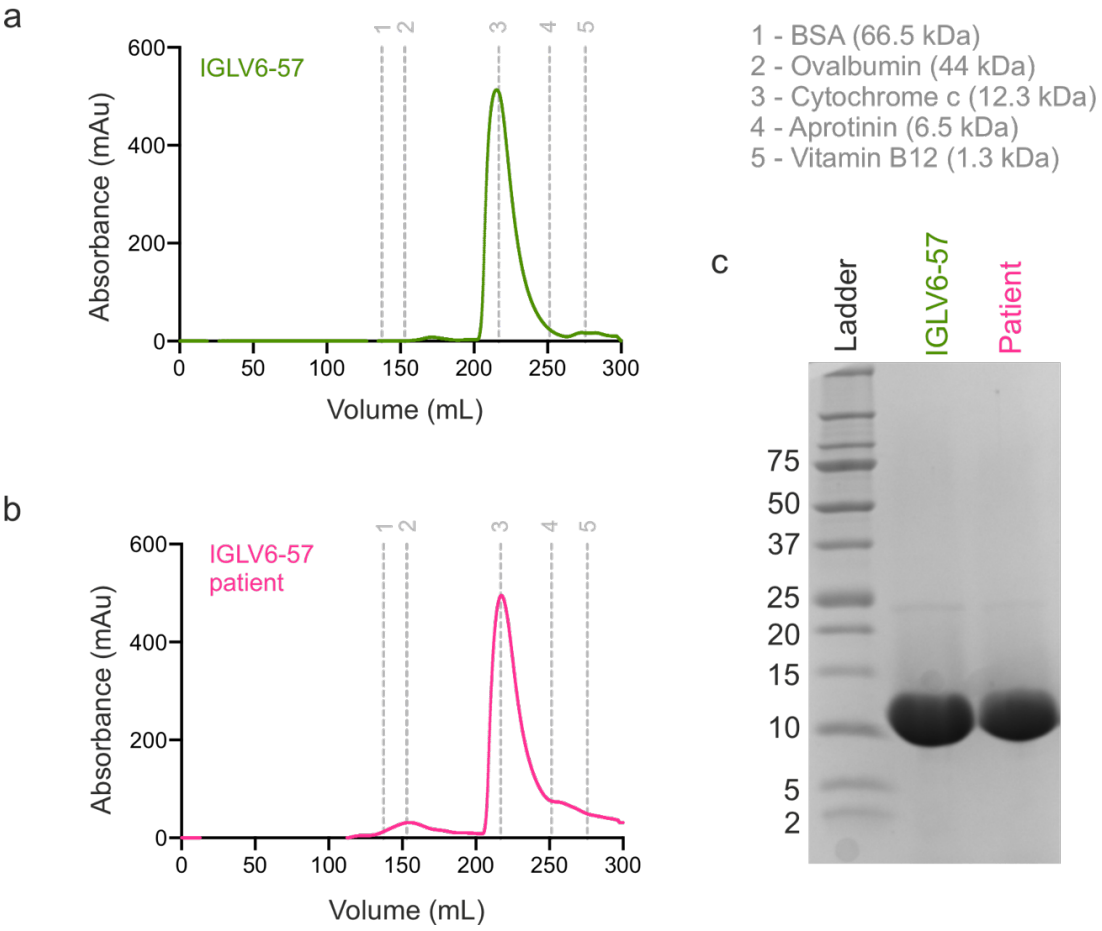


Figure 5.4 Purification of V_L domains. Size exclusion chromatography traces of a) germline IGLV6-57 and b) IGLV6-57 patient protein. Dashed lines (grey) represent standards of known molecular weight (listed in the top left). c) SDS-PAGE conformation of SEC peaks from a) and b) for each protein in comparison to the molecular weight marker (kDa).

To investigate the aggregation of the V_L domains a fluorescent dye Thioflavin T (ThT) was utilised. In an unbound state, the benzylamine and benzathiole rings rotate around their carbon-carbon bond, which quenches the excited state resulting in a low fluorescence emission (Figure 5.5). The binding of ThT to β-sheet-rich structures, such as the cross-β- sheet found in amyloid fibres, causes immobilisation of the benzylamine and benzathiole rings, preserving the exited state and results in enhanced fluorescence.

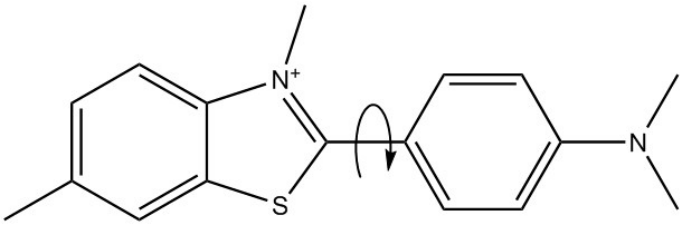


Figure 5.5 Structure of Thioflavin T. Benzylamine and benzathiole rings can rotate around the carbon-carbon bond.

Since the formation of amyloid fibres can occur on long-time scales and may involve protein unfolding, the presence of SDS was included in the ThT to accelerate this process[325]. The ThT fluorescence assay identified that both the germline and patient IGLV6-57 V_L aggregate *in vitro* (Figure 5.6a). The presence of fibrils was confirmed by transmission electron microscopy (TEM) for both proteins (Figure 5.6b). The fibrils formed from each V_L domain had distinct morphologies, IGLV6-57 formed short fibrils that clumped together whereas IGLV6-57 patient formed elongated twisted fibres.

To gain further insight into the difference between the germline and patient proteins, the thermal stability was monitored using far-UV circular dichroism (CD) (Section 2.7.9). The loss of secondary structure was monitored upon heating, and T_m at which 50 % of the protein were determined to be 50.5 °C for the IGLV6-57 germline and 43.3 °C for the IGLV6-57 patient protein (Figure 5.6c).

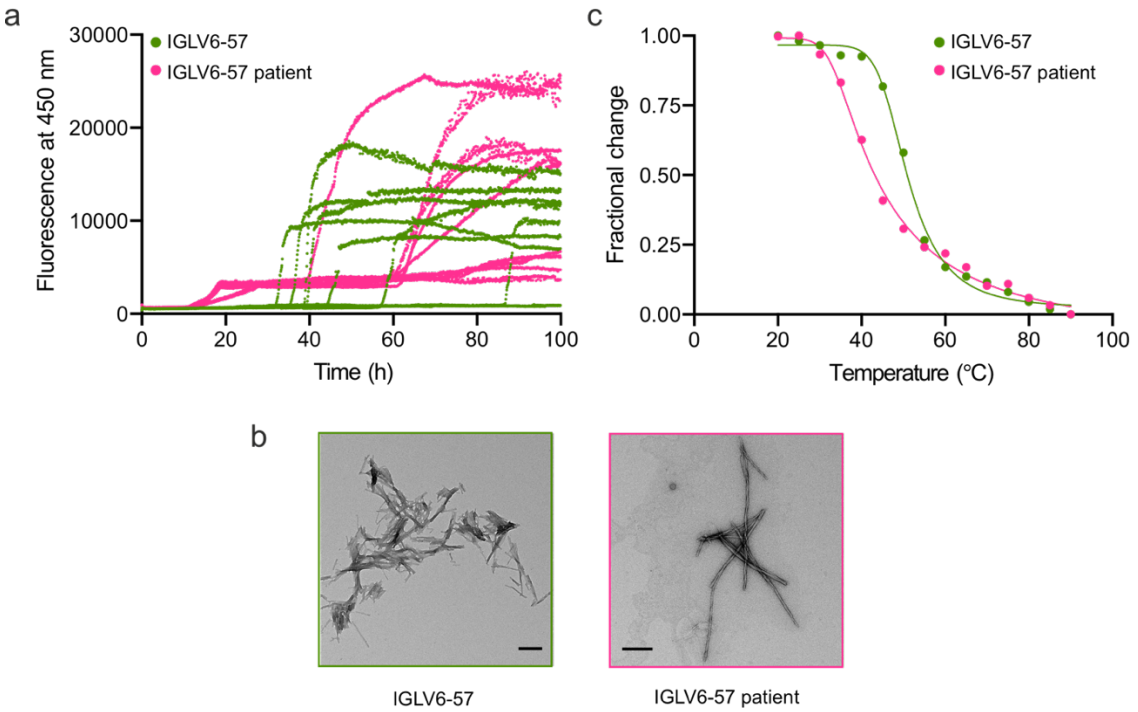


Figure 5.6 *In vitro* characterisation of IGLV6-57 germline and patient V_L domains. a) ThT fluorescence assay of IGLV6-57 germline (green) and patient (pink). Incubations were carried out at 37 °C, 600 rpm in the presence of 0.05 mM SDS. b) The corresponding TEM images were taken at the end points of the incubation from a) and are colour coded as in a). Scale bar = 200 nm. c) Temperature induced unfolding transitions of IGLV6-57 germline (green) and patient (pink). The T_m for IGLV6-57 germline and patient were calculated to be 50.5 °C and 43.3 °C respectfully. Full thermal denaturation CD spectra for the V_L domains can be found in Appendix 7.16.

The results from the *in vitro* characterisation of the V_L domains confirm that indeed the proteins aggregate as predicted by the TPBLA. The slight lower *in vivo* growth observed for the patient proteins in comparison to the germline may be due to a reduction in stability. The *in vitro* characterisation further highlights the complexity of understanding the consequences of the patient mutations in AL amyloidosis.

5.4 Identification of hotspots in germline and patient derived V_L domains

To understand the aggregation relationship between the germline and patient sequences, the four V_L domains were subjected to directed evolution screening in the TPBLA. Mutant libraries were synthesised using epPCR and golden gate assembly (Section 2.2.7) to create libraries estimated contain 1×10^9 clones. The libraries were

screened at an appropriate ampicillin concentration for each protein (IGLV6-57 220 μg/mL, IGLV6-57 patient 180 μg/mL, IGLV1-44 260 μg/mL and IGVL1-44 patient 220 μg/mL ampicillin) and the colonies from the selected plates were scraped into glycerol stocks. To allow for a greater analysis of mutations, next generation sequencing (NGS) was employed to sample a larger number of sequences in comparison to the Sanger sequencing approach taken in Chapters 3 and 4. Libraries for NGS were prepared to add Illumina adaptor sequences (Section 2.5.4) and sequenced using 2 × 250 bp Illumina MiSeq. The reads from NGS were filtered and aligned to the wildtype sequence to identify mutational hotspots for each protein (Section 2.8.3).

The resultant profiles identified striking differences between the four different proteins (Figure 5.7). Overall, both germline sequences had more residues mutated over the 2σ threshold in comparison to the patient samples and the majority of the mutations were located the C-terminus of the protein. For several V_L proteins, using hydrogen deuterium exchange and molecular dynamics, it has been found that the C-terminus is a highly dynamic region of the protein[300,316,325] which can decrease the domain stability. Evolution of C-terminal residues may decrease the conformational dynamics of the protein and reduce the aggregation propensity.

The germline IGLV1-44 V_L had eight residues identified through directed evolution as problematic (Figure 5.7a). All of the mutations were solvent exposed and located on the edges of the CDRs (Figure 5.7a). The C-terminus of the protein contained five hotspots, the substitutions that occurred at these positions generally reduced the amino acid hydrophobicity (Figure 5.8) and presumably reduce the aggregation propensity of this V_L domain by reducing the hydrophobic patch on the surface of the protein.

The patient IGLV1-44 protein had a very simplified mutational landscape in comparison to the germline sequence (Figure 5.7b). Only P7 was identified through evolution as a problematic residue and was found to be mutated to serine in the majority of sequences (Figure 5.9). The cis-trans alteration of P7 has been found to reshape the dimer interface and promote amyloid formation[326]. Other studies have also found that the amyloidogenicity of V_L domains are dependent upon a 'cryptic epitope' that occurs within the first 18 residues of the N-terminus and requires conformational rearrangement around a conserved proline at position 7[327,328]. Evolution of the neighbouring proline in this patient sample may reduce the rigidity of the N-terminus and disfavour misfolding around P7. In relation to the fibril structure observed for the IGLV1-44 patient (Figure 5.1), the N-terminus of the fibril

core starts at residue G15 and therefore the mechanism of P7S preventing amyloid formation could be explained by this cryptic epitope preventing aggregation.

For the second germline sequence, IGLV6-57, eleven hotspots were identified through directed evolution (Figure 5.7c and Figure 5.10). The profile identified by the TPBLA correlates with the oligomerisation hotspots identified for this germline sequence from relaxation dispersion NMR experiments for the conversion of the unfolded V_L domain into fibrils[316]. In contrast to the IGLV1-44 germline, this protein had mutations located in CDR1. The CDR1 for this sequence has previously been identified as a fibrillogenic hotspot for the protein through scanning proline mutagenesis and the generation of synthetic peptides[329]. This study also utilised a proteolysis-based strategy that identified T83 as a proteolytic site that generated amyloidogenic peptides[329]. This residue, T83, was also identified by the TPBLA through evolution (Figure 5.7c), therefore T83D may reduce proteolytic cleavage and therefore prevents cleavage of the β-lactamase construct in the TPBLA allowing *E. coli* expressing this mutant to grow in the presence of β-lactam antibiotics (Figure 5.10).

Despite their differences, the germline sequences contained a few similarities. Both IGLV1-44 and IGLV6-57 germline sequences contained a hotspot at phenylalanine position at F101 (IGLV1-44) and F102 (IGLV6-57) (Figure 5.7a and c). This residue has been shown to stabilise the V_L-V_L dimerization interface[309] as a protective mechanism of amyloid formation, however the formation of dimers in the TPBLA will cause the *E. coli* to become susceptible to ampicillin. Therefore, the mutations identified by the TPBLA at F101/F103 may perturb this V_L dimerization interface. Furthermore, P45 was also identified by the TPBLA in both germline sequences, however the role of P45S in reducing aggregation remains elusive.

The patient protein from IGLV6-57 contained fewer hotspots than its germline sequence, with only four residues identified over 2σ (Figure 5.7d). S28 was identified in the fibrillogenic CDR1 hotspot previously described for the germline sequence[329]. In the native structure, each of the four hotspots are solvent exposed (Figure 5.7d) and each of these residues are present in the amyloid core (Figure 5.2c). S13 and S28 both point into internal cavities, and their mutation to larger amino acids, S13T and S28N through directed evolution (Figure 5.11) may disfavour fibril formation. The other two hotspots H99 and G104, are located on the surface of the fibril. Through evolution these residues are mutated to H99Q and G104S which would alter the charge and polarity of the fibril surface. Interestingly, the substitution R24G present in the patient IGLV6-57 sequence, which is also found to be present in 25 % of IGLV6-57 mutant sequences[322,323], was not found to be mutated the patient sequence,

which may suggest this mutation is not the driving force for fibril formation for this patient protein.

Surprisingly, for both patient proteins, the mutations that were found in the patient relative to the germline protein were not altered during evolution (Figure 5.9 and Figure 5.11). This suggests that the effects of somatic hypermutation within the context of germlines causes non-trivial changes that are not caused by the introduction of more aggregation prone CDRs.

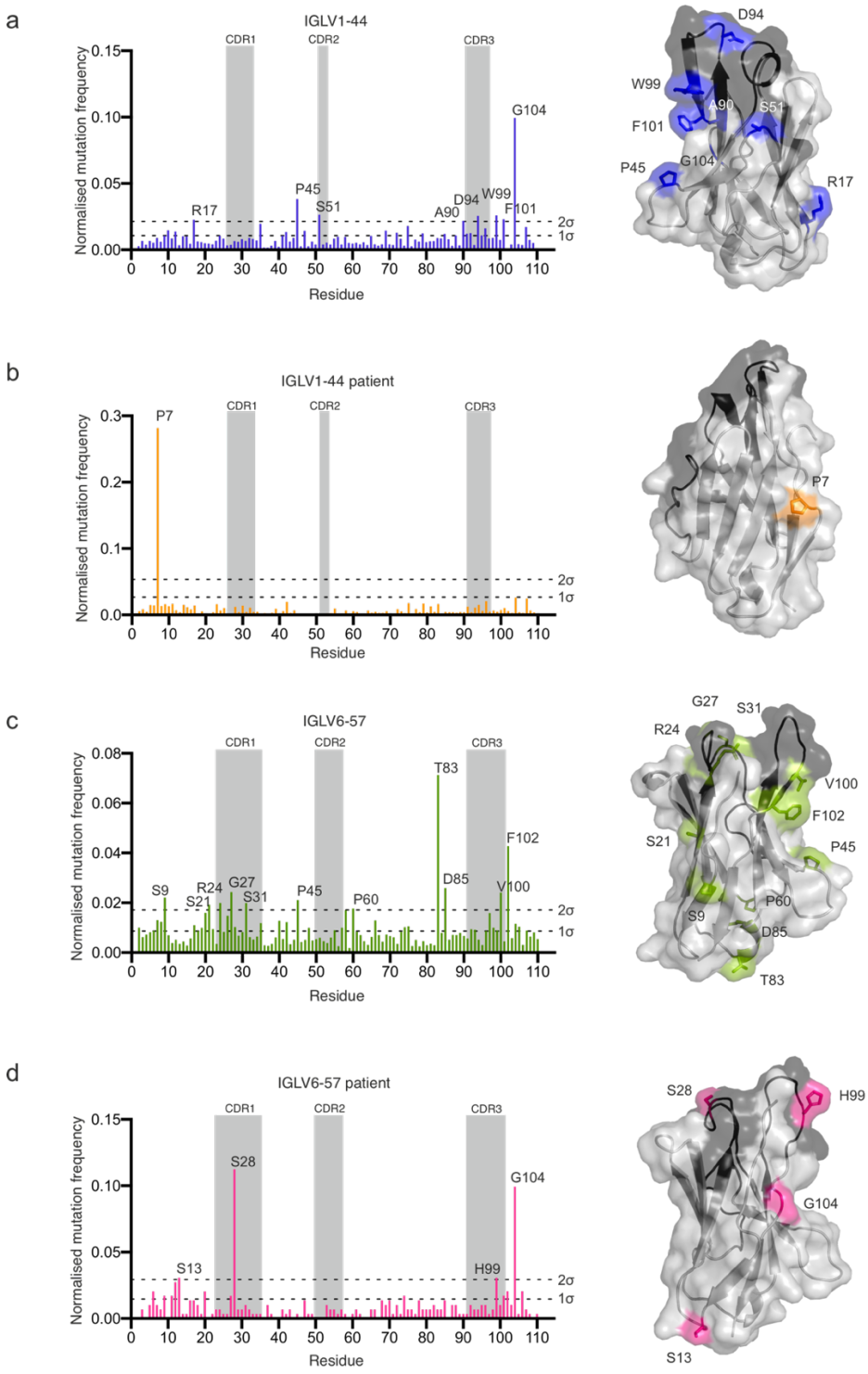


Figure 5.7 Mutational frequency profiles of evolved of V_L domains. Directed evolution of a) IGLV1-44 (15,846 sequences), b) IGLV1-44 patient (14,425 sequences), c) IGLV6-57 (17,724 sequences), and d) IGLV6-57 patient (16,187 sequences). Profiles are normalised to the sum of one. Residues over two standard deviations (2σ) are labelled on the frequency profile and native V_L structure. Grey boxes and dark grey on structure highlight the CDRs.

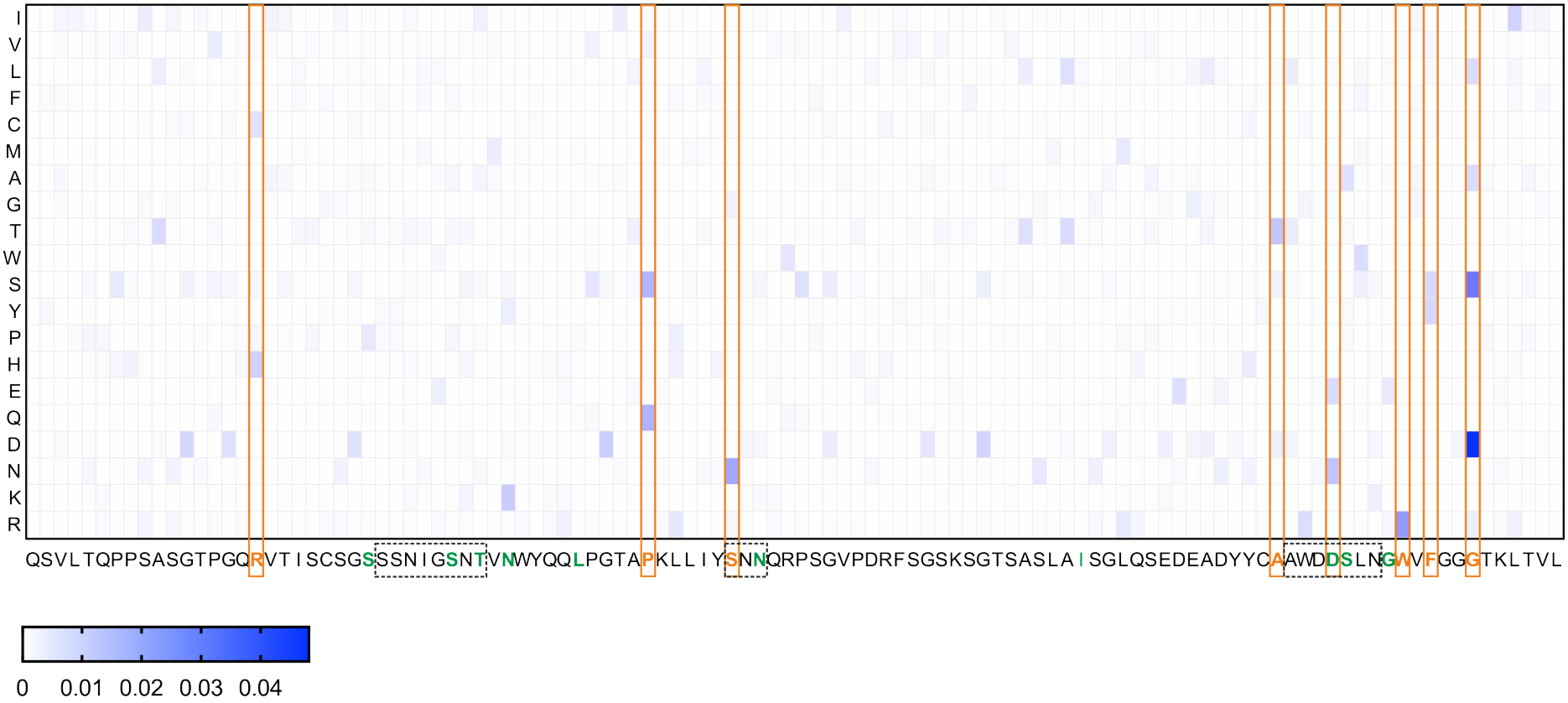


Figure 5.8 Mutational landscape of IGLV1-44 germline. Heatmap showing the identity of mutations from directed evolution of IGLV1-44. Wildtype sequence is shown with residues in green highlighting those mutated in the patient protein. Orange residues and bars highlight the residues over two standard deviations (2σ) from Figure 5.7a. Grey dash boxes highlight residues that form the CDRs. The vertical axis indicates the identity of the substitution and are ordered according to the Kyte-Doolittle scale[271]. Data are normalised to the sum of one.

Figure 5.9 Mutational landscape of IGLV1-44 patient. Heatmap showing the identity of mutations from directed evolution of IGLV1-44 patient. Wildtype sequence is shown with residues in green highlighting those mutated from the germline. Orange residues and bars highlight the residues over two standard deviations (2σ) from Figure 5.7b. Grey dash boxes highlight residues that form the CDRs. The vertical axis indicates the identity of the substitution and are ordered according to the Kyte-Doolittle scale[271]. Data are normalised to the sum of one.

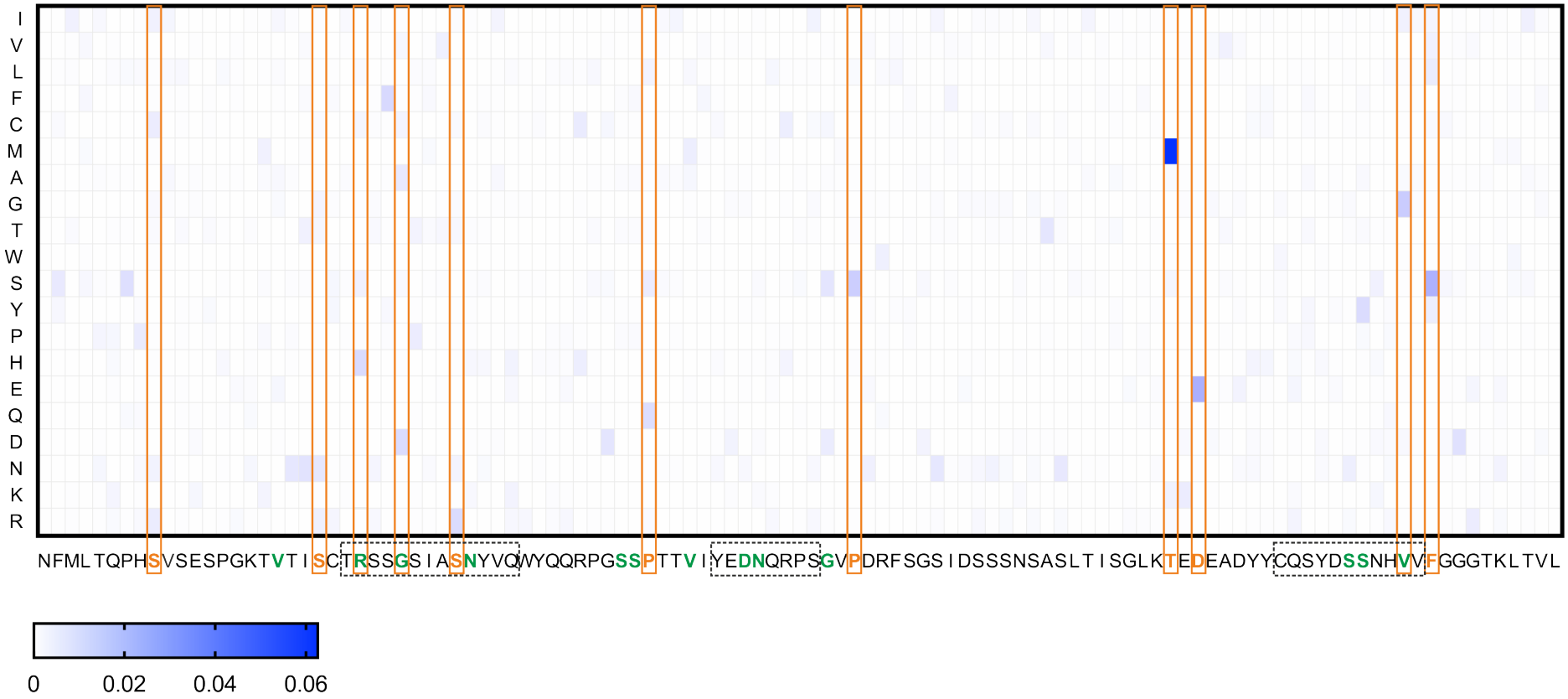


Figure 5.10 Mutational landscape of IGLV6-57 germline. Heatmap showing the identity of mutations from directed evolution of IGLV6-57. Wildtype sequence is shown with residues in green highlighting those mutated in the patient protein. Orange residues and bars highlight the residues over two standard deviations (2σ) from Figure 5.7c. Grey dash boxes highlight residues that form the CDRs. The vertical axis indicates the identity of the substitution and are ordered according to the Kyte-Doolittle scale[271]. Data are normalised to the sum of one.

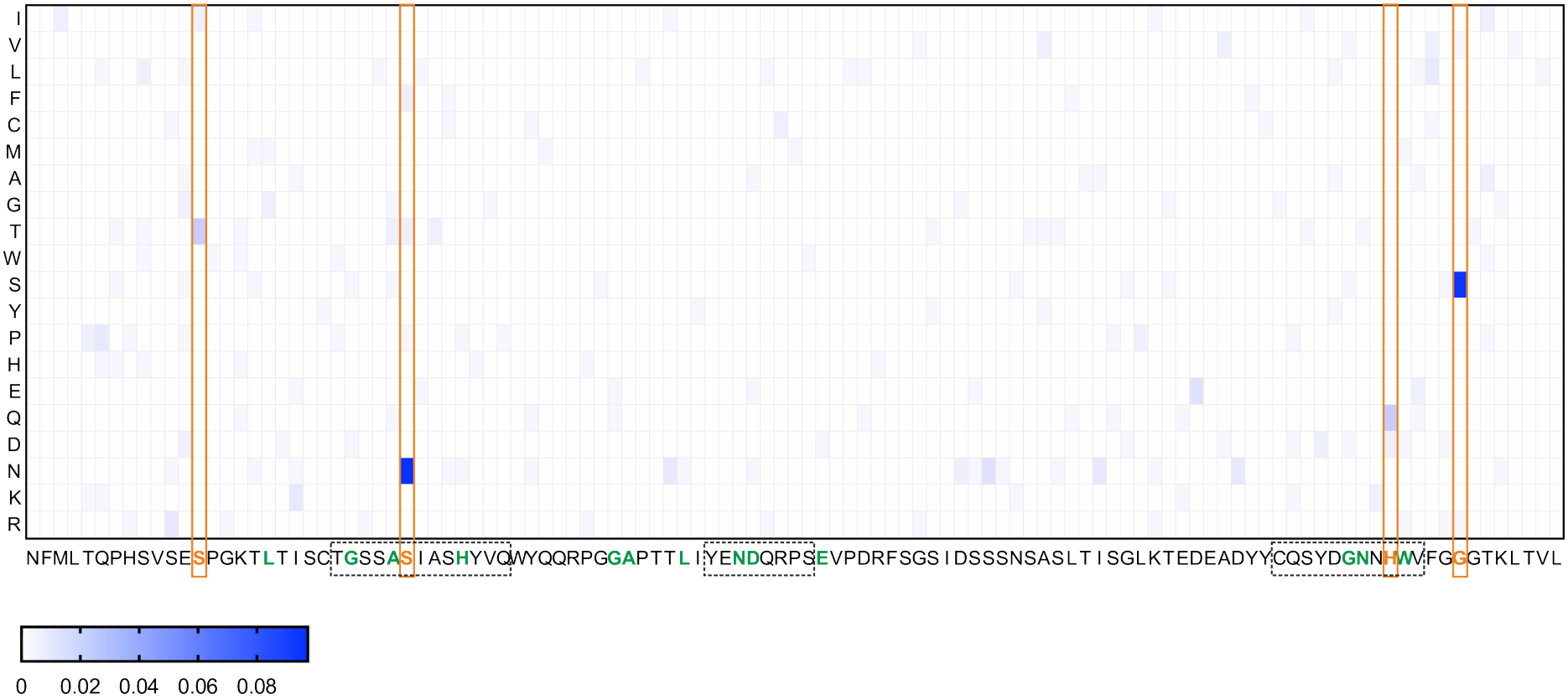


Figure 5.11 Mutation landscape of IGLV6-57 patient. Heatmap showing the identity of mutations from directed evolution of IGLV6-57 patient. Wildtype sequence is shown with residues in green highlighting those mutated from the germline. Orange residues and bars highlight the residues over two standard deviations (2σ) from Figure 5.7d. Grey dash boxes highlight residues that form the CDRs. The vertical axis indicates the identity of the substitution and are ordered according to the Kyte-Doolittle scale[271]. Data are normalised to the sum of one.

5.5 Discussion

AL amyloidosis is a complex disease, in which patients have unique protein sequences which can impede the diagnosis and treatment. The mechanism of aggregation in AL amyloidosis remains unclear and is hampered by the presence of various mutations in V_L domain of patients relative to the germline sequence. This chapter explored the use of the TPBLA as a novel method to identify aggregation prone residues within germline and patient V_L domains.

This study investigated two families of germline and patient samples from AL amyloidosis. The IGLV1-44 germline and patient differed by ten residues (90 % identity) and the IGLV6-57 germline and patient varied by thirteen mutations (88 % identity). The screening of these V_L domains in the TPBLA identified small differences between the *in vivo* growth of the patient sample in comparison to the germline sequence. The results from the TPBLA correlated with those attained *in vitro* in which both proteins formed amyloid as observed by ThT fluorescence and TEM. Although both proteins formed amyloid, the patient IGLV6-57 thermal stability was lowered by 7.2 °C relative to the germline sequence. Taken together, these results may suggest that germline sequences are inherently problematic resulting in a complex interplay of interactions responsible for the pathogenicity of patient proteins.

Each set of germline and patient protein were therefore taken forward for directed evolution studies to observe the similarities and differences in the mutational hotspots. In this chapter, NGS was employed to collect a larger sequencing dataset than can be achieved with Sanger sequencing as used in Chapters 3 and 4 (10,000 sequences by NGS in comparison to 100 sequences by Sanger sequencing). The resultant profiles from this experiment identified notable differences between the germline and patient proteins.

Evolution of the IGLV1-44 patient protein identified a single residue, P7, that was thought to be instrumental in the aggregation of this protein. Proline isomerisation at this position has previously been shown to control V_L dimerization[326], and conformational rearrangement that induces unfolding[327,328]. Interestingly, the mutational profile for the IGLV1-44 germline sequence did not identify this residue (P7) as problematic. Instead, many of the mutations were observed in the C-terminus of the protein, which reduce the hydrophobicity of the protein surface.

The IGLV6-57 patient and germline proteins had slight similarities in that mutations were observed in the CDR1, that has previously been identified to be the fibrillogenic region of the protien[329]. The mutational hotspots for this germline are in accord with work by Rennella *et al.*[316], through which the non-specific profile may reflect that

aggregation is driven from the unfolded state by interactions between APRs throughout the structure. In this case, the TBPLA may select for sequences with both decreased aggregation propensity and increased local or global thermodynamic stability (which decreases the population/lifetime of solvent exposed APRs).

Overall, the complex differences between the mutational landscapes identified suggest that each protein may have different limiting factor for aggregation and/or stability. Multiple selections may be occurring simultaneously during the evolution experiment whereby the hotspot residues identified may impact aggregation, stability or ones that destabilise the final fibril structure. Moreover, the naturally occurring mutations in the patient samples were not selected during directed evolution which may suggest that somatic hypermutation alters protein epistasis and conformational dynamics.

To further understand the mechanistic differences between germline and patient samples, a deep mutational scanning approach could be employed to observe the specific mutational effects of evolution. This method would involve a library containing mutations to every other amino acid at each residue of the protein. Through sampling a range of ampicillin conditions would aid the identification of key residues that rescue the protein from aggregation in the TPBLA.

In summary, the results presented in this chapter demonstrate that the TPBLA can be used for evolving disease relevant antibody domains. This provides a new method to understand the molecular mechanism for aggregation in AL amyloidosis. Furthermore, through combining the TPBLA with NGS thousands of mutational variants can be screened which has the potential aid the prediction of sequences that may cause AL amyloidosis.

Chapter 6
Concluding remarks and future directions

Understanding how and why proteins aggregate is of great importance to both the biopharmaceutical industry and for human disease[15,289,330]. For biopharmaceuticals, recombinant production of proteins often results in escalating expenditures due to the inherent nature of many proteins to aggregate, impinging the ability to produce lifesaving protein therapeutics rapidly and economically[16]. Additionally, protein aggregation pervades human disease and mortality with implications in more than 50 human diseases, including Alzheimer's, Parkinson's and type II diabetes[27]. This poses an ever-increasing risk in the developed world with an aging population having enormous social and economic burdens.

Understanding the molecular mechanisms of protein aggregation is challenging, given the array of competing interactions that control solubility, stability, cooperativity, and aggregation propensity. Various methods have been developed to interrogate protein aggregation, such as computational algorithms that can identify aggregation-prone regions, and biophysical assays to quantify aggregation[18]. Investigating protein aggregation and stability however can be laborious, due solely to the inability of the variants to be expressed and purified for *in vitro* analysis. There is therefore a need to be able to identify protein sequences that may have undesired properties and to engineer their sequences to improve their properties without the need for protein purification, to aid the rapid development of biopharmaceuticals and to further our understanding of protein aggregation.

To address this, this thesis developed a platform to characterise the aggregation propensity of candidate biopharmaceuticals that circumvents the need for recombinant expression of each variant. In the TPBLA, the test protein is inserted between two domains of the periplasmic-based reporter enzyme TEM-1 β-lactamase. Upon correct folding of the test protein in the periplasm of *E. coli*, the two halves of β-lactamase are brought into close proximity to form a functional enzyme, such that the bacteria are resistant to β-lactam antibiotics. If the test protein aggregates, however, the β-lactamase domains will be prevented from associating and the bacteria lose their resistance to the antibiotics. In contrast to other *in vivo* systems for studying protein aggregation, the fusion proteins are expressed in the oxidative periplasm of *E. coli*, allowing the correct formation of disulfide bonds such as those found in IgGs and their derivatives. Most importantly, no perturbant such as increased temperature,

pH or chemical denaturant is used to accelerate aggregation, allowing identification of sequence characteristics that trigger innate aggregation pathways.

In Chapter 3, three pairs of test proteins with varying degrees of aggregation-propensity were assessed in the TPBLA. This demonstrated that the TPBLA was able to differentiate between aggregation prone and non-aggregating protein variants for a range of diverse biopharmaceutically-relevant protein scaffolds. Moreover, the platform was found to have a high sensitivity to single mutations to scFv-WFL, that reflected the results observed for the full-length IgG variants *in vitro*.

As scFvs are commonly grafted into well characterised proprietary IgG scaffolds, and scFv formats are used in phage display, the TPBLA could be integrated into the development pipeline to identify developable sequences directly after discovery and affinity maturation. It could also be used to optimise a wide variety of biologics, including dAbs, scFabs, scFc and bispecifics (in scFv format) all of which are poorly characterised in terms of developability relative to platform IgGs.

The work in Chapter 3 also investigated the feasibility of using a directed evolution approach in combination with the TPBLA as a novel strategy to modulate the aggregation propensity of protein scaffolds. Sequencing of colonies revealed mutational hotspots within the protein sequence of scFv-WFL, that are presumably the residues that contribute to the aggregation. Therefore, in addition to passive screening, the TPBLA can be used as a directed evolution screen to identify mutational hotspots that limit the proteins behaviour. In the future, this directed evolution platform may allow for the rational design of inherently manufacturable IgGs, by identifying target residues to mutate reduce the aggregation potential.

The proteins that were identified through directed evolution were further explored in Chapter 4. A panel of evolved proteins were selected and characterised for their reduced aggregation *in vitro* that revealed the TPBLA had engineered new proteins with reduced aggregation as measured by HP-SEC and AC-SINS. Screening a randomised scFv library of an unrelated IgG sequence, the anti-LINGO1 antibody (Li33), identified variants with improved solubility. These data suggest that the TPBLA assay can be used to screen for a variety of limiting factors such as thermodynamic stability, protein self-association and/or protein solubility.

Unsurprisingly, given their importance in determining epitope binding affinity, the majority of hotspot residues identified by the TPBLA for both scFvs WFL and Li33 are located in, or close to, the CDRs. At first this may appear to be problematic to the maintenance of a successful candidate profile, however as observed in this study, binding was maintained concomitantly with a significant improvement in aggregation performance, at least for variants of IgG-WFL. To overcome this caveat, libraries

could be created for directed evolution that conserve the residues involved in binding to the antigen to maintain the binding affinity.

Despite the ability of the TPBLA to generate greatly improved candidate sequences, one disadvantage is that it can be difficult to understand, at a fundamental level, the underlying cause of aggregation and to rationalise the substitutions made. However, the analysis of the relatively large number of sequences by the TPBLA (relative to rational approaches), each containing multiple substitutions allows identification of more complex, multi-partite interactions which cannot be identified using standard mutational or *in silico* methods. The use of algorithms to predict the aggregation of scFv-WFL was investigated in Chapter 3, each of which yield different predictions, confusing the choice of residues to mutate in any rational approach to improve protein behaviour. While not allowing a molecular understanding, machine learning can potentially be used in the future to identify novel indicators of aggregation. With this in mind, generation of larger datasets using both negative and positive selection of diverse IgG sequences is underway and may in future aid the development of computational algorithms to predict aggregation.

The TPBLA has advantages beyond the application to industrial relevant IgG scaffolds, and the work in Chapter 5 investigated the use of this platform for disease-related antibody aggregation. The long-term goal of the work in this chapter was to facilitate understanding the mechanism of disease-related aggregating proteins. The low sequence identity between germline light chain domains and those implicated in light chain amyloidosis makes it difficult to predict which sequences are likely to cause disease. The study of germline and patient samples in the TPBLA identified that germline sequences are inherently problematic that may result in a complex interplay of interactions responsible for the pathogenicity of patient proteins. Directed evolution identified notable differences between the germline and patient proteins and the naturally occurring mutations in the patient samples were not selected during directed evolution. Combining the TPBLA with NGS allowed the screening of thousands of mutational variants which has the potential aid the prediction of sequences that may cause AL amyloidosis. Work is underway to explore the use of the TPBLA as a platform for deep mutational scanning, that will enable a broader understanding of the relationship between sequence and aggregation mechanism. Furthermore, the ability to rapidly survey the aggregation propensity of large numbers of highly homologous sequences using deep mutational scanning together with statistical and machine learning methods will guide future protein engineering experiments and again could be used for the development of new predictive algorithms.

In future, other disease-related proteins could be examined using the TPBLA to aid the understanding of the molecular mechanisms that underpin the disease. For example, the multitude of intramolecular interactions made by α-synuclein that modulate its aggregation propensity[200] render the identification of key residues to target therapeutic strategies extremely challenging.

The TPBLA assay developed in this thesis has been shown to rapidly predict protein aggregation, rectify problematic sequences and to identify mutational hotspots that limit protein behaviour. As the biopharmaceutical industry moves away from traditional IgG scaffolds to novel modalities, the TPBLA may be instrumental in predicting the aggregation propensity and manufacturability of these molecules, of which protein aggregation is uncharted territory. Finally, the TPBLA may aid the quest of understanding the molecular mechanism of protein aggregation that underpins disease, ultimately leading to earlier diagnosis in patients and new therapeutic strategies.

Chapter 7

Appendices

7.1 β-lactamase construct sequences and plasmid maps

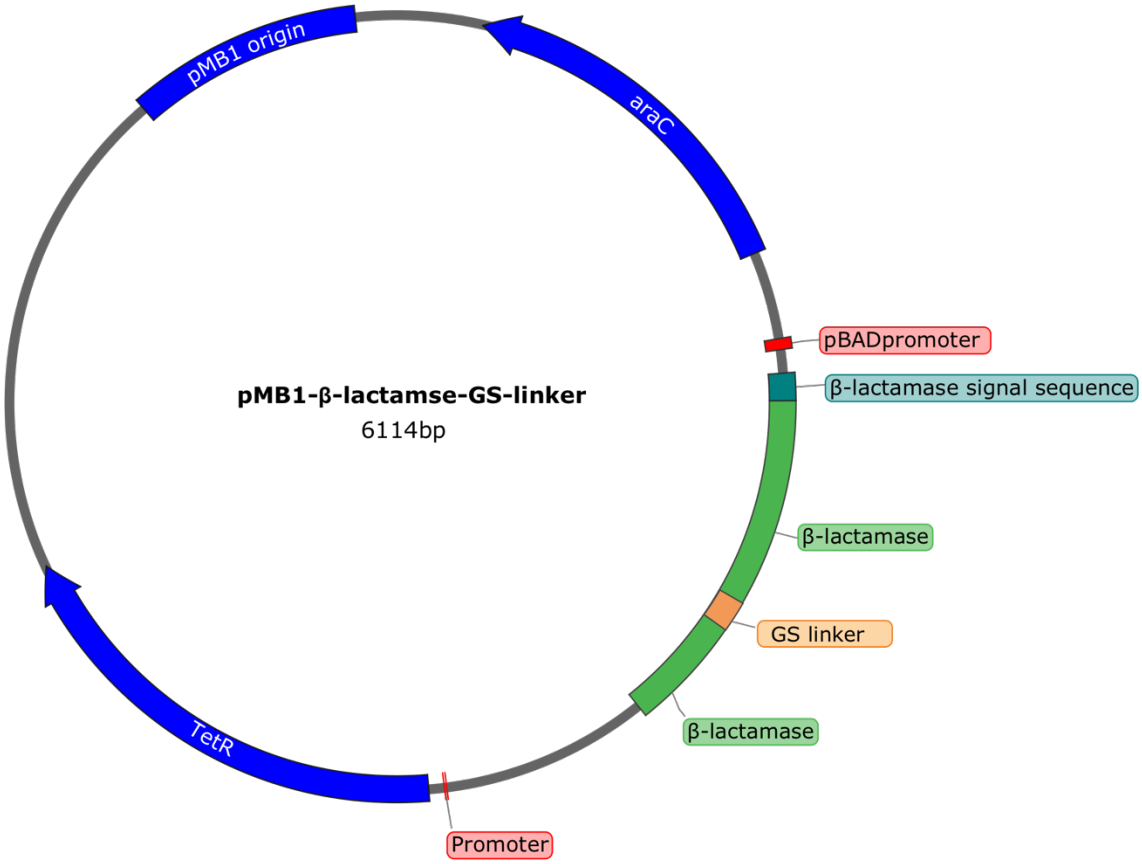
7.1.1 β-lactamase 28 GS linker

ATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCC
TGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTG
CACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGC
CCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATT
ATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATG
ACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGA
GAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGAC
AACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAA
CTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGAC
ACCACGATGCCTGCAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTA**GG**
TGGTGGTGGTTCTGGTGGTGGTGGCTCGAGCTCA**GGATCC****GGGAGCGGTTCCGGAAGCG**
GAGGAGGTGGTTCAGGCGGAGGTGGAAGCTTGACTCTAGCTAGCCGGCAGCAGCTCATA
GACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGG
CTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAG
CACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAG
GCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCA
TTGGTAA

Appendix 7.1 DNA sequence of β-lactamase 28 GS linker. The periplasmic signal sequence is shown in green. The GS linker is shown in bold and the *Xho*I and *Bam*HI restriction sites are shown in blue and orange respectfully. The start and stop codons are underlined.

MSIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAEDQLGARVGYIELDLNSGKILESFR
PEERFPMMSTFKVLLCGAVLSRVDAGQEQLGRRIHYSQNDLVEYSPVTEKHLTDGMTVR
ELCSAAITMSDNTAANLLLTTIGGPKELTAFLHNMGDHVTRLDRWEPELNEAIPNDERD
TTMPAAMATTLRKLLTGELG**GGGSGGGGSSSGSGSGSGSGGGGSGGGGS**LTLASRQQLI
DWMEADKVAGPLLRSALPAGWFIADKSGAGERGSRGIIAALGPDGKPSRIVVIYTTGSQ
ATMDERNRQIAEIGASLIKHW

Appendix 7.2 Protein sequence of β-lactamase 28 GS linker. The periplasmic signal sequence is shown in green. The GS linker is shown in bold. The signal sequence is cleaved after translocation into the periplasm.



Appendix 7.3 Plasmid map of pMB1-β-lactamase-GS-linker. Plasmid was kindly provided by Professor Jim Bardwell (University of Michigan, USA).

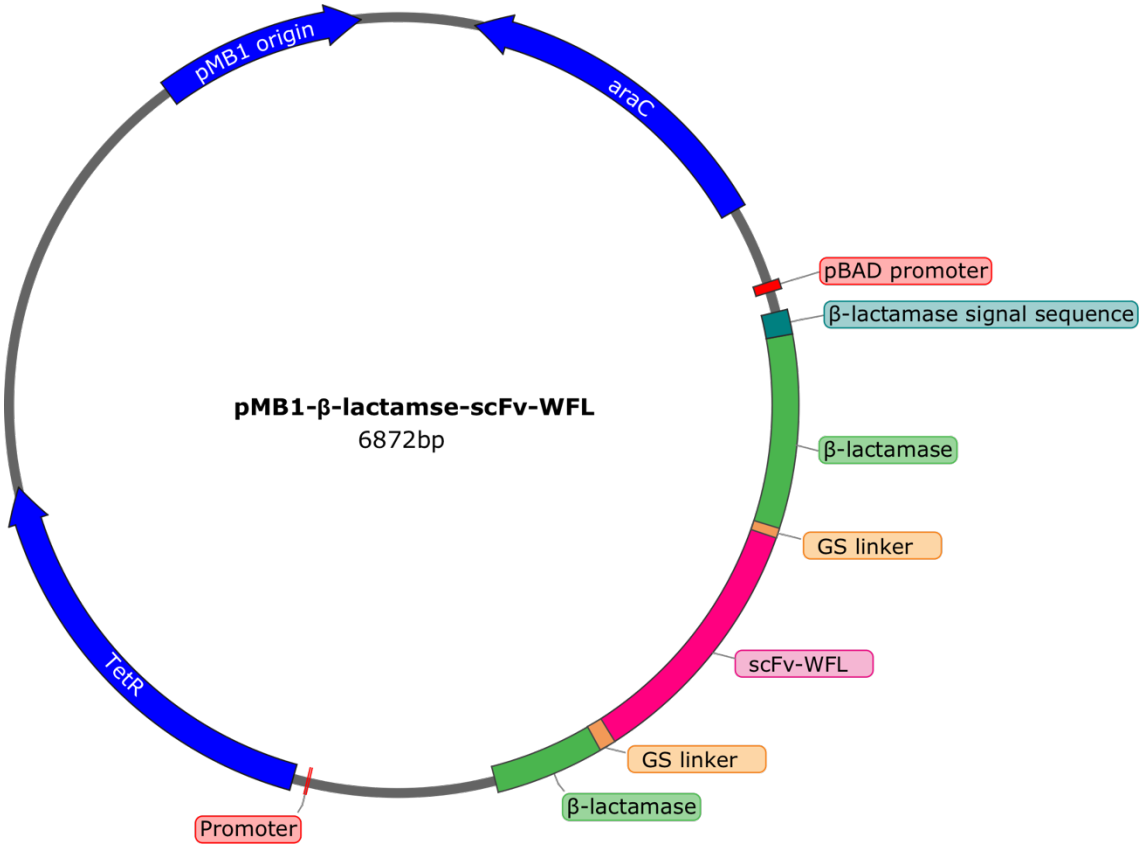
7.1.2 β-lactamase-scFv-WFL

ATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCC
TGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTG
CACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGC
CCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATT
ATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATG
ACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGA
GAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGAC
AACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAA
CTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGAC
ACCACGATGCCTGCAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTA**GG**
TGGTGGTGGTTCTGGTGGTGGTGGCTCGAGCCAGGTTCAGCTTGTGCAGAGCGGTGCGG
AGGTCAAAAAACCCGGCAGCTCTGTAAAAGTTAGCTGCAAAGCGAGTGGCGGTACGTTT
TGGTTTGGGGCCTTTACTTGGGTTCGTCAAGCGCCGGGCCAGGGCTTGGAATGGATGGG
TGGCATTATCCCTATTTTCGGCCTCACAAACCTGGCGCAAAACTTTCAAGGTCGCGTTA
CCATTACGGCGGACGAAAGCACCAGTACCGTCTATATGGAGCTGTCAAGCCTGCGCTCA
GAAGACACCGCAGTTTACTACTGTGCGCGTAGCAGCCGCATTTACGACTTGAATCCTAG
CCTCACAGCGTACTACGACATGGATGTGTGGGGGCAGGGCACCATGGTTACGGTGTCGA
GTGGTGGTGGGAGCAGTGGTGGAGGTGGGTCCGGGGGCGGCGGCGGCGCGCAAAGCGTA
TTAACTCAGCCGCCGAGCGTGAGCGCAGCCCCTGGGCAGAAAGTCACCATTTCATGCAG
CGGCTCCTCCAGCGATATCGGCAACAATTACGTGTCCTGGTATCAGCAGCTGCCTGGCA
CTGCGCCGAAGCTGTTGATTTATGACAACAATAAGCGTCCCTCGGGTATTCCAGATCGT
TTTTCTGGCTCTAAAAGCGGGACATCAGCGACACTGGGCATCACCGGGCTGCAGACGGG
GGATGAAGCCGATTATTACTGCGGGACCTGGGATAGTTCCCTGAGCGCGTGGGTGTTTG
GCGGGGGCACCAAACTCACCGTGCTG**GGATCC**GGGAGCGGTTCCGGAAGCGGAGGAGGT
GGTTCAGGCGGAGGTGGAAGCTTGACTCTAGCTAGCCGGCAGCAGCTCATAGACTGGAT
GGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTA
TTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGG
CCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTAT
GGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAA

Appendix 7.4 DNA sequence of β-lactamase-scFv-WFL. The periplasmic signal sequence is shown in green. The scFv-WFL sequence is shown in pink with the *Xho*I and *BamH*I restriction sites shown in blue and orange, respectively. The glycine-serine linker is highlighted in bold. Codons for residues W, F and L mutated to STT in β-lactamase-scFv-STT are highlighted in yellow. Start and stop codons are underlined.

MSIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAEDQLGARVGYIELDLNSGKILESFR
PEERFPMMSTFKVLLCGAVLSRVDAGQEQLGRRIHYSQNDLVEYSPVTEKHLTDGMTVR
ELCSAAITMSDNTAANLLLTTIGGPKELTAFLHNMGDHVTRLDRWEPELNEAIPNDERD
TTMPAAMATTLRKLLTGEL**GGGGSGGGGSS**QVQLVQSGAEVKKPGSSVKVSCKASGGTF
WFGAFTWVRQAPGQGLEWMGGIIPIFGLTNLAQNFQGRVTITADESTSTVYMELSSLRS
EDTAVYYCARSSRIYDLNPSLTAYYDMDVWGQGTMVTVSSGGGSSGGGGSGGGGGAQSV
LTQPPSVSAAPGQKVTISCSGSSSDIGNNYVSWYQQLPGTAPKLLIYDNNKRPSGIPDR
FSGSKSGTSATLGITGLQTGDEADYYCGTWDSSLSAWVFGGGTKLTVL**GSGSGSGSGGG
GSGGGGS**LTLASRQQLIDWMEADKVAGPLLRSALPAGWFIADKSGAGERGSRGIIAALG
PDGKPSRIVVIYTTGSQATMDERNRQIAEIGASLIKHW

Appendix 7.5 Protein sequence of β-lactamase-scFv-WFL. The periplasmic signal sequence is shown in green, which is cleaved after translocation into the periplasm. The GS linker is shown in bold. Residues W, F and L mutated to STT in β-lactamase-scFv-STT are highlighted in yellow. The heavy and light chains are shown in purple and pink, respectively.



Appendix 7.6 Plasmid map of pMB1-β-lactamase-scFv-WFL. Plasmid was kindly provided by Dr Janet Saunders (University of Leeds).

7.2 Protein expression vectors

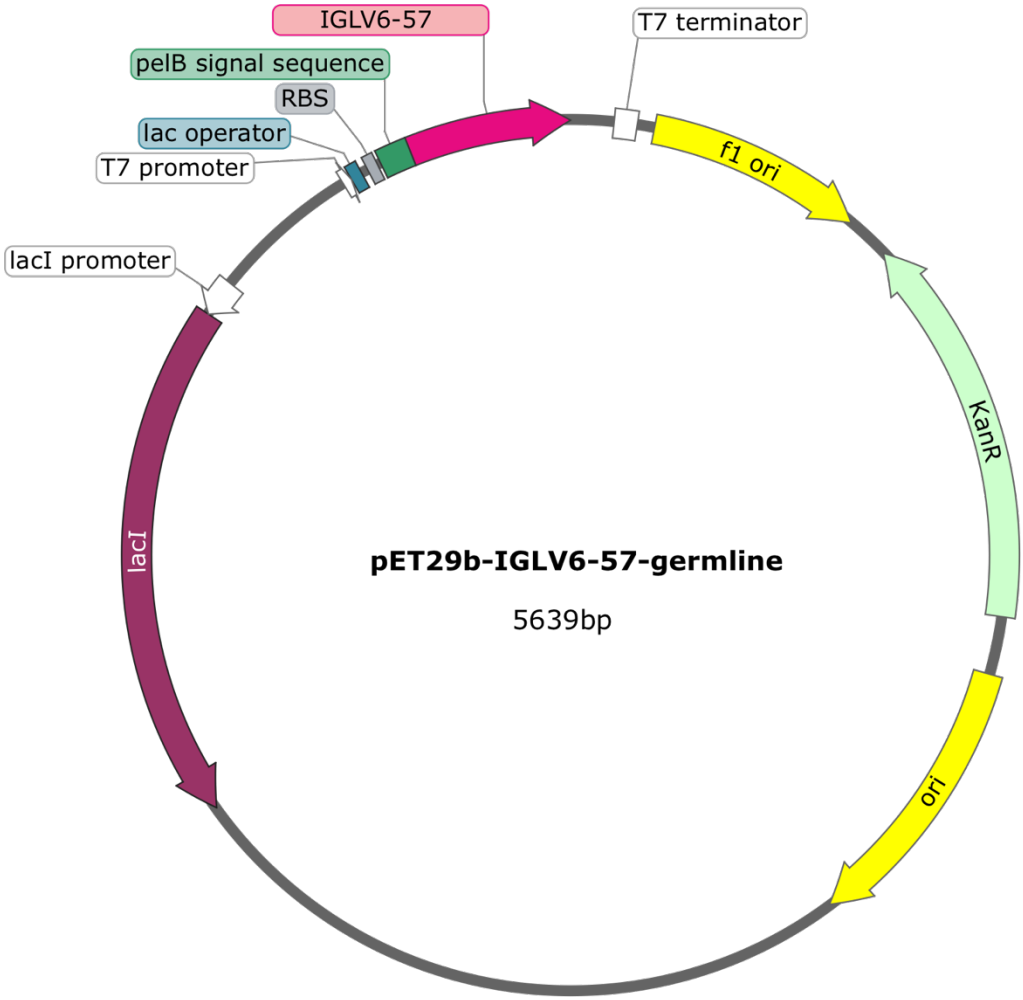
7.2.1 pET29b-IGLV6-57-germline

AAATACCTGCTGCCGACCGCTGCTGCTGGTCTGCTGCTCCTCGCTGCCCAGCCGGCGAT
GGCCATGAACTTTATGTTGACCCAGCCGCACAGTGTATCAGAATCTCCTGGAAAAACGG
TAACCATCAGCTGTACCCGCAGTTCTGGCTCAATTGCGAGCAACTACGTCCAGTGGTAC
CAACAGCGCCCAGGCTCCTCCCCGACCACCGTGATCTATGAAGACAACCAGCGTCCAAG
CGGTGTGCCCGATCGGTTTTCTGGCAGCATTGACAGTAGCAGTAACAGCGCCAGCTTGA
CCATCTCTGGACTTAAAACGGAAGATGAGGCGGACTATTATTGTCAATCCTATGATAGC
TCCAACCACGTCGTCTTTGGTGGCGGGACCAAGTTGACTGTTCTGTAA

Appendix 7.7 DNA sequence of pelB-IGLV6-57-germline. PelB signal sequence is shown in green. Stop codon is underlined.

KYLLPTAAAGLLLLAAQPAMAMNFMLTQPHSVSESPGKTVTISCTRSSGSIASNYVQWY
QQRPGSSPTTVIYEDNQRPSGVPDRFSGSIDSSSNSASLTISGLKTEDEADYYCQSYDS
SNHVVFGGGTKLTVL

Appendix 7.8 Protein sequence of pelB-IGLV6-57-germline. PelB signal sequence is shown in green which is cleaved after translocation to the periplasm.



Appendix 7.9 Plasmid map of pET29b-IGLV6-57-germline.

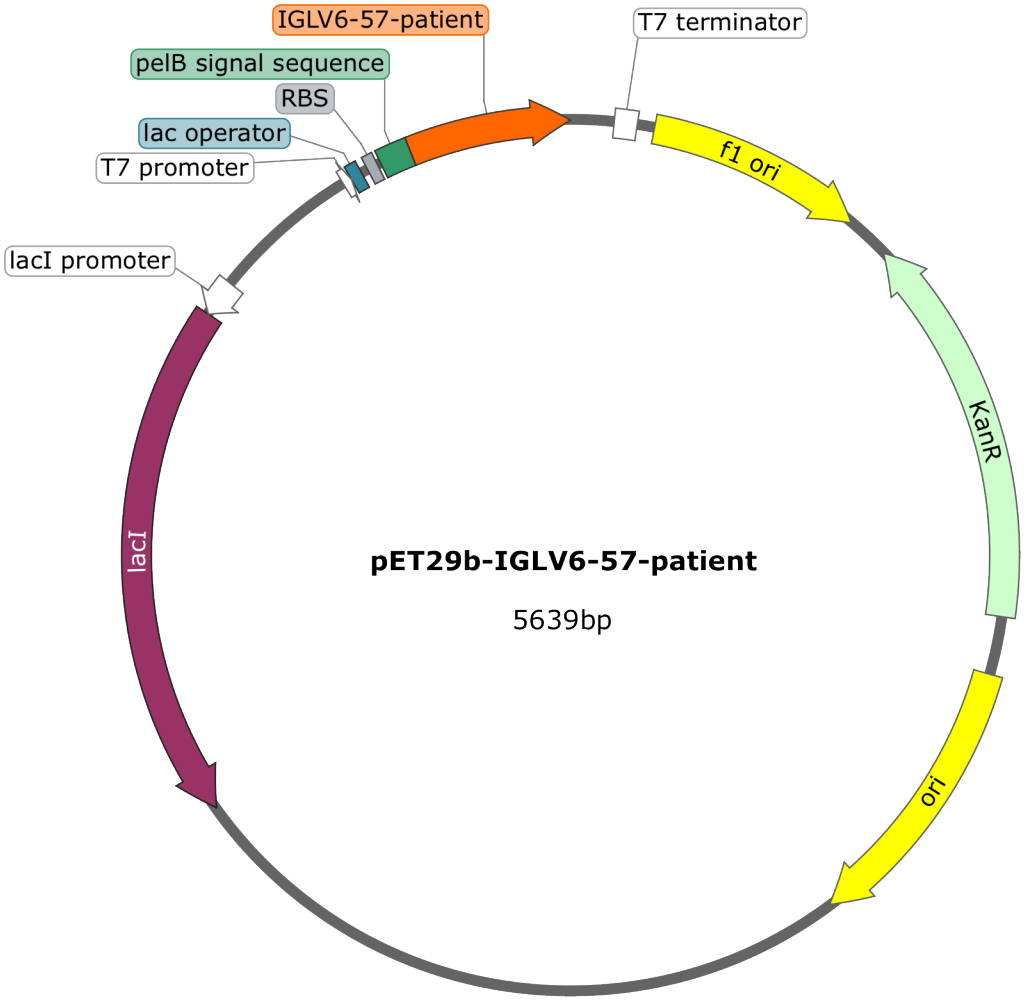
7.2.2 pET29b-IGLV6-57-pateint

AAATACCTGCTGCCGACCGCTGCTGCTGGTCTGCTGCTCCTCGCTGCCCAGCCGGCGAT
GGCCATGAATTTTATGTTGACACAGCCTCACTCGGTCAGCGAAAGCCCCGGAAAGACTC
TGACTATCTCTTGCACAGGCAGTTCGGCCAGCATCGCCTCCCACTATGTGCAATGGTAC
CAACAGCGTCCTGGTGGGGCTCCCACTACCCTCATTTACGAGAACGATCAGCGCCCGAG
TGAAGTTCCGGATCGCTTTTCCGGATCTATCGATTCCAGCAGTAATTCAGCGTCCCTGA
CCATTTCCGGCCTGAAAACGGAGGACGAAGCCGATTATTATTGCCAGTCATACGATGGT
AACAATCATTGGGTGTTCGGCGGCGGTACCAAATTAACTGTGCTGTAA

Appendix 7.10 DNA sequence of pelB-IGLV6-57-patient. PelB signal sequence is shown in green. Stop codon is underlined.

KYLLPTAAAGLLLLAAQPAMAMNFMLTQPHSVSESPGKTLTISCTGSSASIASHYVQWY
QQRPGGAPTTLIYENDQRPSEVPDRFSGSIDSSSNSASLTISGLKTEDEADYYCQSYDG
NNHWVFGGGTKLTVL

Appendix 7.11 Protein sequence of pelB-IGLV6-57-patient. PelB signal sequence is shown in green which is cleaved after translocation to the periplasm.



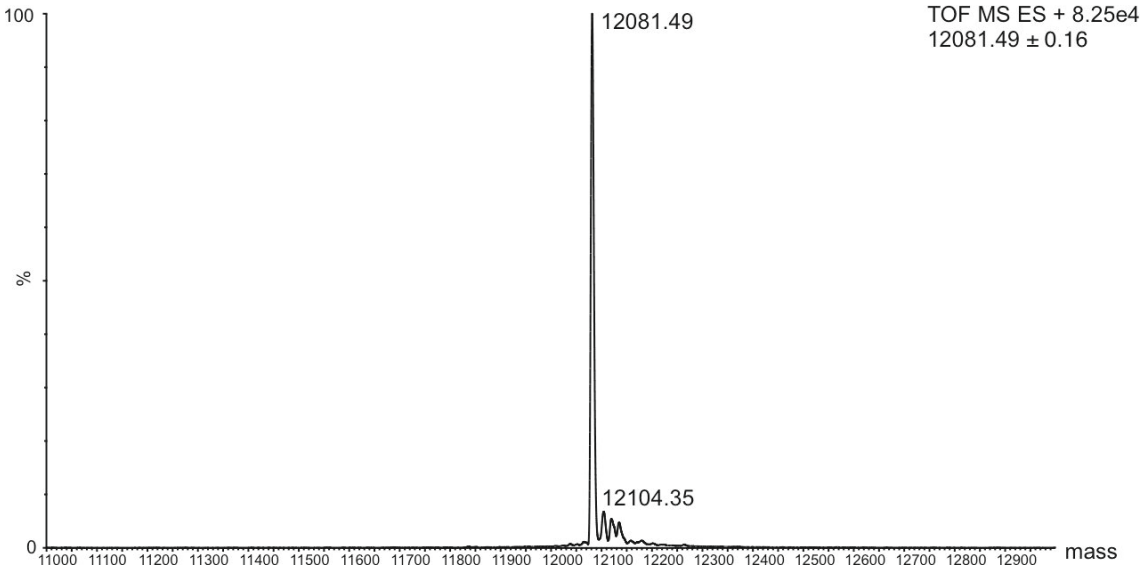
Appendix 7.12 Plasmid map of pET29b-IGLV6-57-patient.

7.3 Protein sequences

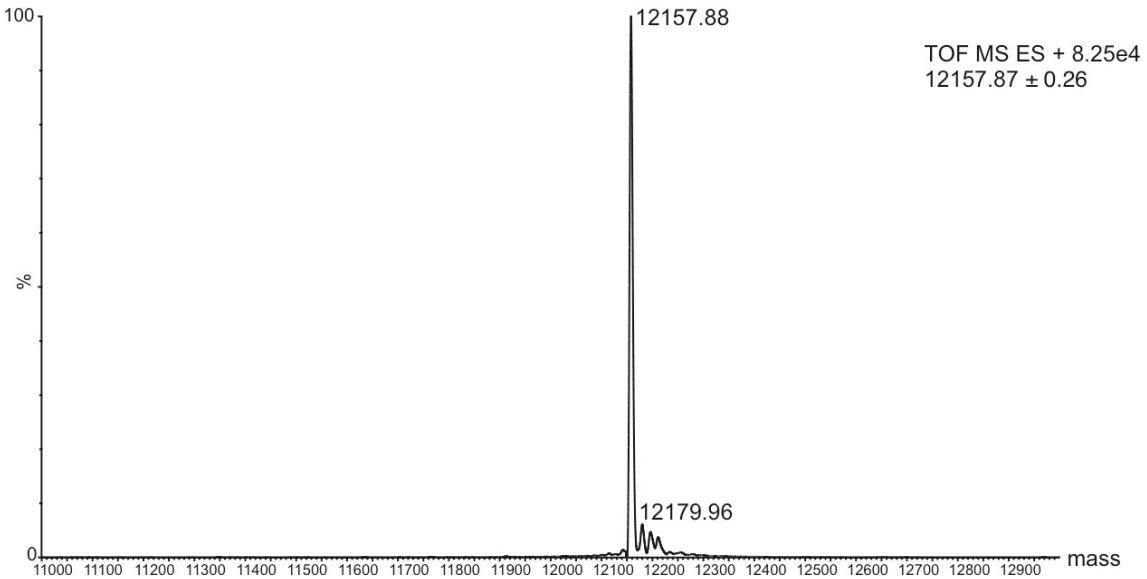
Construct	Amino acid sequence
scFv WFL	QVQLVQSGAEVKKPGSSVKVSCKASGGTFWFGAFTWVRQAPGQGLEWMG GIIPIFGLTNLAQNFQGRVTITADESTSTVYMELSSLRSEDTAVYYCAR SSRIYDLNPSLTAYYDMDVWGQGTMVTVSSGGGSSGGGGSGGGGGAQSV LTQPPSVSAAPGQKVTISCSGSSSDIGNNYVSWYQQLPGTAPKLLIYDN NKRPSGIPDRFSGSKSGTSATLGITGLQTGDEADYYCGTWDSSLSAWVF GGGTKLTVL
scFv STT	QVQLVQSGAEVKKPGSSVKVSCKASGGTFSTGAFTWVRQAPGQGLEWMG GIIPIFGLTNLAQNFQGRVTITADESTSTVYMELSSLRSEDTAVYYCAR SSRIYDLNPSLTAYYDMDVWGQGTMVTVSSGGGSSGGGGSGGGGGAQSV LTQPPSVSAAPGQKVTISCSGSSSDIGNNYVSWYQQLPGTAPKLLIYDN NKRPSGIPDRFSGSKSGTSATLGITGLQTGDEADYYCGTWDSSLSAWVF GGGTKLTVL
GCSF	MTPLGPASSLPQSFLLKCLEQVRKIQGDGAALQEKLCATYKLCHPEELV LLGHSLGIPWAPLSSCPSQALQLAGCLSQLHSGLFLYQGLLQALEGISP ELGPTLDTLQLDVADFATTIWQQMEELGMAPALQPTQGAMPAFASAFQR RAGGVLVASHLQSFLEVSYRVLRHLAQP
GCSF-C3	MTPLGPASSLPQSFLLKGLEQVRKIQGDGAALQEKLCATYKLCHPEELV LLGHSLGIPRAPLSSCPSQALRLAGCLSQLHSGLLLYQGLLQALEGISP ELGPTLDTLQLDVADFATTIWQQMEELGMAPALQPTQGAMPAFASAFQR RAGGVLVASHLQSFLEVSYRVLRHLAQP
Dp47d	EVQLLESGGGLVQPGGSLRLSCAASGFTFSSYAMSWVRQAPGKGLEWVS AISGSGGSTYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAK SYGAFDYWGQGTLVTVSS
HEL4	EVQLLESGGGLVQPGGSLRLSCAASGFRISDEDMGWVRQAPGKGLEWVS SIYGPSGSTYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAS ALEPLSEPLGFWGQGTLVTVSS
scFv Li33	EVQLLESGGGLVQPGGSLRLSCAASGFTFSIYPMFWVRQAPGKGLEWVS WIGPSGGITKYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTATYYCAR EGHNDWYFDLWGRGTLVTVSSGGGGSGGGGSGGGGSGGGGSDIQMTQSP GTLSLSPGERATLSCRASQSVSSYLAWYQQKPGQAPRLLIYDASNRATG IPARFSGSGSGTEFTLTISSLQSEDFAVYYCQQYDKWPLTFGGGTKVEI K
IGLV1-44-germline	QSVLTQPPSASGTPGQRVTISCSGSSSNIGSNTVNWYQQLPGTAPKLLI YSNNQRPSGVPDRFSGSKSGTSASLAISGLQSEDEADYYCAAWDDSLNG WVFGGGTKLTVL
IGLV1-44-patient	QSVLTQPPSASGTPGQRVTISCSGRSSNIGRNLVKWYQQFPGTAPKLLI YSNDQRPSGVPDRFSGSKSGTSASLAVSGLQSEDEADYYCAAWDATLNA WVFGGGTKLTVL
IGLV6-57-germline	NFMLTQPHSVSESPGKTVTISCTRSSGSIASNYVQWYQQRPGSSPTTVI YEDNQRPSGVPDRFSGSIDSSSNSASLTISGLKTEDEADYYCQSYDSSN HVVFGGGTKLTVL
IGLV6-57-patient	NFMLTQPHSVSESPGKTLTISCTGSSASIASHYVQWYQQRPGGAPTTLI YENDQRPSEVPDRFSGSIDSSSNSASLTISGLKTEDEADYYCQSYDGNN HWVFGGGTKLTVL

Appendix 7.13 Protein sequences used in this study.

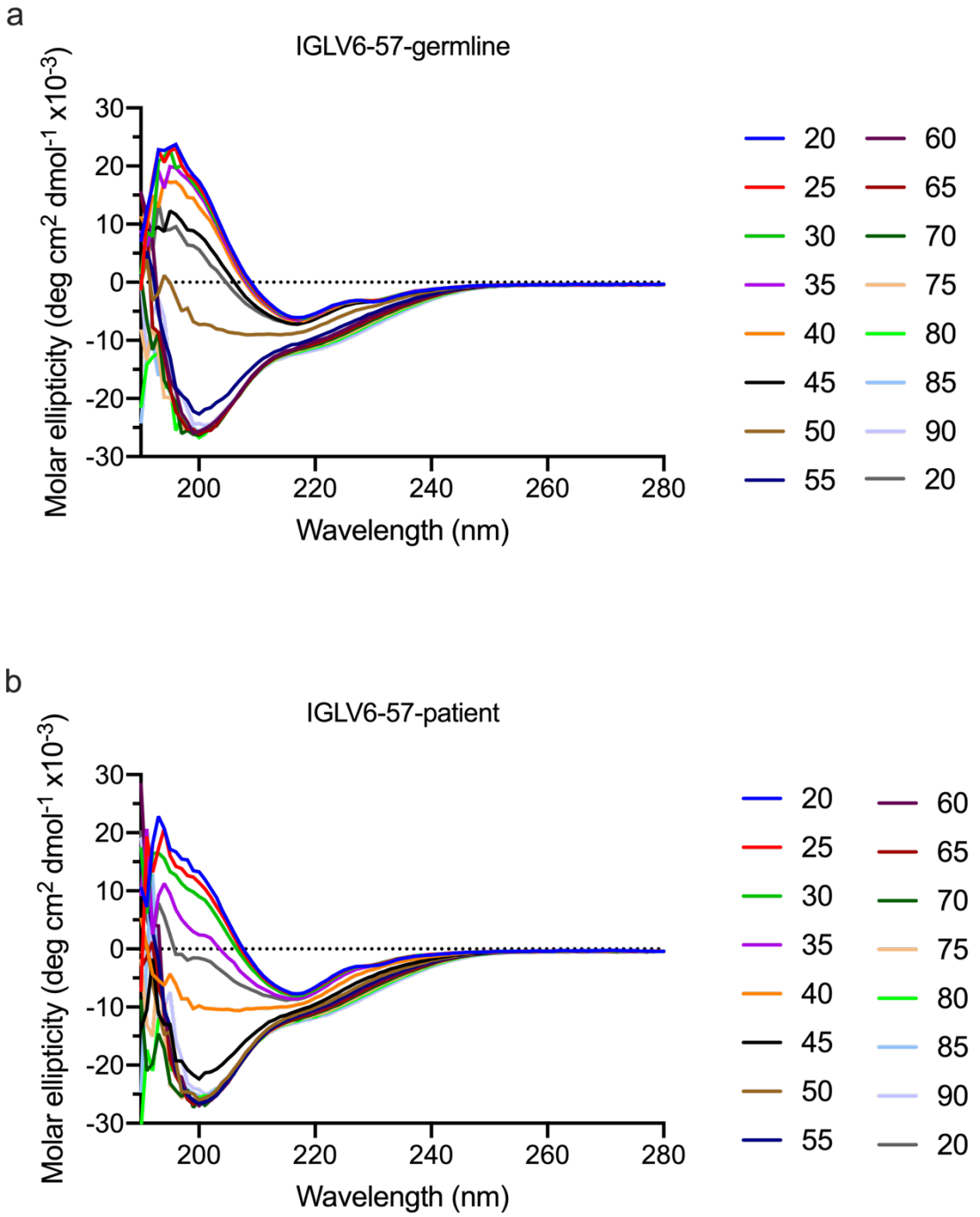
7.4 Mass spectrometry



Appendix 7.14 Mass spectrum of IGLV6-57-germline V_L domain. Expected molecular mass = 12082.16 Da. Measured mass = 12081.49 Da. Data collected by The Biomolecular Mass Spectrometry Facility (University of Leeds).



Appendix 7.15 Mass spectrum of IGLV6-57 patient VL domain. Expected molecular mass = 12158.26 Da. Measured mass = 12158.26 Da. Data collected by The Biomolecular Mass Spectrometry Facility (University of Leeds).



Appendix 7.16 Full CD spectra of V_L domains during thermal denaturation. Thermal denaturation was performed by setting up a gradient from 20-90 ºC in 5 ºC steps (colours in key) for a) IGLV6-57 germline and b) IGLV6-57 patient. Data from 210 nm was used to calculate the fractional change in Figure 5.6c.

References

1. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The protein folding problem. *Annu. Rev. Biophys.* **37**, 289–316 (2008).

2. Dill, K. A. & MacCallum, J. L. The protein-folding problem, 50 years on. *Science* vol. 338 1042–1046 (2012).

3. Anfinsen, C. B., Haber, E., Sela, M. & White, F. H. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.* **47**, 1309–14 (1961).

4. Levinthal, C. How to fold graciously. in *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois.* 22–24 (University of Illinois Press, 1969).

5. Zwanzig, R., Szabo, A. & Bagchi, B. Levinthal's paradox. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 20–2 (1992).

6. Dill, K. A. & Chan, H. S. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**, 10–19 (1997).

7. Jahn, T. R. & Radford, S. E. Folding versus aggregation: Polypeptide conformations on competing pathways. *Arch. Biochem. Biophys.* **469**, 100–

117 (2008).

8. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).

9. Brockwell, D. J. & Radford, S. E. Intermediates: ubiquitous species on folding energy landscapes? *Current Opinion in Structural Biology* vol. 17 30–37 (2007).

10. Jahn, T. R. & Radford, S. E. The Yin and Yang of protein folding. *FEBS J.* **272**, 5962–70 (2005).

11. Chiti, F. & Dobson, C. M. Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* **75**, 333–366 (2006).

12. Chi, E. Y., Krishnan, S., Randolph, T. W. & Carpenter, J. F. Physical stability of proteins in aqueous solution: Mechanism and driving forces in nonnative protein aggregation. *Pharmaceutical Research* vol. 20 1325–1336 (2003).

13. Englander, S. W. & Mayne, L. The nature of protein folding pathways. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 15873–15880 (2014).

14. Dobson, C. M. Protein folding and misfolding. *Nature* **426**, 884–90 (2003).

15. Wang, W. & Roberts, C. J. Protein aggregation – Mechanisms, detection, and control. *Int. J. Pharm.* **550**, 251–268 (2018).

16. Roberts, C. J. Therapeutic protein aggregation: mechanisms, design, and control. *Trends Biotechnol.* **32**, 372–380 (2014).

17. Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **22**, 1302–1306 (2004).

18. Meric, G., Robinson, A. S. & Roberts, C. J. Driving forces for nonnative protein aggregation and approaches to predict aggregation-prone regions. *Annu. Rev. Chem. Biomol. Eng.* **8**, 139–159 (2017).

19. Sassano, M. F., Doak, A. K., Roth, B. L. & Shoichet, B. K. Colloidal aggregation causes inhibition of G protein-coupled receptors. *J. Med. Chem.* **56**, 2406–2414 (2013).

20. Brummitt, R. K. *et al.* Nonnative aggregation of an IgG1 antibody in acidic conditions: Part 1. Unfolding, colloidal interactions, and formation of high-molecular-weight aggregates. *J. Pharm. Sci.* **100**, 2087–2103 (2011).

21. Austerberry, J. I. *et al.* Arginine to lysine mutations increase the aggregation stability of a single-chain variable fragment through unfolded-state interactions. *Biochemistry* **58**, 3413–3421 (2019).

22. Esfandiary, R., Parupudi, A., Casas-Finet, J., Gadre, D. & Sathish, H. Mechanism of reversible self-asssociation of a monoclonal antibody: role of electrostatic and hydrophobic interactions. *J. Pharm. Sci.* **104**, 577–586 (2015).

23. Matsui, D., Nakano, S., Dadashipour, M. & Asano, Y. Rational identification of aggregation hotspots based on secondary structure and amino acid hydrophobicity. *Sci. Rep.* **7**, 1–12 (2017).

24. Knowles, T. P. J., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **15**, 384–396 (2014).

25. Chatani, E. & Yamamoto, N. Recent progress on understanding the mechanisms of amyloid nucleation. *Biophysical Reviews* vol. 10 527–534 (2018).

26. Törnquist, M. *et al.* Secondary nucleation in amyloid formation. *Chem. Commun.* **54**, 8667–8684 (2018).

27. Dobson, C. M., Knowles, T. P. J. & Vendruscolo, M. The amyloid phenomenon and its significance in biology and medicine. *Cold Spring Harb. Perspect. Biol.* **12**, a033878 (2020).

28. Iadanza, M. G., Jackson, M. P., Hewitt, E. W., Ranson, N. A. & Radford, S. E. A new era for understanding amyloid structures and disease. *Nat.*

Rev. Mol. Cell Biol. **19**, 755–773 (2018).

29. Jackson, M. P. & Hewitt, E. W. Why are functional amyloids non-toxic in humans? *Biomolecules* vol. 7 (2017).

30. Ulamec, S. M., Brockwell, D. J. & Radford, S. E. Looking beyond the core: the role of flanking regions in the aggregation of amyloidogenic peptides and proteins. *Front. Neurosci.* **14**, 1216 (2020).

31. Maji, S. K. *et al.* Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. *Science (80-.).* **325**, 328–332 (2009).

32. Roan, N. R. *et al.* Peptides released by physiological cleavage of semen coagulum proteins form amyloids that enhance HIV infection. *Cell Host Microbe* **10**, 541–550 (2011).

33. Walsh, G. *Biopharmaceuticals: Biochemistry and biotechnology second edition*. *Wiley* (2003). doi:10.1590/s1516-93322005000200017.

34. Leader, B., Baca, Q. J. & Golan, D. E. Protein therapeutics: a summary and pharmacological classification. *Nat. Rev. Drug Discov.* **7**, 21–39 (2008).

35. Jackson, D. A., Symons, R. H. & Berg, P. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose

operon of Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **69**, 2904–2909 (1972).

36. Cohen, S. N., Chang, A. C. Y., Boyer, H. W. & Helling, R. B. Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 3240–3244 (1973).

37. Goeddel, D. V *et al.* Expression in Escherichia coli of chemically synthesized genes for human insulin. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 106–10 (1979).

38. Human insulin receives FDA approval. *FDA Drug Bull.* **12**, 18–19 (1982).

39. Köhler, G. & Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **256**, 495–497 (1975).

40. Leavy, O. Therapeutic antibodies: past, present and future. *Nat. Rev. Immunol.* **10**, 297–297 (2010).

41. Walsh, G. Biopharmaceutical benchmarks 2018. *Nat. Biotechnol.* **36**, 1136–1145 (2018).

42. Behring, E. von & Kitasato, S. Ueber das zustandekommen der diphtherie-immunität und der tetanus-immunität bei thieren. *Dtsch. Medizinische Wochenschrift* **49**, 1113–1114 (1980).

43. Chaplin, D. D. Overview of the immune response. *J. Allergy Clin. Immunol.* **125**, S3 (2010).

44. Schroeder, H. W., Cavacini, L. & Cavacini, L. Structure and function of immunoglobulins. *J. Allergy Clin. Immunol.* **125**, S41-52 (2010).

45. Vidarsson, G., Dekkers, G. & Rispens, T. IgG subclasses and allotypes: From structure to effector functions. *Front. Immunol.* **5**, (2014).

46. Dobson, C. L. *et al.* Engineering the surface properties of a human monoclonal antibody prevents self-association and rapid clearance in vivo. *Sci. Rep.* **6**, 38644 (2016).

47. Fields, C. *et al.* Creation of recombinant antigen-binding molecules derived from hybridomas secreting specific antibodies. *Nat. Protoc.* **8**, 1125–1148 (2013).

48. Wurch, T., Pierre, A. & Depil, S. Novel protein scaffolds as emerging therapeutic proteins: from discovery to clinical proof-of-concept. *Trends Biotechnol.* **30**, 575–582 (2012).

49. Better, M., Chang, C., Robinson, R. & Horwitz, A. Escherichia coli secretion of an active chimeric antibody fragment. *Science (80-.).* **240**, (1988).

50. Bird, R. E. *et al.* Single-chain antigen-binding proteins. *Science (80-.).* **242**,

423–426 (1988).

51. Bates, A. & Power, C. A. David vs. Goliath: The structure, function, and clinical prospects of antibody fragments. *Antibodies* **8**, 28 (2019).

52. Todorovska, A. *et al.* Design and application of diabodies, triabodies and tetrabodies for cancer targeting. *Journal of Immunological Methods* vol. 248 47–66 (2001).

53. Ward, E. S., Güssow, D., Griffiths, A. D., Jones, P. T. & Winter, G. Binding activities of a repertoire of single immunoglobulin variable domains secreted from Escherichia coli. *Nature* **341**, 544–546 (1989).

54. Hamers-Casterman, C. *et al.* Naturally occurring antibodies devoid of light chains. *Nature* **363**, 446–448 (1993).

55. Greenberg, A. S. *et al.* A new antigen receptor gene family that undergoes rearrangement and extensive somatic diversification in sharks. *Nature* **374**, 168–173 (1995).

56. Muyldermans, S., Atarhouch, T., Saldanha, J., Barbosa, J. A. R. G. & Hamers, R. Sequence and structure of VH domain from naturally occurring camel heavy chain immunoglobulins lacking light chains. *Protein Eng. Des. Sel.* **7**, 1129–1135 (1994).

57. Bannas, P., Hambach, J. & Koch-Nolte, F. Nanobodies and nanobody-

based human heavy chain antibodies as antitumor therapeutics. *Frontiers in Immunology* vol. 8 1603 (2017).

58. Frenzel, A., Hust, M. & Schirrmann, T. Expression of recombinant antibodies. *Front. Immunol.* **4**, 217 (2013).

59. Fernandes, J. C. Therapeutic application of antibody fragments in autoimmune diseases: current state and prospects. *Drug Discovery Today* vol. 23 1996–2002 (2018).

60. Li, Z. *et al.* Influence of molecular size on tissue distribution of antibody fragments. *MAbs* **8**, 113–119 (2016).

61. Yokota, T., Milenic, D. E., Whitlow, M. & Schlom, J. Rapid tumor penetration of a single-chain Fv and comparison with other immunoglobulin forms. *Cancer Res.* **52**, (1992).

62. Weidle, U. H., Auer, J., Brinkmann, U., Georges, G. & Tiefenthaler, G. The emerging role of new protein scaffold-based agents for treatment of cancer. *Cancer Genomics Proteomics* **10**, 155–68 (2013).

63. Mitragotri, S., Burke, P. A. & Langer, R. Overcoming the challenges in administering biopharmaceuticals: formulation and delivery strategies. *Nat. Rev. Drug Discov.* **13**, 655–672 (2014).

64. Kaplon, H., Muralidharan, M., Schneider, Z. & Reichert, J. M. Antibodies

to watch in 2020. *MAbs* **12**, (2020).

65. Zhang, C. Hybridoma technology for the generation of monoclonal antibodies. *Methods in Molecular Biology* vol. 901 117–135 (2012).

66. Little, M., Kipriyanov, S. M., Le Gall, F. & Moldenhauer, G. Of mice and men: Hybridoma and recombinant antibodies. *Immunology Today* vol. 21 364–370 (2000).

67. Osborn, M. J., Freeman, M. & Huennekens, F. M. Inhibition of dihydrofolic reductase by aminopterin and amethopterin. *Exp. Biol. Med.* **97**, 429–431 (1958).

68. Murray, A. W. The biological significance of purine salvage. *Annu. Rev. Biochem.* **40**, 811–826 (1971).

69. Arnold, W. J. & Kelley, W. N. Human hypoxanthine-guanine phosphoribosyltransferase purification and subunit structure. *J. Biol. Chem.* **246**, 7398–7404 (1971).

70. Parray, H. A. *et al.* Hybridoma technology a versatile method for isolation of monoclonal antibodies, its applicability across species, limitations, advancement and future perspectives. *International Immunopharmacology* vol. 85 106639 (2020).

71. Morrison, S. L., Johnson, M. J., Herzenberg, L. A. & Oi, V. T. Chimeric

human antibody molecules: Mouse antigen-binding domains with human constant region domains. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 6851–6855 (1984).

72. Boulianne, G. L., Hozumi, N. & Shulman, M. J. Production of functional chimaeric mouse/human antibody. *Nature* **312**, 643–646 (1984).

73. Harding, F. A., Stickler, M. M., Razo, J. & DuBridge, R. B. The immunogenicity of humanized and fully human antibodies: Residual immunogenicity resides in the CDR regions. *MAbs* **2**, 256–265 (2010).

74. Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S. & Winter, G. Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature* **321**, 522–525 (1986).

75. Chames, P., Van Regenmortel, M., Weiss, E. & Baty, D. Therapeutic antibodies: successes, limitations and hopes for the future. *Br. J. Pharmacol.* **157**, 220–33 (2009).

76. Smith, G. P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315–1317 (1985).

77. McCafferty, J., Griffiths, A. D., Winter, G. & Chiswell, D. J. Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* **348**, 552–554 (1990).

78. Nixon, A. E., Sexton, D. J. & Ladner, R. C. Drugs derived from phage display: from candidate identification to clinical practice. *MAbs* **6**, 73–85 (2014).

79. Ledsgaard, L., Kilstrup, M., Karatt-Vellatt, A., McCafferty, J. & Laustsen, A. H. Basics of antibody phage display technology. *Toxins (Basel).* **10**, (2018).

80. Jespers, L. S., Roberts, A., Mahler, S. M., Winter, G. & Hoogenboom, H. R. Guiding the selection of human antibodies from phage display repertoires to a single epitope of an antigen. *Bio/Technology* **12**, 899–903 (1994).

81. Kempeni, J. Preliminary results of early clinical trials with the fully human anti-TNFalpha monoclonal antibody D2E7. *Ann. Rheum. Dis.* **58 Suppl 1**, I70-2 (1999).

82. Mease, P. J. Adalimumab in the treatment of arthritis. *Ther. Clin. Risk Manag.* **3**, 133–48 (2007).

83. Top 15 Best-Selling Drugs of 2019. https://www.genengnews.com/topics/drug-discovery/top-15-best-selling-drugs-of-2019/.

84. Schaffitzel, C., Hanes, J., Jermutus, L. & Plückthun, A. Ribosome display: an in vitro method for selection and evolution of antibodies from

libraries. *J. Immunol. Methods* **231**, 119–35 (1999).

85. Elgundi, Z., Reslan, M., Cruz, E., Sifniotis, V. & Kayser, V. The state-of-play and future of antibody therapeutics. *Adv. Drug Deliv. Rev.* (2016) doi:10.1016/j.addr.2016.11.004.

86. Hanes, J. & Plckthun, A. In vitro selection and evolution of functional proteins by using ribosome display. *Biochemistry* **94**, 4937–4942 (1997).

87. He, M. & Taussig, M. J. Antibody-ribosome-mRNA (ARM) complexes as efficient selection particles for in vitro display and evolution of antibody combining sites. *Nucleic Acids Res.* **25**, 5132–4 (1997).

88. A, R., RJ, H. & BE, P. Ribosome display for improved biotherapeutic molecules. *Expert Opin. Biol. Ther.* **6**, 177–187 (2006).

89. Galán, A. *et al.* Library-based display technologies: where do we stand? *Mol. Biosyst.* **12**, 2342–2358 (2016).

90. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.* **15**, 553–557 (1997).

91. Chao, G. *et al.* Isolating and engineering human antibodies using yeast surface display. *Nat. Protoc.* **1**, 755–768 (2006).

92. Cherf, G. M. & Cochran, J. R. Applications of yeast surface display for protein engineering. *Methods Mol. Biol.* **1319**, 155–175 (2015).

93. Hackel, B. J., Ackerman, M. E., Howland, S. W. & Wittrup, K. D. Stability and CDR composition biases enrich binderfunctionality landscapes. *J. Mol. Biol.* **401**, 84–96 (2010).

94. Boder, E. T. & Wittrup, K. D. Yeast surface display for directed evolution of protein expression, affinity, and stability. *Methods Enzymol.* **328**, 430–444 (2000).

95. Tiller, K. E. & Tessier, P. M. Advances in antibody design. *Annu. Rev. Biomed. Eng.* **17**, 191–216 (2015).

96. I. Razinkov, V., J. Treuheit, M. & W. Becker, G. Methods of high throughput biophysical characterization in biopharmaceutical development. *Curr. Drug Discov. Technol.* **10**, 59–70 (2013).

97. Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M. & Granger, C. B. *Fundamentals of clinical trials*. *Fundamentals of Clinical Trials* (Springer International Publishing, 2015). doi:10.1007/978-3-319-18539-2.

98. Bailly, M. *et al.* Predicting antibody developability profiles through early stage discovery screening. *MAbs* **12**, 1743053 (2020).

99. Shukla, A. A. & Thömmes, J. Recent advances in large-scale production of monoclonal antibodies and related proteins. *Trends Biotechnol.* **28**, 253–261 (2010).

100. Tripathi, N. K. & Shrivastava, A. Recent developments in bioprocessing of recombinant proteins: expression hosts and process development. *Front. Bioeng. Biotechnol.* **7**, 420 (2019).

101. Andersen, D. C. & Krummen, L. Recombinant protein expression for therapeutic applications. *Current Opinion in Biotechnology* vol. 13 117–123 (2002).

102. Butler, M. & Meneses-Acosta, A. Recent advances in technology supporting biopharmaceutical production from mammalian cells. *Appl. Microbiol. Biotechnol.* **96**, 885–894 (2012).

103. Kim, J. Y., Kim, Y. G. & Lee, G. M. CHO cells in biotechnology for production of recombinant proteins: Current state and further potential. *Appl. Microbiol. Biotechnol.* **93**, 917–930 (2012).

104. Thomas, P. & Smart, T. G. HEK293 cell line: A vehicle for the expression of recombinant proteins. *J. Pharmacol. Toxicol. Methods* **51**, 187–200 (2005).

105. Hunter, M., Yuan, P., Vavilala, D. & Fox, M. Optimization of protein expression in mammalian cells. *Curr. Protoc. Protein Sci.* **95**, (2019).

106. Liu, H. F., Ma, J., Winter, C. & Bayer, R. Recovery and purification process development for monoclonal antibody production. *MAbs* **2**, 480–499 (2010).

107. Shukla, A. A., Hubbard, B., Tressel, T., Guhan, S. & Low, D. Downstream processing of monoclonal antibodies-Application of platform approaches. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* vol. 848 28–39 (2007).

108. Hober, S., Nord, K. & Linhult, M. Protein A chromatography for antibody purification. *J. Chromatogr. B* **848**, 40–47 (2007).

109. Marichal-Gallardo, P. A. & Álvarez, M. M. State-of-the-art in downstream processing of monoclonal antibodies: Process trends in design and validation. *Biotechnol. Prog.* **28**, 899–916 (2012).

110. Rathore, N. & Rajan, R. S. Current perspectives on stability of protein drug products during formulation, fill and finish operations. *Biotechnol. Prog.* **24**, 504–514 (2008).

111. Jozala, A. F. *et al.* Biopharmaceuticals from microorganisms: from production to purification. *Brazilian J. Microbiol.* **47**, 51–63 (2016).

112. Vázquez-Rey, M. & Lang, D. A. Aggregates in monoclonal antibody manufacturing processes. *Biotechnol. Bioeng.* **108**, 1494–1508 (2011).

113. Rosenberg, A. S. Effects of protein aggregates: an immunologic perspective. *AAPS J.* **8**, E501–E507 (2006).

114. Sauerborn, M., Brinks, V., Jiskoot, W. & Schellekens, H. Immunological mechanism underlying the immune response to recombinant human protein therapeutics. *Trends Pharmacol. Sci.* **31**, 53–59 (2010).

115. Gunn, G. R. *et al.* From the bench to clinical practice: understanding the challenges and uncertainties in immunogenicity testing for biopharmaceuticals. *Clin. Exp. Immunol.* **184**, 137–146 (2016).

116. Cromwell, M. E. M., Hilario, E. & Jacobson, F. Protein aggregation and bioprocessing. *AAPS J.* **8**, E572–E579 (2006).

117. Mahler, H.-C., Friess, W., Grauschopf, U. & Kiese, S. Protein Aggregation: Pathways, induction factors and analysis. *J. Am. Pharm. Assoc.* **98**, 2909–2934 (2010).

118. Wang, W. Protein aggregation and its inhibition in biopharmaceutics. *Int. J. Pharm.* **289**, 1–30 (2005).

119. Speed, M. A., King, J. & Wang, D. I. C. Polymerization mechanism of polypeptide chain aggregation. *Biotechnol. Bioeng.* **54**, 333–343 (1997).

120. Luo, Q. *et al.* Chemical modifications in therapeutic protein aggregates generated under different stress conditions. *J. Biol. Chem.* **286**, 25134–

25144 (2011).

121. Sahin, E., Grillo, A. O., Perkins, M. D. & Roberts, C. J. Comparative effects of pH and ionic strength on protein–protein interactions, unfolding, and aggregation for IgG1 antibodies. *J. Pharm. Sci.* **99**, 4830–4848 (2010).

122. Bee, J. S., Davis, M., Freund, E., Carpenter, J. F. & Randolph, T. W. Aggregation of a monoclonal antibody induced by adsorption to stainless steel. *Biotechnol. Bioeng.* **105**, 121–129 (2010).

123. Gerhardt, A. *et al.* Protein aggregation and particle formation in prefilled glass syringes. *J. Pharm. Sci.* **103**, 1601–1612 (2014).

124. Sharma, B. Immunogenicity of therapeutic proteins. Part 2: Impact of container closures. *Biotechnol. Adv.* **25**, 318–324 (2007).

125. Bee, J. S. *et al.* Production of particles of therapeutic proteins at the air-water interface during compression/dilation cycles. *Soft Matter* **8**, 10329–10335 (2012).

126. Koepf, E., Eisele, S., Schroeder, R., Brezesinski, G. & Friess, W. Notorious but not understood: How liquid-air interfacial stress triggers protein aggregation. *Int. J. Pharm.* **537**, 202–212 (2018).

127. Dobson, J. *et al.* Inducing protein aggregation by extensional flow. *Proc.*

Natl. Acad. Sci. U. S. A. **114**, 4673–4678 (2017).

128. Munishkina, L. A., Ahmad, A., Fink, A. L. & Uversky, V. N. Guiding protein aggregation with macromolecular crowding. *Biochemistry* **47**, 8993–9006 (2008).

129. Shire, S. J., Shahrokh, Z. & Liu, J. Challenges in the development of high protein concentration formulations. *J. Pharm. Sci.* **93**, 1390–1402 (2004).

130. Neergaard, M. S. *et al.* Viscosity of high concentration protein formulations of monoclonal antibodies of the IgG1 and IgG4 subclass - Prediction of viscosity through protein-protein interaction measurements. *Eur. J. Pharm. Sci.* **49**, 400–410 (2013).

131. Fitzroy Willis, L. *The Effects of Flow on Therapeutic Protein Aggregation*. (2018).

132. Hamrang, Z., Rattray, N. J. W. & Pluen, A. Proteins behaving badly: Emerging technologies in profiling biopharmaceutical aggregation. *Trends Biotechnol.* **31**, 448–458 (2013).

133. Rousseau, F., Serrano, L. & Schymkowitz, J. W. H. How evolutionary pressure against protein aggregation shaped chaperone specificity. *J. Mol. Biol.* **355**, 1037–1047 (2006).

134. Monsellier, E., Ramazzotti, M., Taddei, N. & Chiti, F. Aggregation propensity of the human proteome. *PLoS Comput. Biol.* **4**, (2008).

135. Houben, B. *et al.* Autonomous aggregation suppression by acidic residues explains why chaperones favour basic residues. *EMBO J.* (2020) doi:10.15252/embj.2019102864.

136. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382–W388 (2005).

137. De Baets, G., Van Durme, J., Van Der Kant, R., Schymkowitz, J. & Rousseau, F. Solubis: optimize your protein. *Bioinformatics* **31**, 2580–2582 (2015).

138. Van Durme, J. *et al.* Solubis: a webserver to reduce protein aggregation through mutation. *Protein Eng. Des. Sel.* **29**, 285–289 (2016).

139. van der Kant, R. *et al.* Prediction and reduction of the aggregation of monoclonal antibodies. *J. Mol. Biol.* **429**, 1244–1261 (2017).

140. Conchillo-Solé, O. *et al.* AGGRESCAN: a server for the prediction and evaluation of 'hot spots'; of aggregation in polypeptides. *BMC Bioinformatics* **8**, 65 (2007).

141. Zambrano, R. *et al.* AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Res.* **43**, W306–W313 (2015).

142. Kuriata, A. *et al.* CABS-flex 2.0: a web server for fast simulations of

flexibility of protein structures. *Nucleic Acids Res.* **46**, W338–W343 (2018).

143. Kuriata, A. *et al.* Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. *Nucleic Acids Res.* **47**, W300–W307 (2019).

144. Sormanni, P., Aprile, F. A. & Vendruscolo, M. The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* **427**, 478–490 (2015).

145. Sormanni, P., Amery, L., Ekizoglou, S., Vendruscolo, M. & Popovic, B. Rapid and accurate in silico solubility screening of a monoclonal antibody library. *Sci. Rep.* **7**, 8200 (2017).

146. Wolf Pérez, A. M. *et al.* In vitro and in silico assessment of the developability of a designed monoclonal antibody library. *MAbs* **11**, 388–400 (2019).

147. Hebditch, M., Carballo-Amador, M. A., Charonis, S., Curtis, R. & Warwicker, J. Protein–Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics* **33**, 3098–3100 (2017).

148. Hebditch, M. & Warwicker, J. Web-based display of protein surface and pH-dependent properties for assessing the developability of biotherapeutics. *Sci. Rep.* **9**, 1969 (2019).

149. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B. & Trout, B. L.

Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 11937–11942 (2009).

150. Voynov, V., Chennamsetty, N., Kayser, V., Helk, B. & Trout, B. L. Predictive tools for stabilization of therapeutic proteins. *MAbs* **1**, 580–2 (2009).

151. Chennamsetty, N., Helk, B., Voynov, V., Kayser, V. & Trout, B. L. Aggregation-prone motifs in human Immunoglobulin G. *J. Mol. Biol.* **391**, 404–413 (2009).

152. Raybould, M. I. J. *et al.* Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci.* **116**, 4025–4030 (2019).

153. Jain, T. *et al.* Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 944–949 (2017).

154. Lowe, D. *et al.* Aggregation, stability, and formulation of human antibody therapeutics. in *Advances in protein chemistry and structural biology* vol. 84 41–61 (2011).

155. Estep, P. *et al.* An alternative assay to hydrophobic interaction chromatography for high-throughput characterization of monoclonal antibodies. *MAbs* **7**, 553–561 (2015).

156. Kohli, N. *et al.* A novel screening method to assess developability of antibody-like molecules. *MAbs* **7**, 752–758 (2015).

157. Jacobs, S. A., Wu, S. J., Feng, Y., Bethea, D. & O'Neil, K. T. Cross-interaction chromatography: A rapid method to identify highly soluble monoclonal antibody candidates. *Pharm. Res.* **27**, 65–71 (2010).

158. Schuck, P., Perugini, M. A., Gonzales, N. R., Howlett, G. J. & Schubert, D. Size-distribution analysis of proteins by analytical ultracentrifugation: strategies and application to model systems. *Biophys. J.* **82**, 1096–1111 (2002).

159. Zhao, H., Brautigam, C. A., Ghirlando, R. & Schuck, P. Overview of current methods in sedimentation velocity and sedimentation equilibrium analytical ultracentrifugation. *Curr. Protoc. Protein Sci.* **0 20**, (2013).

160. Gabrielson, J. P. *et al.* Precision of protein aggregation measurements by sedimentation velocity analytical ultracentrifugation in biopharmaceutical applications. *Anal. Biochem.* **396**, 231–241 (2010).

161. Li, Y., Lubchenko, V. & Vekilov, P. G. The use of dynamic light scattering and Brownian microscopy to characterize protein aggregation. *Rev. Sci. Instrum.* **82**, 053106 (2011).

162. Stetefeld, J., McKenna, S. A. & Patel, T. R. Dynamic light scattering: a practical guide and applications in biomedical sciences. *Biophys. Rev.* **8**,

409–427 (2016).

163. Johnson, C. M. Differential scanning calorimetry as a tool for protein folding and stability. *Arch. Biochem. Biophys.* **531**, 100–109 (2013).

164. Semisotnov, G. V. *et al.* Study of the 'molten globule' intermediate state in protein folding by a hydrophobic fluorescent probe. *Biopolymers* **31**, 119–128 (1991).

165. Shi, S., Semple, A., Cheung, J. & Shameem, M. DSF method optimization and its application in predicting protein thermal aggregation kinetics. *J. Pharm. Sci.* **102**, 2471–2483 (2013).

166. Lang, B. E. & Cole, K. D. Differential scanning calorimetry and fluorimetry measurements of monoclonal antibodies and reference proteins: Effect of scanning rate and dye selection. *Biotechnol. Prog.* **33**, 677–686 (2017).

167. Thiagarajan, G., Semple, A., James, J. K., Cheung, J. K. & Shameem, M. A comparison of biophysical characterization techniques in predicting monoclonal antibody stability. *MAbs* **8**, 1088–1097 (2016).

168. Liu, Y. *et al.* High-throughput screening for developability during early-stage antibody discovery using self-interaction nanoparticle spectroscopy. *MAbs* **6**, 483–492 (2014).

169. Sule, S. V, Dickinson, C. D., Lu, J., Chow, C.-K. & Tessier, P. M. Rapid analysis of antibody self-association in complex mixtures using immunogold conjugates. *Mol. Pharm.* **10**, 1322–1331 (2013).

170. Nishi, H. *et al.* Fc domain mediated self-association of an IgG1 monoclonal antibody under a low ionic strength condition. *J. Biosci. Bioeng.* **112**, 326–332 (2011).

171. Sule, S. V. *et al.* High-throughput analysis of concentration-dependent antibody self-association. *Biophys. J.* **101**, 1749–1757 (2011).

172. Tessier, P. M., Jinkoji, J., Cheng, Y. C., Prentice, J. L. & Lenhoff, A. M. Self-interaction nanoparticle spectroscopy: A nanoparticle-based protein interaction assay. *J. Am. Chem. Soc.* **130**, 3106–3112 (2008).

173. Hötzel, I. *et al.* A strategy for risk mitigation of antibodies with fast clearance. *MAbs* **4**, 753–760 (2012).

174. Rabia, L. A., Desai, A. A., Jhajj, H. S. & Tessier, P. M. Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility. *Biochemical Engineering Journal* vol. 137 365–374 (2018).

175. Kamerzell, T. J., Esfandiary, R., Joshi, S. B., Middaugh, C. R. & Volkin, D. B. Protein–excipient interactions: Mechanisms and biophysical characterization applied to protein formulation development. *Adv. Drug Deliv. Rev.* **63**, 1118–1159 (2011).

176. Ohtake, S., Kita, Y. & Arakawa, T. Interactions of formulation excipients with proteins in solution and in the dried state. *Adv. Drug Deliv. Rev.* **63**, 1053–1073 (2011).

177. Yamaguchi, H. & Miyazaki, M. Refolding techniques for recovering biologically active recombinant proteins from inclusion bodies. *Biomolecules* **4**, 235–251 (2014).

178. Baynes, B. M., Wang, D. I. C. & Trout, B. L. Role of arginine in the stabilization of proteins against aggregation. *Biochemistry* **44**, 4919–4925 (2005).

179. Das, U. *et al.* Inhibition of protein aggregation: supramolecular assemblies of arginine hold the key. *PLoS One* **2**, e1176 (2007).

180. Arakawa, T. & Tsumoto, K. The effects of arginine on refolding of aggregated proteins: not facilitate refolding, but suppress aggregation. *Biochem. Biophys. Res. Commun.* **304**, 148–52 (2003).

181. Timasheff, S. N. Protein-solvent preferential interactions, protein hydration, and the modulation of biochemical reactions by solvent components. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 9721–9726 (2002).

182. McNally, E. J. & Hastedt, J. E. *Protein formulation and delivery*. vol. 175 (CRC Press, 2008).

183. Razinkov, V. I., Treuheit, M. J. & Becker, G. W. Accelerated formulation development of monoclonal antibodies (mAns) and mAb-based modalities: review of methods and tools. *J. Biomol. Screen.* **20**, 468–483 (2015).

184. Ebo, J. S., Guthertz, N., Radford, S. E. & Brockwell, D. J. Using protein engineering to understand and modulate aggregation. *Current Opinion in Structural Biology* vol. 60 157–166 (2020).

185. Franey, H., Brych, S. R., Kolvenbach, C. G. & Rajan, R. S. Increased aggregation propensity of IgG2 subclass over IgG1: Role of conformational changes and covalent character in isolated aggregates. *Protein Sci.* **19**, 1601–1615 (2010).

186. Hari, S. B., Lau, H., Razinkov, V. I., Chen, S. & Latypov, R. F. Acid-induced aggregation of human monoclonal IgG1 and IgG2: Molecular mechanism and the effect of solution composition. *Biochemistry* **49**, 9328–9338 (2010).

187. Arosio, P., Rima, S. & Morbidelli, M. Aggregation mechanism of an IgG2 and two IgG1 monoclonal antibodies at low pH: From oligomers to larger aggregates. *Pharm. Res.* **30**, 641–654 (2013).

188. Pepinsky, R. B. *et al.* Improving the solubility of anti-LINGO-1 monoclonal antibody Li33 by isotype switching and targeted mutagenesis. *Protein Sci.* **19**, 954–66 (2010).

189. Wu, H., Kroe-Barrett, R., Singh, S., Robinson, A. S. & Roberts, C. J. Competing aggregation pathways for monoclonal antibodies. *FEBS Lett.* **588**, 936–941 (2014).

190. Perchiacca, J. M., Lee, C. C. & Tessier, P. M. Optimal charged mutations in the complementarity-determining regions that prevent domain antibody aggregation are dependent on the antibody scaffold. *Protein Eng. Des. Sel.* **27**, 29–39 (2014).

191. Antibody Engineering | Absolute Antibody. https://absoluteantibody.com/custom-services/antibody-engineering/.

192. Teplyakov, A. *et al.* Epitope mapping of anti-interleukin-13 neutralizing antibody CNTO607. *J. Mol. Biol.* **389**, 115–123 (2009).

193. Wu, S.-J. *et al.* Structure-based engineering of a monoclonal antibody for improved solubility. *Protein Eng. Des. Sel.* **23**, 643–651 (2010).

194. Meisl, G., Yang, X., Dobson, C. M., Linse, S. & Knowles, T. P. J. Modulation of electrostatic interactions to reveal a reaction network unifying the aggregation behaviour of the Aβ42 peptide and its variants. *Chem. Sci.* **8**, 4352–4362 (2017).

195. Simeonov, P., Berger-Hoffmann, R., Hoffmann, R., Strater, N. & Zuchner, T. Surface supercharged human enteropeptidase light chain shows improved solubility and refolding yield. *Protein Eng. Des. Sel.* **24**,

261–268 (2011).

196. Lawrence, M. S., Phillips, K. J. & Liu, D. R. Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.* **129**, 10110–10112 (2007).

197. Miklos, A. E. *et al.* Structure-based design of supercharged, highly thermoresistant antibodies. *Chem. Biol.* **19**, 449–455 (2012).

198. Austerberry, J. I. *et al.* The effect of charge mutations on the stability and aggregation of a human single chain Fv fragment. *Eur. J. Pharm. Biopharm.* **115**, 18–30 (2017).

199. Lee, C. C. *et al.* Design and optimization of anti-amyloid domain antibodies specific for β-amyloid and islet amyloid polypeptide. *J. Biol. Chem.* **291**, 2858–2873 (2016).

200. Doherty, C. P. A. *et al.* A short motif in the N-terminal region of α-synuclein is critical for both aggregation and function. *Nat. Struct. Mol. Biol.* **27**, 249–259 (2020).

201. Rouhani, M., Khodabakhsh, F., Norouzian, D., Cohan, R. A. & Valizadeh, V. Molecular dynamics simulation for rational protein engineering: Present and future prospectus. *J. Mol. Graph. Model.* **84**, 43–53 (2018).

202. Yu, H., Yan, Y., Zhang, C. & Dalby, P. A. Two strategies to engineer

flexible loops for improved enzyme thermostability. *Sci. Rep.* **7**, 41212 (2017).

203. Zhang, C. *et al.* Computational design to reduce conformational flexibility and aggregation rates of an antibody Fab fragment. *Mol. Pharm.* **15**, 3079–3092 (2018).

204. Codina, N. *et al.* An expanded conformation of an antibody Fab region by X-ray scattering, molecular dynamics, and smFRET identifies an aggregation mechanism. *J. Mol. Biol.* **431**, 1409–1425 (2019).

205. Yu, H. & Dalby, P. A. Exploiting correlated molecular-dynamics networks to counteract enzyme activity–stability trade-off. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E12192–E12200 (2018).

206. Yu, H. & Dalby, P. A. Coupled molecular dynamics mediate long- and short-range epistasis between mutations that affect stability and aggregation kinetics. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11043–E11052 (2018).

207. Chen, K. & Arnold, F. H. Tuning the activity of an enzyme for unusual environments: Sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 5618–5622 (1993).

208. Farinas, E. T., Bulter, T. & Arnold, F. H. Directed enzyme evolution.

Curr. Opin. Biotechnol. **12**, 545–551 (2001).

209. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379–394 (2015).

210. Waldo, G. S. Genetic screens and directed evolution for protein solubility. *Curr. Opin. Chem. Biol.* **7**, 33–38 (2003).

211. Sachsenhauser, V. & Bardwell, J. C. Directed evolution to improve protein folding in vivo. *Curr. Opin. Struct. Biol.* **48**, 117–123 (2018).

212. Cadwell, R. C. & Joyce, G. F. Randomization of genes by PCR mutagenesis. *Genome Res.* **2**, 28–33 (1992).

213. Stemmer, W. P. C. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370**, 389–391 (1994).

214. Greener, A., Callahan, M. & Jerpseth, B. An efficient random mutagenesis technique using an E. coli mutator strain. *Appl. Biochem. Biotechnol. - Part B Mol. Biotechnol.* **7**, 189–195 (1997).

215. Lai, Y. P., Huang, J., Wang, L. F., Li, J. & Wu, Z. R. A new approach to random mutagenesis in vitro. *Biotechnol. Bioeng.* **86**, 622–627 (2004).

216. Reetz, M. T. & Carballeira, J. D. Iterative saturation mutagenesis (ISM)

for rapid directed evolution of functional enzymes. *Nat. Protoc.* **2**, 891–903 (2007).

217. Waldo, G. S., Standish, B. M., Berendzen, J. & Terwilliger, T. C. Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.* **17**, 691–5 (1999).

218. Wurth, C., Guimard, N. K. & Hecht, M. H. Mutations that reduce aggregation of the Alzheimer's Aβ42 peptide: An unbiased search for the sequence determinants of Aβ amyloidogenesis. *J. Mol. Biol.* **319**, 1279–1290 (2002).

219. Matis, I. *et al.* An integrated bacterial system for the discovery of chemical rescuers of disease-associated protein misfolding. *Nat. Biomed. Eng.* **1**, 838–852 (2017).

220. Van Den Berg, S., Löfdahl, A., Härd, T. & Berglund, H. Improved solubility of TEV protease by directed evolution. *J. Biotechnol.* **121**, 291–298 (2006).

221. Dyson, M. R. *et al.* Identification of soluble protein fragments by gene fragmentation and genetic selection. *Nucleic Acids Res.* **36**, 51 (2008).

222. Sieber, V., Martinez, C. A. & Arnold, F. H. Libraries of hybrid proteins from distantly related sequences. *Nat. Biotechnol.* **19**, 456–460 (2001).

223. Cabantous, S., Terwilliger, T. C. & Waldo, G. S. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat. Biotechnol.* **23**, 102–107 (2005).

224. Remy, I. & Michnick, S. W. Clonal selection and in vivo quantitation of protein interactions with protein-fragment complementation assays. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 5394–5399 (1999).

225. Pelletier, J. N., Campbell-Valois, F. X. & Michnick, S. W. Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 12141–12146 (1998).

226. Cabantous, S. & Waldo, G. S. In vivo and in vitro protein solubility assays using split GFP. *Nat. Methods* **3**, 845–854 (2006).

227. Roodveldt, C., Aharoni, A. & Tawfik, D. S. Directed evolution of proteins for heterologous expression and stability. *Curr. Opin. Struct. Biol.* **15**, 50–56 (2005).

228. Wang, T., Badran, A. H., Huang, T. P. & Liu, D. R. Continuous directed evolution of proteins with improved soluble expression. *Nat. Chem. Biol.* **14**, 972–980 (2018).

229. Jespers, L., Schon, O., Famm, K. & Winter, G. Aggregation-resistant domain antibodies selected on phage by heat denaturation. *Nat. Biotechnol.*

22, 1161–1165 (2004).

230. Jung, S., Honegger, A. & Plu, A. *Selection for Improved Protein Stability by Phage Display*. http://www.mrc-cpe.cam.ac.uk/imt-doc/ (1999).

231. Famm, K., Hansen, L., Christ, D. & Winter, G. Thermodynamically stable aggregation-resistant antibody domains through directed evolution. *J. Mol. Biol.* **376**, 926–931 (2008).

232. Vollmer, W., Blanot, D. & De Pedro, M. A. Peptidoglycan structure and architecture. *FEMS Microbiology Reviews* vol. 32 149–167 (2008).

233. Saunders, J. C. An in vivo platform for identifying protein aggregation inhibitors. (University of Leeds, 2014).

234. Jelsch, C., Lenfant, F., Masson, J. M. & Samama, J. P. Crystallization and preliminary crystallographic data on Escherichia coli TEM1 β-lactamase. *J. Mol. Biol.* **223**, 377–380 (1992).

235. Strynadka, N. C. J. *et al.* Molecular structure of the acyl-enzyme intermediate in β-lactam hydrolysis at 1.7 Å resolution. *Nature* **359**, 700–705 (1992).

236. Galarneau, A., Primeau, M., Trudeau, L.-E. & Michnick, S. W. β-Lactamase protein fragment complementation assays as in vivo and in vitro sensors of protein–protein interactions. *Nat. Biotechnol.* **20**, 619–622

(2002).

237. Hallet, B., Sherratt, D. J. & Hayes, F. Pentapeptide scanning mutagenesis: random insertion of a variable five amino acid cassette in a target protein. *Nucleic Acids Res.* **25**, 1866–1867 (1997).

238. Edwards, W. R. *et al.* Regulation of β-lactamase activity by remote binding of heme: functional coupling of unrelated proteins through domain insertion. *Biochemistry* **49**, 6541–6549 (2010).

239. Edwards, W. R., Busse, K., Allemann, R. K. & Jones, D. D. Linking the functions of unrelated proteins using a novel directed evolution domain insertion method. *Nucleic Acids Res.* **36**, e78–e78 (2008).

240. D'Angelo, S. *et al.* Filtering 'genic' open reading frames from genomic DNA samples for advanced annotation. *BMC Genomics* **12**, S5 (2011).

241. Secco, P. *et al.* Antibody library selection by the β-lactamase protein fragment complementation assay. *Protein Eng. Des. Sel.* **22**, 149–158 (2009).

242. Foit, L. *et al.* Optimizing protein stability in vivo. *Mol. Cell* **36**, 861–871 (2009).

243. Pugsley, A. P. The complete general secretory pathway in gram-negative bacteria. *Microbiological Reviews* vol. 57 50–108 (1993).

244. Xiong, P. *et al.* Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. *Nat. Commun.* **5**, 1–9 (2014).

245. Wang, J. *et al.* Recurring sequence-structure motifs in (βα)8-barrel proteins and experimental optimization of a chimeric protein designed based on such motifs. *Biochim. Biophys. Acta - Proteins Proteomics* **1865**, 165–175 (2017).

246. Saunders, J. C. *et al.* An in vivo platform for identifying inhibitors of protein aggregation. *Nat. Chem. Biol.* **12**, 94–101 (2016).

247. Hailu, T. T., Foit, L. & Bardwell, J. C. A. In vivo detection and quantification of chemicals that enhance protein stability. *Anal. Biochem.* **434**, 181–6 (2013).

248. Niklasson, M. *et al.* Robust and convenient analysis of protein thermal and chemical stability. *Protein Science* vol. 24 2055–2062 (2015).

249. Brooks, B. R. *et al.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217 (1983).

250. Miller, S., Janin, J., Lesk, A. M. & Chothia, C. Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–656 (1987).

251. Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol. Biol.* **1151**, 165–188 (2014).

252. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

253. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

254. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

255. Cock, P. J. A. *et al.* Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

256. Devine, P. W. A. Defining the mechanism behind the self-association of therapeutic monoclonal antibodies using mass spectrometric techniques. (University of Leeds, 2016).

257. Lefranc, M.-P. *et al.* IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* **27**, 55–77 (2003).

258. Dunbar, J. & Deane, C. M. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* **32**, btv552 (2015).

259. Jespers, L., Schon, O., James, L. C., Veprintsev, D. & Winter, G. Crystal Structure of HEL4, a Soluble, Refoldable Human VH Single Domain with a Germ-line Scaffold. *J. Mol. Biol.* **337**, 893–903 (2004).

260. Roberts, A. W. G-CSF: A key regulator of neutrophil production, but that's not all! *Growth Factors* **23**, 33–41 (2005).

261. Bendall, L. J. & Bradstock, K. F. G-CSF: From granulopoietic stimulant to bone marrow stem cell mobilizing agent. *Cytokine Growth Factor Rev.* **25**, 355–367 (2014).

262. Mehta, H. M., Malandra, M. & Corey, S. J. G-CSF and GM-CSF in Neutropenia. *J. Immunol.* **195**, 1341–1349 (2015).

263. Hill, C. P., Osslund, T. D. & Eisenberg, D. The structure of granulocyte-colony-stimulating factor and its relationship to other growth factors. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 5167–71 (1993).

264. Zink, T. *et al.* Structure and dynamics of the human granulocyte colony-stimulating factor determined by NMR spectroscopy. Loop mobility in a four-helix-bundle protein. *Biochemistry* **33**, 8453–8463 (1994).

265. Raso, S. W. *et al.* Aggregation of granulocyte-colony stimulating factor in

vitro involves a conformationally altered monomeric state. *Protein Sci.* **14**, 2246–57 (2005).

266. Chi, E. Y. *et al.* Roles of conformational stability and colloidal stability in the aggregation of recombinant human granulocyte colony-stimulating factor. *Protein Sci.* **12**, 903–913 (2003).

267. Thirumangalathu, R., Krishnan, S., Brems, D. N., Randolph, T. W. & Carpenter, J. F. Effects of pH, temperature, and sucrose on benzyl alcohol-induced aggregation of recombinant human granulocyte colony stimulating factor. *J. Pharm. Sci.* **95**, 1480–1497 (2006).

268. Vanz, A. L. S. *et al.* Human granulocyte colony stimulating factor (hG-CSF): Cloning, overexpression, purification and characterization. *Microb. Cell Fact.* **7**, 13 (2008).

269. Buchanan, A. *et al.* Improved drug-like properties of therapeutic proteins by directed evolution. *Protein Eng. Des. Sel.* **25**, 631–638 (2012).

270. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).

271. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* (1982) doi:10.1016/0022-2836(82)90515-0.

272. Langenberg, T. *et al.* Thermodynamic and evolutionary coupling between the native and amyloid state of globular proteins. *Cell Rep.* **31**, (2020).

273. Morell, M., de Groot, N. S., Vendrell, J., Avilés, F. X. & Ventura, S. Linking amyloid protein aggregation and yeast survival. *Mol. Biosyst.* **7**, 1121–8 (2011).

274. Espargaró, A., Sabate, R. & Ventura, S. Thioflavin-S staining coupled to flow cytometry. A screening tool to detect in vivo protein aggregation. *Mol. Biosyst.* **8**, 2839 (2012).

275. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).

276. Bolognesi, B. *et al.* The mutational landscape of a prion-like domain. *Nat. Commun.* **10**, 4162 (2019).

277. Gray, V. E. *et al.* Elucidating the molecular determinants of Aβ aggregation with deep mutational scanning. *G3 Genes, Genomes, Genet.* **11**, 3683–3689 (2019).

278. Schmidt-Dannert, C. & Arnold, F. H. Directed evolution of industrial enzymes. *Trends Biotechnol.* **17**, 135–136 (1999).

279. Julian, M. C., Li, L., Garde, S., Wilen, R. & Tessier, P. M. Efficient affinity maturation of antibody variable domains requires co-selection of

compensatory mutations to maintain thermodynamic stability. *Sci. Rep.* **7**, 45259 (2017).

280. Kazlauskas, R. Engineering more stable proteins. *Chem. Soc. Rev.* **47**, 9026–9045 (2018).

281. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: A biophysical view of protein evolution. *Nature Reviews Genetics* vol. 6 678–687 (2005).

282. Borrok, M. J. *et al.* An 'Fc-Silenced' IgG1 format with extended half-life designed for improved stability. *J. Pharm. Sci.* **106**, 1008–1017 (2017).

283. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* vol. 19 596–604 (2009).

284. Niesen, F. H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat. Protoc.* **2**, 2212–2221 (2007).

285. Sikosek, T. & Chan, H. S. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of the Royal Society Interface* vol. 11 (2014).

286. Mathis, G. Probing molecular interactions with homogeneous techniques based on rare earth cryptates and fluorescence energy transfer. *Clin. Chem.*

41, 1391–1397 (1995).

287. Degorce, F. *et al.* HTRF: A technology tailored for drug discovery - a review of theoretical aspects and recent applications. *Curr. Chem. Genomics* **3**, 22–32 (2009).

288. Harlow, D. E., Honce, J. M. & Miravalle, A. A. Remyelination therapy in multiple sclerosis. *Frontiers in Neurology* vol. 6 257 (2015).

289. Chiti, F. & Dobson, C. M. Protein misfolding, amyloid formation, and human disease: A summary of progress over the last decade. *Annu. Rev. Biochem.* **86**, 27–68 (2017).

290. Benson, M. D. *et al.* Amyloid nomenclature 2020: update and recommendations by the International Society of Amyloidosis (ISA) nomenclature committee. *Amyloid* **27**, 217–222 (2020).

291. Eisenberg, D. & Jucker, M. The amyloid state of proteins in human diseases. *Cell* vol. 148 1188–1203 (2012).

292. Merlini, G., Wechalekar, A. D. & Palladini, G. Systemic light chain amyloidosis: An update for treating physicians. *Blood* **121**, 5124–5130 (2013).

293. Blancas-Mejia, L. M. *et al.* Immunoglobulin light chain amyloid aggregation. *Chem. Commun.* **54**, 10664–10674 (2018).

294. Sanchorawala, V. Light-chain (AL) amyloidosis: diagnosis and treatment. *Clinical journal of the American Society of Nephrology : CJASN* vol. 1 1331–1341 (2006).

295. Merlini, G. *et al.* Systemic immunoglobulin light chain amyloidosis. *Nat. Rev. Dis. Prim.* **4**, 1–19 (2018).

296. Schatz, D. G. & Ji, Y. Recombination centres and the orchestration of V(D)J recombination. *Nature Reviews Immunology* vol. 11 251–263 (2011).

297. Feige, M. J., Hendershot, L. M. & Buchner, J. How antibodies fold. *Trends Biochem. Sci.* **35**, 189–198 (2010).

298. Jung, D. & Alt, F. W. Unraveling V(D)J recombination: Insights into gene regulation. *Cell* vol. 116 299–311 (2004).

299. Odegard, V. H. & Schatz, D. G. Targeting of somatic hypermutation. *Nature Reviews Immunology* vol. 6 573–583 (2006).

300. Rottenaicher, G. J. *et al.* Molecular mechanism of amyloidogenic mutations in hypervariable regions of antibody light chains. *J. Biol. Chem.* **0**, 100334 (2021).

301. Merlini, G. AL amyloidosis: From molecular mechanisms to targeted therapies. *Hematology* **2017**, 1–12 (2017).

302. Buxbaum, J. Aberrant immunoglobulin synthesis in light chain amyloidosis. Free light chain and light chain fragment production by human bone marrow cells in short-term tissue culture. *J. Clin. Invest.* **78**, 798–806 (1986).

303. Olsen, K. E., Sletten, K. & Westermark, P. Fragments of the constant region of immunoglobulin light chains are constituents of AL-amyloid proteins. *Biochem. Biophys. Res. Commun.* **251**, 642–647 (1998).

304. Klimtchuk, E. S. *et al.* The critical role of the constant region in thermal stability and aggregation of amyloidogenic immunoglobulin light chain. *Biochemistry* **49**, 9848–9857 (2010).

305. Enqvist, S., Sletten, K. & Westermark, P. Fibril protein fragmentation pattern in systemic AL-amyloidosis. *J. Pathol.* **219**, 473–480 (2009).

306. Radamaker, L. *et al.* Cryo-EM structure of a light chain-derived amyloid fibril from a patient with systemic AL amyloidosis. *Nat. Commun.* **10**, 1–8 (2019).

307. Swuec, P. *et al.* Cryo-EM structure of cardiac amyloid fibrils from an immunoglobulin light chain AL amyloidosis patient. *Nat. Commun.* **10**, 1–9 (2019).

308. Morgan, G. J. & Kelly, J. W. The kinetic stability of a full-length antibody light chain dimer determines whether endoproteolysis can release

amyloidogenic variable domains. *J. Mol. Biol.* **428**, 4280–4297 (2016).

309. Rennella, E., Morgan, G. J., Kelly, J. W. & Kay, L. E. Role of domain interactions in the aggregation of full-length immunoglobulin light chains. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 854–863 (2019).

310. Weber, B. *et al.* Domain interactions determine the amyloidogenicity of antibody light chain mutants. *J. Mol. Biol.* **432**, 6187–6199 (2020).

311. Kim, Y. S. *et al.* Thermodynamic modulation of light chain amyloid fibril formation. *J. Biol. Chem.* **275**, 1570–1574 (2000).

312. Hurle, M. R., Helms, L. R., Li, L., Chan, W. & Wetzel, R. A role for destabilizing amino acid replacements in light-chain amyloidosis. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 5446–5450 (1994).

313. Wall, J. *et al.* Thermodynamic instability of human λ6 light chains: correlation with fibrillogenicity. *Biochemistry* **38**, 14101–14108 (1999).

314. Baden, E. M., Randles, E. G., Aboagye, A. K., Thompson, J. R. & Ramirez-Alvarado, M. Structural insights into the role of mutations in amyloidogenesis. *J. Biol. Chem.* **283**, 30950–30956 (2008).

315. Nokwe, C. N. *et al.* A residue-specific shift in stability and amyloidogenicity of antibody variable domains∗. *J. Biol. Chem.* **289**, 26829–26846 (2014).

316. Rennella, E., Morgan, G. J., Yan, N., Kelly, J. W. & Kay, L. E. The role of protein thermodynamics and primary structure in fibrillogenesis of variable domains from immunoglobulin light chains. *J. Am. Chem. Soc.* **141**, 13562–13571 (2019).

317. Baden, E. M. *et al.* Altered dimer interface decreases stability in an amyloidogenic protein. *J. Biol. Chem.* **283**, 15853–15860 (2008).

318. Nokwe, C. N. *et al.* A stable mutant predisposes antibody domains to amyloid formation through specific non-native interactions. *J. Mol. Biol.* **428**, 1315–1332 (2016).

319. FDA grants accelerated approval to Darzalex Faspro for newly diagnosed light chain amyloidosis | FDA. https://www.fda.gov/drugs/drug-approvals-and-databases/fda-grants-accelerated-approval-darzalex-faspro-newly-diagnosed-light-chain-amyloidosis.

320. Van De Donk, N. W. C. J. & Usmani, S. Z. CD38 antibodies in multiple myeloma: Mechanisms of action and modes of resistance. *Frontiers in Immunology* vol. 9 2134 (2018).

321. Leem, J., Dunbar, J., Georges, G., Shi, J. & Deane, C. M. ABodyBuilder: Automated antibody structure prediction with data–driven accuracy estimation. *MAbs* **8**, 1259–1268 (2016).

322. Del Pozo Yauner, L., Ortiz, E. & Becerril, B. The CDR1 of the human

λVI light chains adopts a new canonical structure. *Proteins Struct. Funct. Genet.* **62**, 122–129 (2006).

323. González-Andrade, M. *et al.* Mutational and genetic determinants of λ6 light chain amyloidogenesis. *FEBS J.* **280**, 6173–6183 (2013).

324. Del Pozo Yauner, L. *et al.* Influence of the germline sequence on the thermodynamic stability and fibrillogenicity of human lambda 6 light chains. *Proteins Struct. Funct. Genet.* **72**, 684–692 (2008).

325. Kazman, P. *et al.* Fatal amyloid formation in a patient's antibody light chain is caused by a single point mutation. *Elife* **9**, (2020).

326. Zhao, J., Zhang, B., Zhu, J., Nussinov, R. & Ma, B. Structure and energetic basis of overrepresented λ light chain in systemic light chain amyloidosis patients. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1864**, 2294–2303 (2018).

327. O'Nuallain, B. *et al.* Localization of a conformational epitope common to non-native and fibrillar immunoglobulin light chains. *Biochemistry* **46**, 1240–1247 (2007).

328. Piehl, D. W. *et al.* Immunoglobulin light chains form an extensive and highly ordered fibril involving the N- and C-termini. *ACS Omega* **2**, 712–720 (2017).

329. Ruiz-Zamora, R. A. *et al.* The CDR1 and other regions of immunoglobulin light chains are hot spots for amyloid aggregation. *Sci. Rep.* **9**, 1–18 (2019).

330. Hartl, F. U. Protein misfolding diseases. *Annu. Rev. Biochem.* **86**, 21–26 (2017).