From single amino acid deletions to whole domain insertions; Engineering GFP through polypeptide backbone mutations

James Alexander Joseph Arpino

A dissertation submitted to Cardiff University in candidature for the degree of Doctor of Philosophy

December 2011

School of Biosciences
Cardiff University

Submission of thesis declaration and statements

DECLARATION

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed …… (candidate) Date ……

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of ……(insert MCh, MD, MPhil, PhD etc, as appropriate)

Signed …… (candidate) Date ……

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated.
Other sources are acknowledged by explicit references.

Signed …… (candidate) Date ……

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed …… (candidate) Date ……

STATEMENT 4: PREVIOUSLY APPROVED BAR ON ACCESS

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loans **after expiry of a bar on access previously approved by the Graduate Development Committee.**

Signed …… (candidate) Date ……

Contents

Acknowledgements

This Ph.D. was funded by a BBSRC CASE studentship in association with Merck KGaA. I would firstly like to thank my academic supervisor, Dr Dafydd Jones, for all his help and guidance throughout the course of my work. Especially I owe Dafydd a big thank you for his endless patience with my written work and all of his own personal time he put into helping me with my scientific writing. I would secondly like to thank Prof. Matthias Bochtler, Dr. Honorat Czapińska, and Dr. Anna Piasecka for their help and guidance with protein crystallography and structure determination. I would also like to thank Michal Gajda at EMBL for SAXS measurements and data analysis.

I would like to thank Dr Roger Chittock for his interest and enthusiasm throughout my Ph.D. and for organizing my trip to the Merck labs in Darmstadt, Germany to perform fluorescence lifetime measurements. A special thanks must therefore be made to Dr. Dirk Wandschneider and Dr Peter Barnekow for accommodating me in their lab and for all their guidance and help during my stay in Germany.

A big thank you goes to Dr. Wayne Edwards and Dr. Amy Baldwin for all their help around the lab and for the intense tutoring I got when I first started in the lab after a couple of years being out of touch with hands on lab work. I wouldn't have got to where I am now without their guidance. I must mention Sam as I think I quite possibly would have gone potty over my last year in the lab if it wasn't for his light hearted attitude and banter to keep me on my toes. Obvious thanks goes to everyone who has worked in Dafydd's lab that I have not already mentioned over the past 4 years for their help with lab work and buying pints in the Pen and Wig on a Friday evening after a long weeks work, including 'Geordie' Matt, Mike and Andy!

I need to say a special thank you to the love of my life Rachel without whom this thesis probably would not have been finished! The fact you still want to marry me after all the stress you had to put up with, cooking me dinner, doing all the washing up, tidying the house and essentially doing everything for me whilst writing this thesis just makes me love you all that much more.

I would lastly like to thank all of my family, for their constant support through tough times and who have always believed in me. Especially I would like to thank my Nan, Barbara Brown, our weekly chats kept me sane and in touch with reality when I felt things were getting out of my grasp.

Abbreviations

Amino acids have been described throughout the text, figures and tables using the three and one letter code. Mutations are denoted as follows: wild type amino acid, residue number, mutant residue, as in Y39F. A residue that has been deleted is signified with a Δ symbol after the residue number as in G4Δ.

Abbreviations used throughout the text, figures and tables:

A_{xxx}	Absorbance at xxx nm
bp	base pair
BSA	Bovine serum albumin
cfu	Colony forming units
$CG15^{CC}$	CG15 double cysteine mutants
Cyt b_{562}	Cytochrome b_{562}
DTT	Dithiothreitol
dNTP	Equimolar mixture of the four deoxyribonuctleotide triphosphates
EDTA	Ethylenediaminetetraacetic acid
EGFP	Enhanced Green fluorescent protein
EGFPΔ	Single amino acid deletion variants of EGFP
GdmCl	Guanidinium Hydrochloride
GGS	Glycine-Glycine-Serine linker
GFP	Green fluorescent protein
IPTG	Isopropyl β-D-1-thiogalactopyranoside
K	Kelvin
MALDI-MS	Matrix-assisted laser desorption ionization mass spectrometry
MCS	Multiple cloning site
NCS	Non-crystallographic symmetry
NEB	New England Biolabs
NMR	Nuclear magnetic resonance
OD_{xxx}	Optical density at xxx nm
PAGE	Polyacrylamide gel electrophoresis
RMSD	Root mean square deviation
SDM	Site directed mutagenesis
SDS	Sodium dodecyl sulphate
TAE	Tris-acetate-EDTA
TEMED	Tetraethylethylenediamine
TLS	Translation, libration, screw
TND	Triplet nucleotide deletion
TNG	Tris-NaCl-Glycerol
Tris	2-Amino-2-hydroxymethyl-propane-1,3-diol
v/v	Volume per volume
wt	Wild type
w/v	Weight per volume
X	Single random amino acid linker

Abstract

With an ever-expanding protein engineering toolbox different mutational techniques can be used to engineer new or altered function into protein scaffolds without the restriction of sampling just simple substitution mutagenesis. These include approaches that target backbone as well as side chain changes, such as single amino acid deletions or whole domain insertion. The problem with utilizing these mutational approaches is the difficulty in predicting both local and global structural changes on changing the backbone conformation. The aim of this thesis is to demonstrate the tolerance and beneficial influence of backbone targeted mutations using a mutant library-based screening approach, and to provide an understanding of their mechanism of action.

A directed evolution transposon-based approach was used to generate libraries of single amino acid deletion and whole domain insertions into enhanced green fluorescent protein (EGFP). The later involved the insertion of cyt b_{562} as the donating insert domain. Library analysis revealed a wide range of sites were sampled across the backbone of EGFP.

Library screening revealed widespread tolerance of EGFP to single amino acid deletions. Using the crystal structure of EGFP determined here, it was found that that loop regions where particularly tolerant. Two variants with residues G4 or A227 deleted conferred increased protein fluorescence to cell cultures with respect to EGFP. Spectral characterization and unfolding experiments identified that rather than altering the fluorescent properties of EGFP the mutations elicited their effects through altered protein folding and stability.

Screening of the domain insert library revealed that sites spread along the backbone of EGFP were tolerant to cyt b_{562} insertion. Particularly tolerant were loops and the C-terminal end of β-strand 7, with the linker sequences playing a key role. One integral domain fusion scaffold, termed CG6, was identified in which the functions of the two individual domains were highly coupled. CG6 exhibited almost 100% fluorescence quenching upon the binding of haem to the cyt b_{562} domain. CG6 was also shown to potentially act as a sensor for redox state and reactive oxygen species such as H_2O_2 via a haem-dissociation dependent mechanism. The structure of CG6, determined by X-ray crystallography, provided the molecular basis for the functional coupling of the two domains. Critical was the side-by-side domain arrangement caused by differential linker lengths at the pivot position re-enforced by a domain-domain interface that placed the chromophores within 17-18Å of each other.

Further rational design of a cyt b_{562}-EGFP integral fusion scaffold (CG15) was performed to create novel ratiometric fluorescent redox sensors, termed $CG15^{CC}$ variants. The $CG15^{CC}$ variants have been shown to have the most reducing redox midpoint potentials of any protein based redox sensor studied to date. One of the $CG15^{CC}$ variants also has the fastest redox kinetics observed to date.

The survey of the tolerance and influence of single amino acid deletions in EGFP conducted here has highlighted the potential beneficial nature of deletion mutagenesis and has helped provide a molecular understanding of their effect. Through domain insertion mutagenesis and retrospective structure analysis the mechanism behind the functional coupling of two domains has been described and will also help guide future work in the development of novel biomolecular switches.

Chapter 1: Introduction

1.1 Protein structure and function relationship

Proteins are the most functionally diverse of all the macromolecules found in Nature. They are comprised from 20 naturally occurring amino acids [1], and in some rare instances pyrolysine [2] or selenocysteine [3]. The central dogma in molecular biology is that DNA via mRNA encodes the order and composition of amino acids in a given polypeptide chain. The amino acid sequence in turn determines the structure and thus function of a protein; the 3D arrangement of the amino acids brings together amino acids from different regions of amino acid sequence required for the functional conformation of a protein.

The natural protein repertoire displays extraordinary plasticity in terms of both their structure and function. Proteins can take on many different folds (topologies) and have a wide range of functions such as enzymes catalyzing biological reactions, structural cellular components, cellular signaling and in immune response.

The central relationship between protein sequence and structure/function give proteins the inherent characteristic of "programmability". Over the last 20-30 years, protein sequence, structural and functional diversity has been increased in the lab through the use of protein engineering to introduce changes to a target protein. This has proved critical to our molecular understanding of proteins as we can now directly interact and influence a protein rather than being simply passive observers. It has also allowed proteins to be adapted and optimized for a particular application.

1.2 Protein Engineering

1.2.1 General

In 1978 an oligonucleotide mutational method was first demonstrated for altering specific bases in a target gene [4]. This discovery allowed point mutations to be introduced leading to the specific substitution of one target residue to another. This in turn was used as a tool to study the molecular basis of catalysis, substrate specificity, stability and folding of enzymes [5]. One of the first enzymes to be targeted for protein mutagenesis was β-lactamase [6], which confers β-lactam antibiotic resistance to bacteria. Further site directed mutagenesis identified numerous variants of β-lactamase with altered catalytic activity [7, 8], substrate specificity [7, 9,

10] and resistance to inhibitors [7, 10], as well as elucidating the catalytic mechanism [11, 12].

Since its establishment, site directed mutagenesis (SDM) normally based on amino acid substitution, has become a widely used approach to study and adapt proteins. One such protein that has been a target for SDM is the *Aequorea victoria* green fluorescent protein (GFP), a workhorse for cell biology studies. Substitution mutagenesis has been used to alter the fluorescent characteristics of GFP to generate different color variants such as blue [13], cyan and yellow fluorescent proteins [14].

However, amino acid substitutions simply change one side chain for another, which may restrict the sequence and conformational space a particular protein can sample. With an ever-expanding protein engineering toolbox different mutational techniques can be used to engineer new or altered function into protein scaffolds that may not be accessible by simple substitution mutagenesis. This includes deletion of a single amino acid to the insertion of whole protein domains.

1.2.2 Insertion and deletion (InDel) mutagenesis

Insertion and deletion mutations (InDel) are sampled during natural protein evolution but are generally overlooked as a useful tool for protein engineering. This is due to the fundamental difference between substitution and InDel mutations. A substitution mutation only affects the side chain of the residue being mutated but leaves the backbone relatively unperturbed. InDel events however involve the loss or gain of amino acids and therefore alter the length of the protein backbone. The problem with mutations that change the backbone structure is that their effects can be difficult to predict due to the local and global structural changes caused by insertion or removal of an amino acid from the polypeptide. Dogma suggests that loops are likely to be more tolerant to Indel mutations due to the inherent conformational flexibility of these regions, whilst InDel mutations in secondary structures can result in registry shifts (Fig 1.1), which could be considered detrimental to protein structure and therefore function [15].

Whilst both insertion and deletion mutations effect the length of the polypeptide chain insertion mutations generally result in increased conformational flexibility and can be accommodated by ‘looping out’ of the polypeptide chain in secondary structures or loops. Deletion mutagenesis however introduces increased

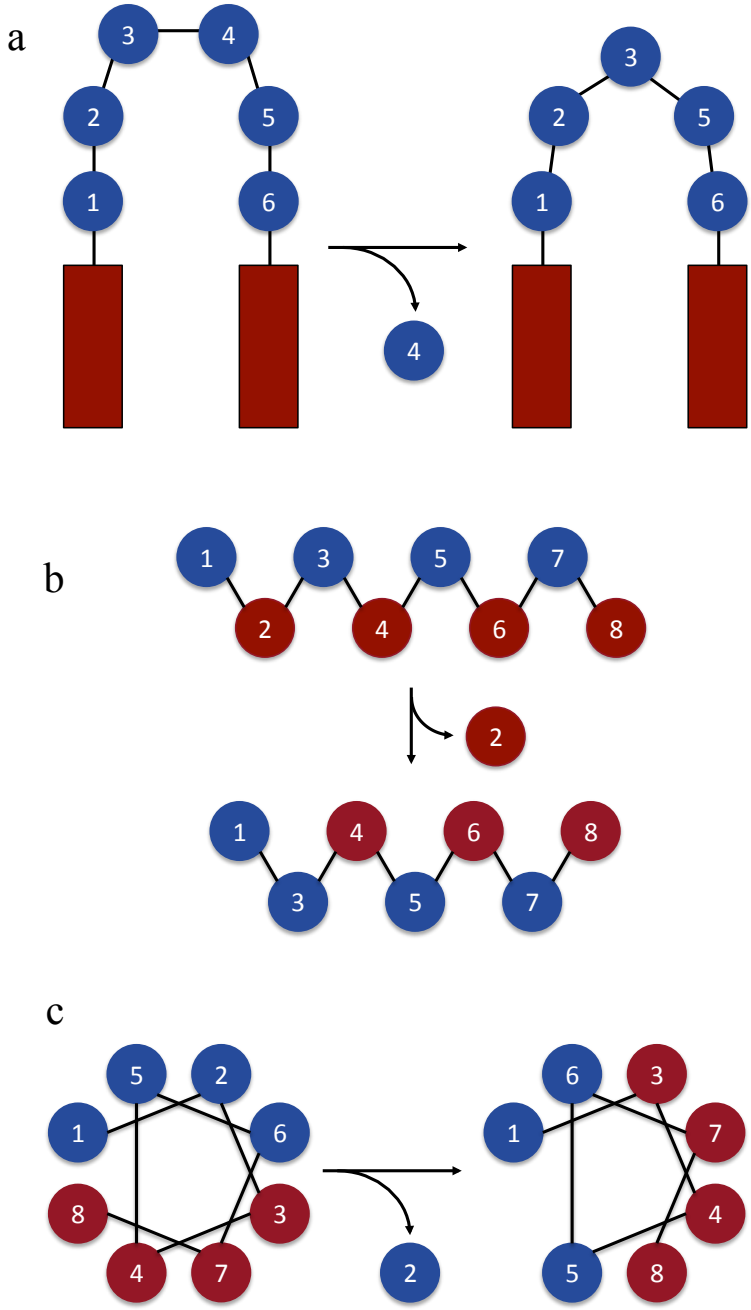


Fig 1.1 Effect of single amino acid deletions on secondary structure registry. a, Deletion of a single amino acid (blue circle) from a loop region connecting two ordered secondary structural elements (red rectangles) is usually accommodated by loop shortening. Deletion of an amino acid (red/blue circle) from **b,** a β-strand or **c,** an α-helix results in registry shifts. Amino acids are coloured red or blue to distinguish between different faces of a secondary structure.

conformational constraints on a polypeptide chain and therefore can be considered to be more detrimental to protein structure and function. Amino acid deletions are also sampled more regularly during natural evolution than insertion mutations due to the mechanism by which Indel mutations are introduced into a gene (described in more detail below). Therefore the effects of deletion mutations have been focused on in this study.

As mentioned above deletion of an amino acid from loops are generally tolerated due to increased conformational flexibility in these regions of a protein and are accommodated through loop shortening (Fig 1.1 a). Deletion of an amino acid from a β-strand could cause the local rearrangement of amino acids in the strand, resulting in a shift of the side chains from one face of the β-strand to the other (Fig 1.1 b), potentially having knock on effects to global structure. For example if the side chains on one face of a surface exposed β-strand were predominantly polar and those on the opposite face were predominantly hydrophobic and buried in the core of the protein, an amino acid deletion may cause a register shift in some of the hydrophobic side chains. The result could be hydrophobic residues becoming solvent exposed and the polar side chains being buried into the core of the protein (Fig 1.1 b). A similar effect may be seen if an amino acid were to be deleted from an α-helix, the potential result being a rotation of all the side chain positions around the α-helix (Fig 1.1 c). As well as affecting registry in organized secondary structures deletion of an amino acid could also result in the loss of secondary structure resulting in the formation of a loop in its place.

During natural protein evolution it has been shown that deletion mutations are more prevalent than insertions [16, 17], which can be explained by the DNA repair mechanism [17]. Either through DNA damage or replication slippage a looped out piece of ssDNA can form, mismatching to its complementary strand, normally at sites in a gene where repeat sequences are present [18]. This DNA loop is more susceptible to cleavage and after DNA repair can result in the insertion or deletion of base pairs from the gene of interest (Fig 1.2). However as outlined in Fig 1.2 deletion mutations can be up to 3 times more prevalent than insertion mutations.

InDel mutations can commonly give rise to frame shifts altering the coding sequence of an entire gene because they can result in a shift of all the subsequent

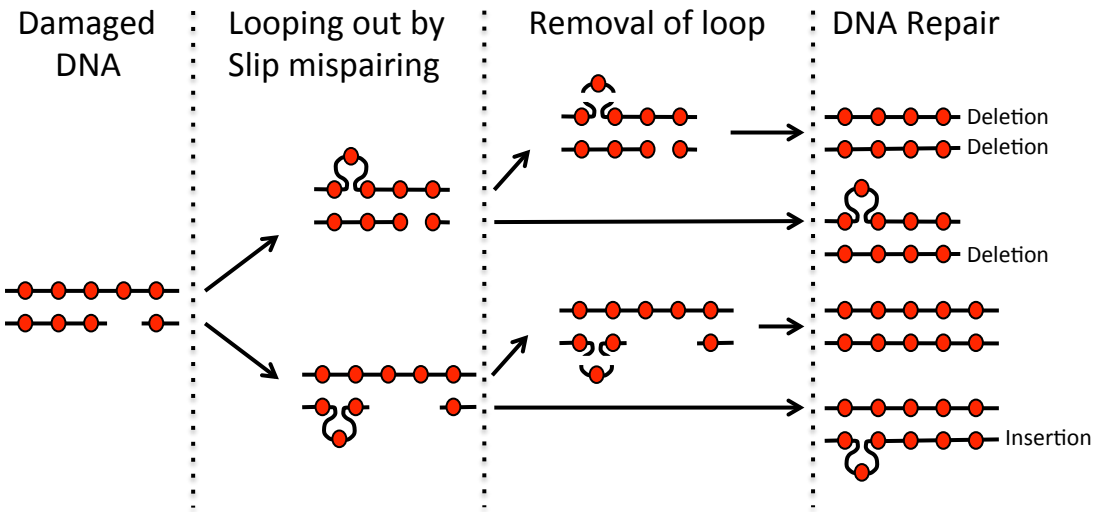


Fig 1.2 Mutational model to explain an excess of deletion mutations. Damaged DNA resulting in the loss of one or more nucleotides can result in looping out of the DNA by slip mispairing. The potential removal of the extrahelical loop leads to an excess of deletion mutations with respect to insertion mutations.

nucleotides changing the codon sequence. However, InDel events that occur in multiples of 3 maintain the reading frame but with the insertion or deletion of amino acids.

Despite the dogma that deletion mutations can be deleterious to protein structure and function they have recently been shown to have positive effects on protein function even when they occur in secondary structures [19]. InDel mutations have been shown to shape the human antibody repertoire by varying epitope surfaces of the heavy and light chains of variable regions on immunoglobulins [18]. InDel mutations also play a key role in the evolution of HIV envelope proteins (encoded by the *env* gene) early in transmission to new hosts allowing the virus to rapidly adapt to its new environment [20].

More recently, protein engineering using deletion mutagenesis has shown beneficial effects of single amino acid deletions. Work carried out on TEM-1 β-lactamase identified many single amino acid deletions that were tolerated within loops and organized secondary structures (Fig 1.3), with several variants displaying increased activity to extended spectrum β-lactam antibiotic susbtrates (Ceftazidime) [19, 21].

Given that InDel mutations can influence protein structure in a manner distinct from that of substitutions, sampling them during protein engineering studies will allow sampling new sequence and conformational space. This in turn may allow the generation of useful protein variants not accessible using substitutions alone.

1.2.3 Integral domain fusion architecture

A protein domain is a structural unit capable of functioning independently of the rest of the polypeptide chain. Many proteins are comprised of multiple domains, with each domain having a discrete function [22, 23]. Structural assignment to gene sequences of complete genomes showed that almost two thirds of prokaryotic and 80% or eukaryotic proteins comprise of multiple domains [24].

The majority of multi-domain proteins are organised in a head to tail fashion with the C-terminal end of one domain leading on to the N-terminus of the next [24] (Fig 1.4 a). However, a significant minority (9%) exhibits integral fusion architecture with one protein domain inserted within another (Fig 1.4 b) [24, 25]. Integral fusion

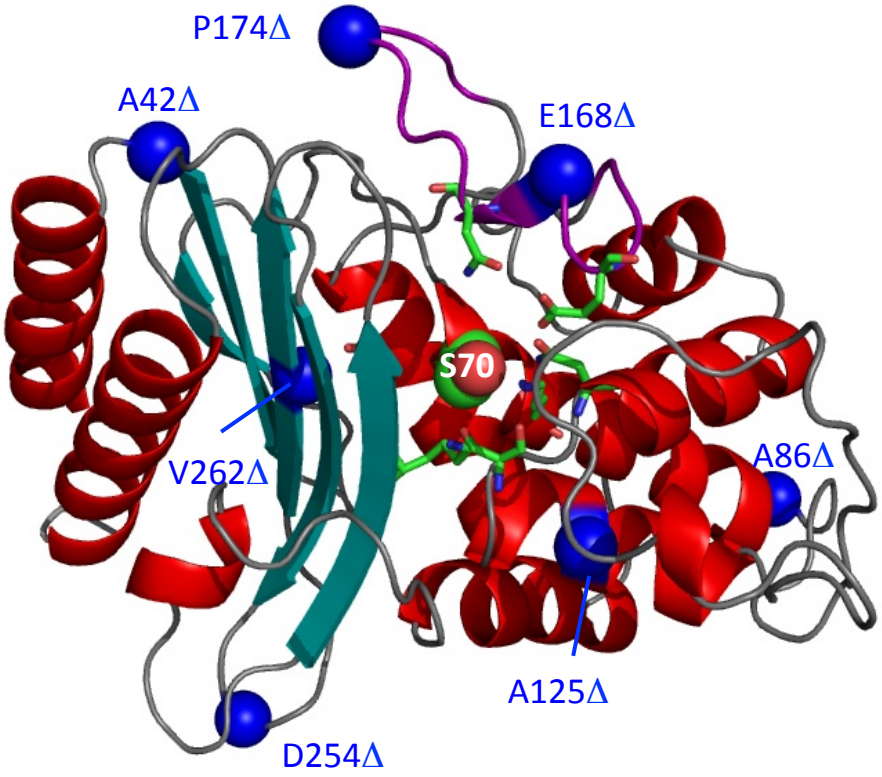


Fig 1.3 Tolerated single amino acid deletions in TEM-1 β-lactamase resulting in increased activity towards ceftazidime. Crystal structure of TEM-1 β-lactamase (pdb:1BTL) in cartoon representation with the active site serine (S70) shown in spacefill with the rest of the active site residues shown as sticks. Positions of tolerated single amino acid deletions are highlighted by blue spheres and labeled, where Δ after a residue number signifies that residue is deleted.

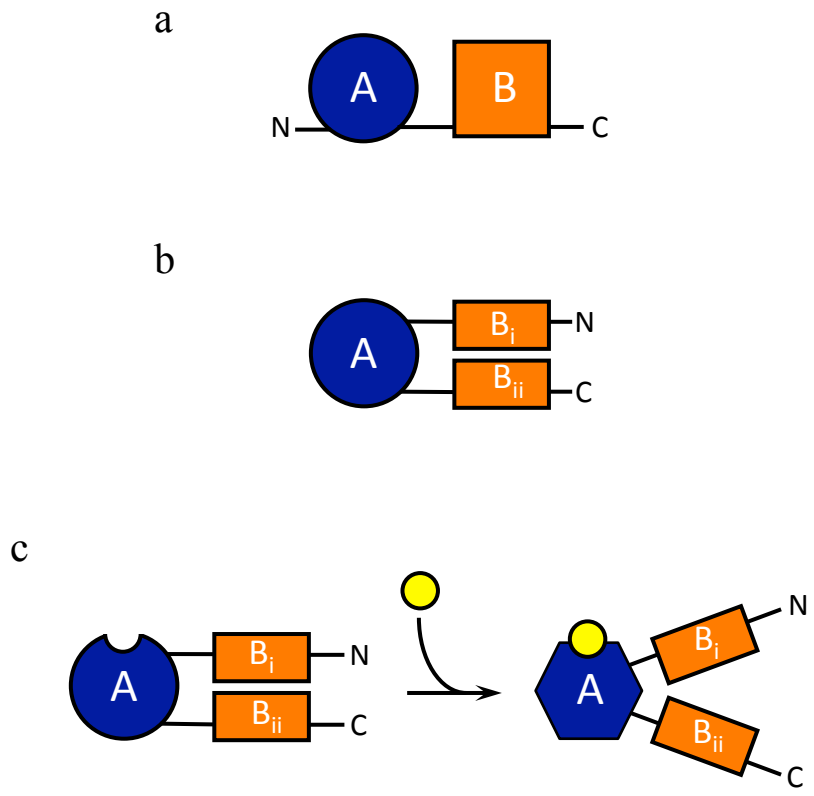


Fig 1.4 Schematic of domain organisation in multi-domain proteins. a, Head-to-tail fusion of two domains (A: blue and B: orange). **b,** Integral fusion domain architecture with the insert domain (A: blue) within the parent domain (Bi, Bii: orange). **c,** Binding of a signal molecule (yellow) to the insert domain triggers a change in conformation of the parent domain. N and C represent the N and C-termini.

architecture decreases the degrees of freedom the 'insert' domain has with respect to the accepting or 'parent' domain, thereby intimately linking the two structures. This intimate linkage has the potential to communicate stimulus-induced changes in one domain through to the other therefore coupling the structure and function of the two proteins (Fig 1.4 c). Analysis of integral domain fusion architecture of natural proteins identified the N- and C-termini of the inserted domain to be juxtaposed with an average distance of ~8 Å from one another, so as not to disrupt the overall structure of the parent domain. This therefore has been considered a prerequisite when considering a protein domain for insertion into another.

A classic example of a domain insert protein arrangement found in nature is the GroEL-GroES chaperone system [26, 27]. The GroEL component is comprised of 3 individual domains, the apical, intermediate and equatorial domain (Fig 1.5). The apical domain is inserted with the intermediate domain, which is in turn inserted within the equatorial domain (Fig 1.5). Communication between the apical and equatorial domain is critical for chaperone function. Polypeptides with a non-native structure bind to a hydrophobic patch on the apical domain with concomitant binding of ATP between the equatorial and intermediate domain [26]. Binding of the cochaperone GroES encapsulates the non-native polypeptide triggering a conformational change (Fig 1.5) resulting in an increase in the size of the cavity, which becomes hydrophilic in nature, thereby releasing the non-native polyeptide into the cavity to commence folding [26]. ATP hydrolysis results in the dissociation of GroES and release of the polypeptide [26]. The integral domain architecture of GroEL is a prerequisite for its conformational switching between a protein capturing to protein folding machine.

1.2.4 Engineering artificial biomolecular switches with integral domain fusion architecture

The ability to construct integral fusion proteins that can modulate the function of one domain by a desired input signal at second domain is an attractive approach for generating novel protein scaffolds for uses in both natural and artificial contexts [28, 29]. Designed protein switches could be used for *in vivo* and *in vitro* sensing of cellular metabolites or redox state [30, 31], as a drug delivery mechanism triggered by specific cellular conditions [28, 32], as a modulator of signal transduction pathways [22] or as energy transfer scaffolds [33].

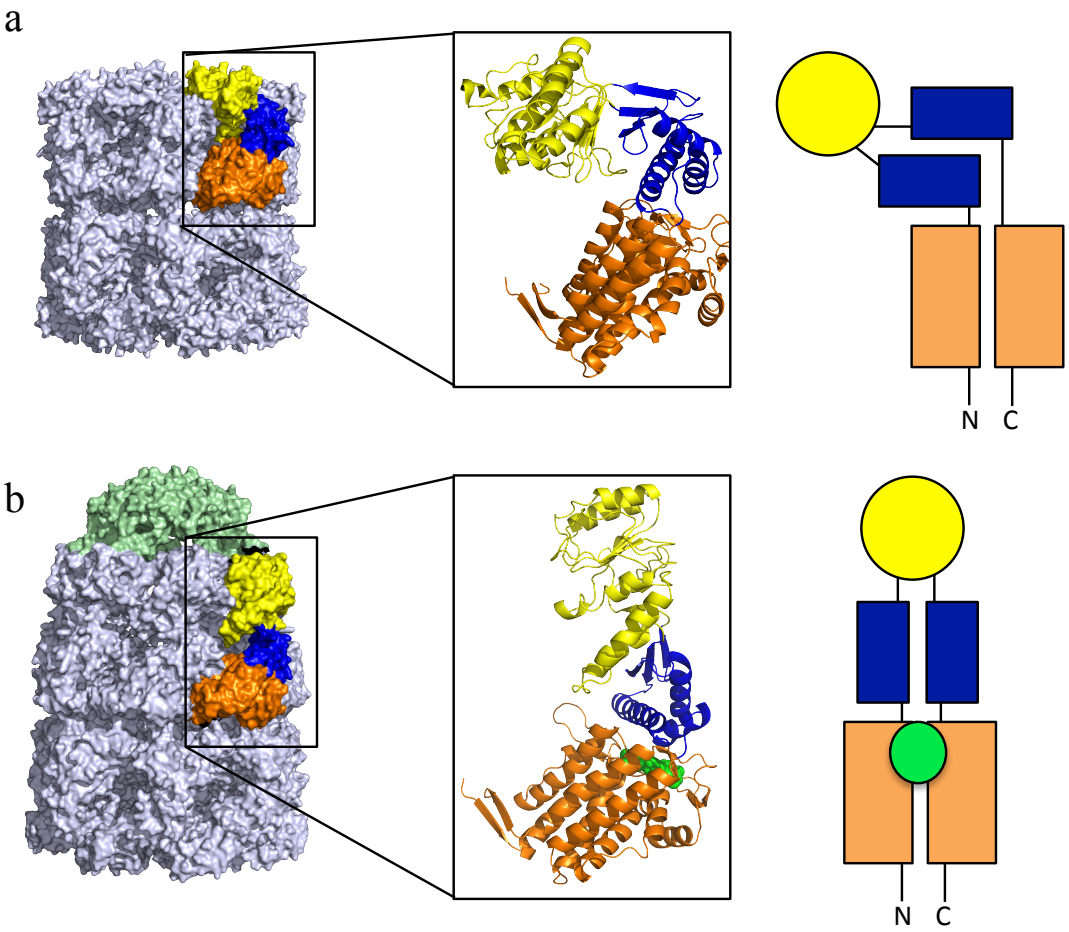


Fig 1.5 Domain insert architecture of the molecular chaperone GroEL. Surface representation (grey) of the multi subunit GroEL with the equatorial (orange), intermediate (blue) and apical (yellow) domains highlighted. Inset, cartoon representation of a single GroEL subunit with a schematic representation of the integral domain architecture shown beside for GroEL in **a,** the trans state and **b,** the cis state. In the cis state the GroEL complex is bound to the accessory complex GroES (pale green). The conformational switch is triggered by ATP (bright green) binding between the equatorial and intermediate domains.

There has been limited success using rational design approaches for the creation of integral fusion proteins that exhibit functional coupling, often resulting in only modest switching magnitudes [32, 34, 35]. The majority of attempts have focused on domain insertion into dynamic loop regions of the parent (accepting) domain so as not to disrupt the tertiary structure and therefore the function of the parent domain [34-36].

However, if the functions of two normally disparate proteins are to be linked simple tolerance to a domain insertion is not enough. Predicting sites within a parent domain that will not only tolerate the insertion of a whole domain but also retain the functions of both domains whilst exhibiting functional coupling with large switching magnitudes is very difficult to achieve. This is because insertion of one domain into a dynamic loop of an accepting domain may be tolerated but to the detriment of functional coupling. A conformational change in the insert domain may be accommodated within the dynamic loop into which the domain is inserted and therefore may not be propogated through to the accepting domain. On the other hand domain insertion into an ordered secondary structure within the accepting domain is less likely to be tolerated but conformational changes in the insert domain are more likely to be propogated through to the accepting domain modulating its function.

An attractive prospect would be to use directed evolution approaches to sample many insertion sites within a parent domain and screen a library for variants displaying switching characteristics. This has been shown previously by coupling various binding events to the activity of β-lactamase [31, 32, 37]. For example, using a transposon-based directed evolution approach, integral domain fusion libraries have been generated by inserting the *E. coli* periplasmic protein cyt b_{562} (Section 1.5) randomly throughout TEM-1 β-lactamase. Several positions within TEM-1 β-lactamase were identified that were tolerant to cyt b_{562} domain insertion. The functions of both domains were retained in the cyt b_{562}-TEM-1 β-lactamase integral fusion scaffolds with functional coupling observed for many of the variants (Fig 1.6). A few variants were identified with up to a 128-fold switch in activity against ampicillin when the cyt b_{562} ligand, haem, was bound to the integral fusion scaffolds.

The domain insertion position can also dictate the nature of the functional couple observed in integral domain fusion scaffolds; a signal induced conformational

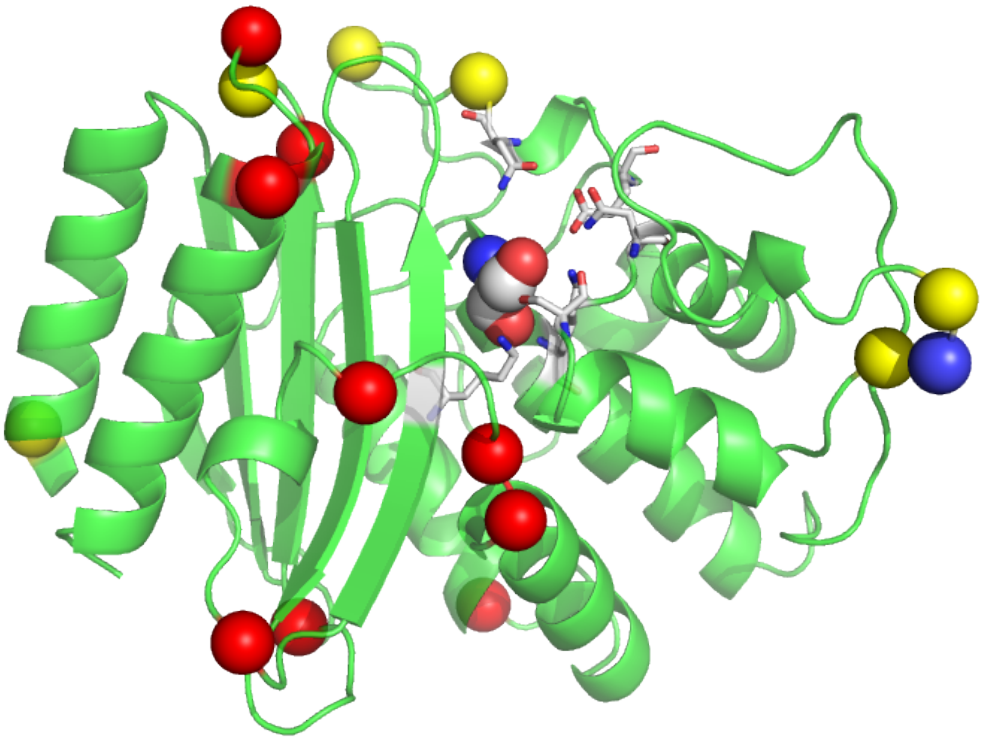


Fig 1.6 Tolerated cyt b_{562} domain insertion positions in TEM-1 β-lactamase. Cartoon representation of TEM-1 β-lactamase (PDB: 1BTL, green) with the active site serine shown in spacefill (CPK colouring) and the rest of the active site residues shown as sticks. Spheres represent positions within TEM-1 β-lactamase tolerant to cyt b_{562} domain insertion. Cyt b_{562} insertion positions resulting in positive or negative modulation of TEM-1 β-lactamase activity in the presence of haem are shown by blue or red spheres respectively. Yellow spheres represent insertion positions where no significant functional coupling was observed between the two domains.

change in the insert domain could modulate the function of the parent domain in a positive or negative manner. For example many cyt b_{562}-TEM-1 β-lactamase integral fusion scaffolds were identified that negatively modulated the TEM-1 β-lactamase activity towards ampicillin in the presence of haem with one variant identified that contained two cyt b_{562} domain inserts and positively modulated TEM-1 β-lactamase activity towards ampicillin (Fig 1.6).

1.3 Directed evolution versus rational design

With an ever-expanding protein engineering toolbox it is possible to design and construct proteins with altered or improved characteristics and to create novel artificial protein scaffolds to perform functions not sampled by the natural protein repertoire. The two main methods of achieving this is by rational site directed mutagenesis or by directed evolution. The limitation of rational protein engineering is often requiring prior knowledge of the target protein structure and function in order to aid the engineering process.

The advancement of directed evolution approaches along with improved screening techniques allows libraries of proteins containing mutations randomly introduced throughout the polypeptide to be created, with variants exhibiting desired characteristics identified for further characterization [38]. The benefit of using a directed evolution approach is that a detailed knowledge of protein structure or function is not always required. Several directed evolution techniques have been developed for creating libraries of mutant protein variants including oligonucleotide based methods, enzyme or chemical mediated DNA cleavage and transposon based approaches [38, 39].

1.3.1 Transposon-based directed evolution approaches

Recent mutagenesis approaches based on the use of an engineered transposon termed Mu has allowed sampling of single amino acid deletions, codon replacements (including for unnatural amino acid incorporation) and whole domain insertion.

1.3.1.1 General structure of Mu transposable elements.

A transposon is a mobile piece of DNA that can be inserted into a target DNA molecule catalyzed by a transposase protein. The Mu transposon from the

bacteriophage Mu has specific inverted repeat sequences at both its 5' and 3' ends, unique to the Mu transposable element [40]. The four terminal regions are specific transposase binding motifs, known as transposase recognition elements (TREs). In Mu they are termed L1, L2, R1 and R2 for the left end (5') or right end (3') TREs respectively (Fig 1.7). The DNA carried between the TREs is essentially non-specific to the transposition reaction and can be any chosen coding sequence such as antibiotic resistance markers for selection of positive insertion events or a DNA cassette encoding desired peptides for insertion (Table 1.1), whole domains [41, 42] or the whole Mu genome [40].

1.3.1.2 Mu transposition mechanism

Although the exact details of the transposition mechanism vary between different transposon/transposase systems the overall mechanism is shared across a number of class II transposable elements. The class II transposons (Mu) can be efficiently inserted into a target DNA sequence via a 'cut and paste' or by 'cleaved donor' meachanism, dependent on whether the transposon comes from a donor DNA or is pre-cleaved respectively (Fig 1.7). The 'cut and paste' mechanism of transposition follows three major reaction steps: (1) donor transposon DNA cleavage, (2) a strand transfer reaction and (3) DNA repair [40]. The 'cleaved donor' mechanism of transposition follows the same reaction steps as the 'cut and paste' mechanism except the first stage concerns the formation of a type I transpososome complex, termed the cleaved donor complex (CDC), with a pre-cleaved transposon [40, 43] (Fig 1.7).

In the first step of the 'cut and paste' mechanism the transposase binds to the TREs and the donor transposon DNA is specifically cleaved, by the transposase nuclease activity, exposing 3'-OH termini at both ends of the transposon (Fig 1.7) [40, 43]. The first step of the 'cleaved donor' mechanism is the formation of a type I transpososome complex with a pre-cleaved transposon already with exposed 3'-OH groups (Fig 1.7).

The transpososome complex is a stable protein bound DNA complex formed by transposase binding to the TREs of transposons and subsequently forming stable transposase oligomers. Several *in vitro* transposon systems have been shown to be able to form stable transpososome complexes including the Mu, Tn5, Tn7 and Tn10

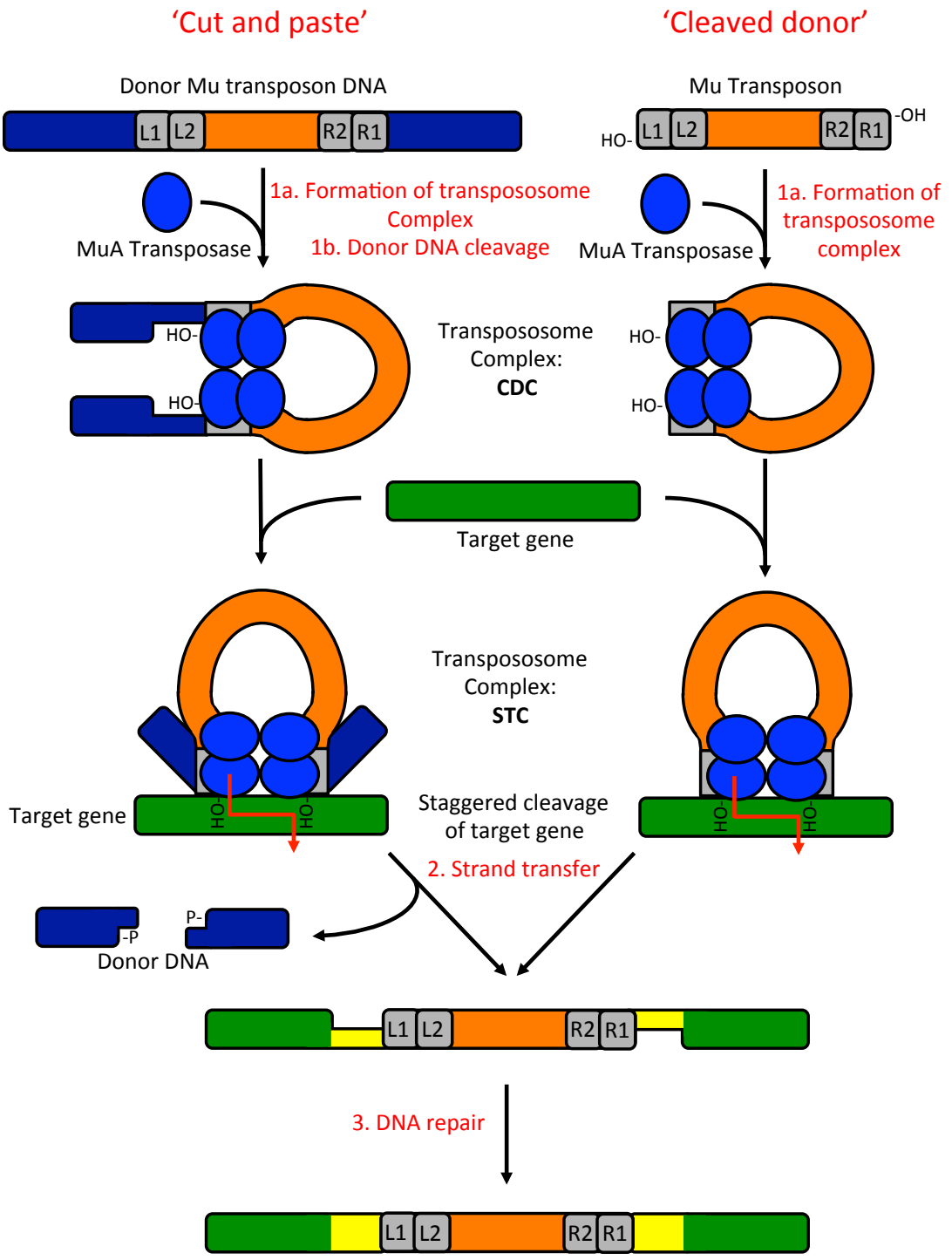


Fig 1.7 Schematic of transposition mechanism. ***Step 1 a*****,** A cleaved donor transpososome complex (CDC) is formed by MuA transposase (blue circle) binding to transposon recognition elements (L1, L2, R1 and R2:grey) at the 5' and 3' ends of the transposon (orange) and forming tetramers. ***Step 1 b,*** The transposon can originate from donor DNA (dark blue), which is cleaved by the transposase revealing the transposon 3'-OH groups ('cut and paste' mechanism), or can be pre-cleaved ('cleaved donor' mechanism). The CDC complex then binds target DNA (green) forming a strand transfer transpososome complex (STC). ***Step 2*****,** The strand transfer reaction takes place producing a staggered cut (red arrow) in the target gene connecting the transposon to the target gene. Due to the staggered cut target site duplications are produced (yellow). ***Step 3*****,** Cellular DNA repair mechanisms fill the gaps in the target gene completing the transposition reaction.

systems [40]. Transpososome complexes can be comprised of transposase dimers bound to the TREs seen in the Tn5, Tn7 and Tn10 systems [44] or transposase tetramers bound to the TREs as in the Mu transposon system described here [43, 44].

The second step for both mechanisms is the strand transfer reaction, which firstly involves the binding of target DNA by the CDC forming a type II transpososome complex [40] termed the strand transfer complex (STC) (Fig 1.7) [43]. Most transposases demonstrate some degree of target site specificity but are still capable of efficiently inserting transposons at many target sites, an exception being the Tn7 transposon, which transposes with high efficiency to a specific site within its bacterial host chromosome [45]. The strand transfer reaction involves the cleavage of the target DNA and joining of phosphoester bonds resulting in connection of the transposon to the target DNA (Fig 1.7) [40]. The strand transfer reaction is achieved by a one-step mechanism by which the exposed 3'-OH groups of the transposon act as nucleophiles attacking phosphoester bonds in the target DNA, resulting in the connection of the 5'-phosphoryl groups of the target DNA to the 3'-OH groups of the transposon [40].

The cleavage sites introduced into the target DNA during the strand transfer reaction are usually staggered due to the relative locations of the exposed transposon 3'-OH groups in the STC with respect to the target DNA [46]. The staggered cleavage of the target DNA is responsible for the target site duplications flanking the transposon (Fig 1.7). The length of the target site duplications is dependent on the distance of the free 3'-OH groups of the transposon from one another and differs depending which transposon system is being used. There is a 5 bp staggered cleavage of the target gene when using the Mu/MuA system. The 5' to 3' exonuclease activity of DNA polymerase I repairs the missing nucleotides reconstituting the target DNA completing the transposition process (Fig 1.7) [40].

1.3.1.3 Use of transposons in directed evolution

Several transposon-based techniques have been utilized to create libraries with insertion events randomly positioned throughout target genes. This relies on the transposon system being used having a low target site specificity. The majority of the methods utilize imperfect restriction digestion, after transposition, to remove the bulk of the transposon from the target gene [47]. The limitation of these methods is that the position of the restriction sites in the transposon often result in the TREs and/or target

site duplications left behind in the target gene (Fig 1.8). Depending on the transposon used and the method of transposon removal DNA sequences left behind can range from 12 - 279 bp. Religation of the single breaks in the target DNA, introduced after transposon removal, can result in the coding sequence producing proteins with peptide insertions ranging from 4 – 93 a.a (Table 1.1).

This technique can be useful for probing general sites within proteins that are tolerant to peptide insertions. However, there is no control over the inserted DNA (TREs and target site duplication) and therefore the resulting encoded peptide insertions are restricted to that encoded by elements of the transposon. Designed DNA cassettes (e.g. encoding desired protein sequences) can be ligated into the random breaks generated after transposon removal, however the linkers connecting the peptides or domains to the target protein would again be encoded by the TREs and target site duplications (Fig 1.8).

There are other methods available for the development of randomly introduced breaks in a target gene. Many of these techniques utilize enzymatic or chemical cleavage of DNA, by DNAseI or Ce(IV)-EDTA respectively, to introduce random breaks in a target gene into which DNA cassettes can then be inserted [39, 48] or potentially used in downstream processes to introduce deletion mutations [39, 48]. However, it is notoriously difficult to generate single cuts in DNA using DNaseI and digestion with this nonspecific nuclease regularly produces tandem duplications and nested deletions of unpredictable sizes within the parent gene [32].

A recently developed transposon-based directed evolution approach [19, 21, 31] can be used to sample different mutagenesis events not normally sampled by traditional approaches: triplet nucleotide deletion (TND) [19, 21], codon replacement [49], domain insertion [31, 37] and unnatural amino acid incorporation [50] (Fig 1.9). This method utilizes the bacteriophage transposon, Mu, and transposase, MuA, which has very low target site preference and can efficiently insert Mu transposons randomly throughout target genes [51]. An engineered Mu transposon, termed MuDel, containing *Mly*I restriction sites 1 bp from its 5’ and 3’ ends has been generated for the purposes of sampling various downstream mutagenesis events. The major advantage of this transposon-based technique over the previously described methods is that *Mly*I restriction digestion after transposition removes MuDel from the library DNA along with its TREs and the target site duplication. A triplet nucleotide

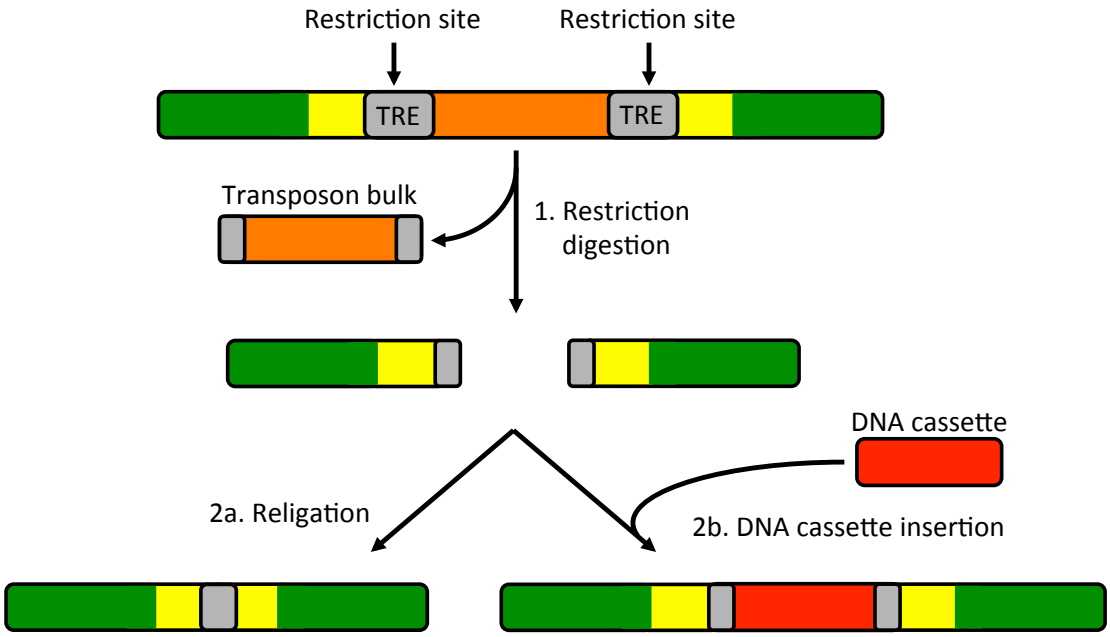


Fig 1.8 Transposon removal by restriction digestion. ***Step 1*****,** Restriction digestion removes the bulk of the transposon (orange) or TREs (grey), depending on the location of the restriction sites, from the target gene (green). This imperfect restriction digestion can leave behind part of the TREs and the target site duplication (yellow). ***Step 2a*****,** Religation of the target gene results in the insertion of part of the transposon, TREs and the target site duplication, of varying sizes depending on the transposon system and removal method used (see Table 1.1). ***Step 2b*****,** Designed DNA cassettes (red) can be ligated into the random breaks produced after transposon removal.

Table 1.1 Transposon systems for peptide insertions

Transposon derivatives	Insertion size (a.a)	Method of transposon removal	Ref
IS21	4 or 11	BglII or SalI restriction	[52]
Tn4430	5	KpnI restriction	[53]
Tn7	5	PmeI restriction	[54]
Mini Mu	5	NotI restriction	[55]
Tn5	24 or 31	NotI or BamHI restriction	[56, 57]
Tn3	45, 89 or 93	Cre-loxP site specific recombination	[58, 59]

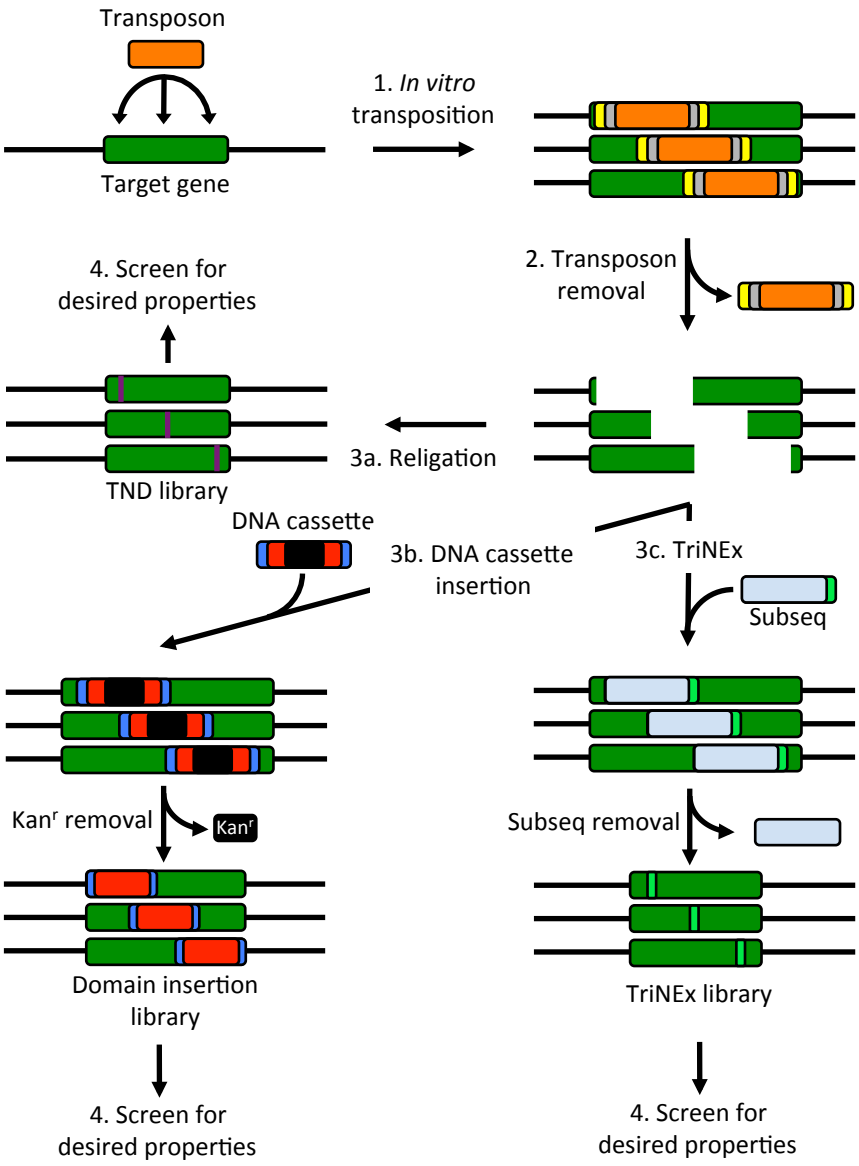


Fig 1.9 Schematic of MuDel transposon based method to library production. ***Step 1*****,** *In vitro* transposition of MuDel (orange) into a target gene (green) results in a library of random insertions with target site duplications (yellow). ***Step 2*****,** *Mly*I restriction digestion removes MuDel from the transposon insertion library, including the TREs (grey), target site duplication and a triplet nucleotide from the target gene, resulting in an insertion site library. ***Step 3a*****,** Religation of the insertion site library results in a triplet nucleotide deletion (TND: purple line) library, for sampling single amino acid deletions. ***Step 3b*****,** Insertion of DNA cassettes encoding a whole protein domain (red) and designed linker sequences (blue), with a kanamycin resistance gene (black) for positive selection, into the random breaks in the insertion site library produces a domain insertion library. ***Step 3c*****,** Insertion of a DNA cassette termed Subseq (pale blue with bright green line) into the insertion site library, followed by removal of the bulk of subseq (pale blue) results in a triplet nucleotide exchange (TriNEx) library, for sampling substitution mutations (bright green line) or incorporation of amber stop codons (TAG) for non-natural amino acid incorporation. ***Step 4*****,** The resulting libraries are then screened for protein variants with desired characteristics.

from the target gene is also removed upon *Mly*I restriction digestion. The mechanism by which this occurs is described in detail in Chapter 3.

The resulting sub library has random breaks throughout the target gene and can be recircularized to produce a triplet nucleotide deletion (TND) library [19]. A DNA cassette encoding a whole protein domain can be ligated into the random breaks for the construction of integral domain fusion scaffolds [37]. Alternatively a DNA cassette, termed Subseq can be inserted within the random breaks in the target gene. Subseq has *Mly*I restriction sites critically placed to remove the bulk of the cassette leaving a triplet nucleotide behind in the target gene resulting in triplet nucleotide exchange (TriNEx) [49, 50] (Fig 1.9).

1.4 Green fluorescent protein (GFP)

1.4.1 General

The discovery of GFP from the jellyfish *Aequorea victoria* has revolutionized real time imaging in live cells [60, 61]. The 27 kDa, 238-residue protein is termed an autofluorescent protein as it exhibits green fluorescence without the need for any additional cofactors. All that is required for fluorescence is correctly folded protein and O_2, which allows the autocatalytic covalent rearrangement of residues S65, Y66 and G67 to form the chromophore: 4-(p-hydroxybenzylidene)imidazolidin-5-one [62]. The autocatalytic formation of the chromophore makes GFP an efficient genetically encoded fluorescent marker [62]. As such it has been used extensively as a tool in cell biology to study gene expression, protein localization, protein trafficking and protein-protein interactions [62]. While there were many envisaged uses of wild type GFP, the inherent properties were not optimized for use in cell biology. Thus, GFP has also been extensively engineered to adapt it for various different uses. This includes, new colour variants, improved brightness, increased stability and biosensing [63].

1.4.2 Chromophore maturation

For chromophore maturation to proceed the polypeptide must undergo four distinct processes: mainchain rearangement (folding), cyclisation, oxidation and dehydration [62, 64, 65]. Work by several research groups have helped to identify the roles of conserved residues, Y66, G67, R96 and E222 in the mechanism of chromophore maturation [66]. There are historically two proposed mechanisms by

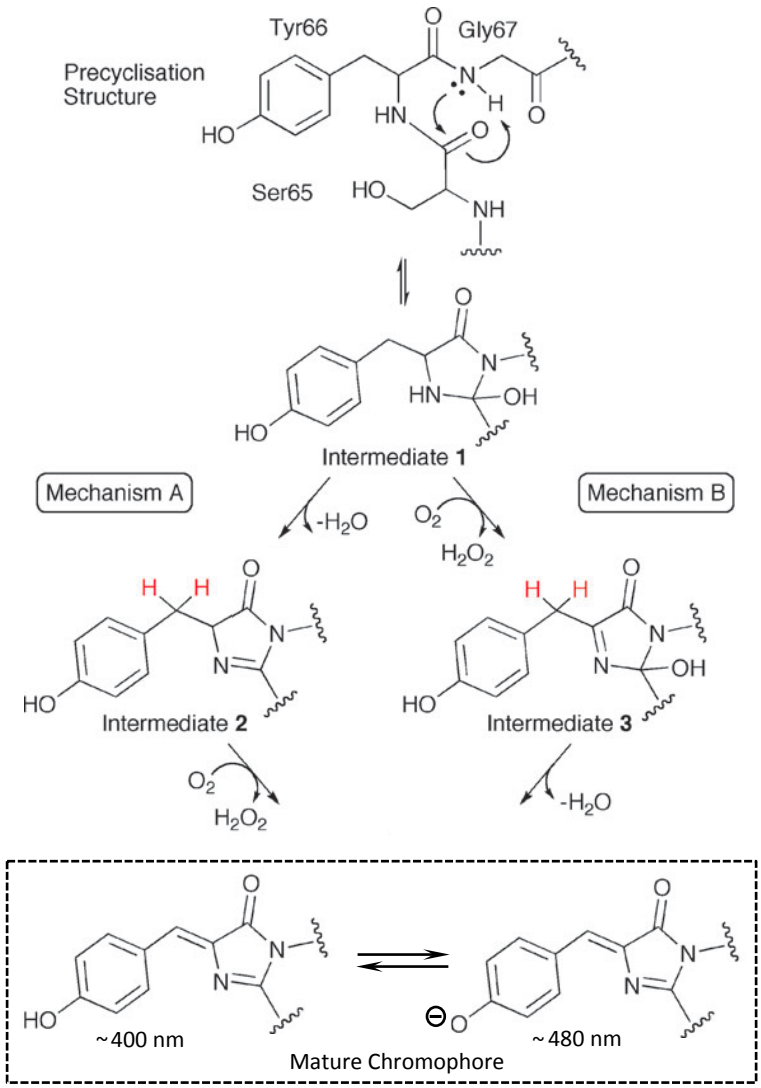


Fig 1.10 Two proposed mechanisms for GFP chromophore maturation. Mainchain rearrangement and cyclisation is either followed by dehydration then oxidation (Mechanism A) or oxidation then dehydration (Mechanism B). Figure adapted from [66]

which the chromophore matures (Fig 1.10) [66]. The first mechanism proposes main chain cyclisation is followed by dehydration of the S65 carbonyl carbon followed by oxidation of the Y66 Cα-Cβ bond introducing full conjugation to the chromophore. Whilst the second proposed mechanism follows main chain cyclisation, oxidation then dehydration [66].

Both mechanisms start with the folding of the protein into the native β-barrel structure with the a distorted α-helix formed to promote main chain reorganization through placement of the carbonyl carbon of S65 in close proximity to the amide nitrogen of G67 (Fig 1.10) [64]. Nucleophilic attack by the amide nitrogen of G67 on the carbonyl carbon of S65 occurs in the precyclized structure forming an imidazalone ring [64]. Conserved residues R96 and E222 have both been implicated in the main chain cyclisation reaction. E222 acts as a general base increasing the nucleophilicity of the G67 amide nitrogen [66], whilst R96 acts to stabilize proposed enolate intermediates during the main chain cyclisation reaction [64]

Computational methods have shown that the main chain cyclisation reaction is energetically unfavourable and therefore requires trapping in a cyclysed conformation either by oxidation or dehydration [67]. It was long thought that the chromophore maturation reaction progressed down the dehydration then oxidation pathway [62]. However, it has recently been shown that the oxidation step can precede dehydration [65]. Oxidation of the enolate intermediate by molecular oxygen results in the production of H_2O_2 in a 1:1 stoichiometry to mature chromophore [65] (Fig 1.10).

Experiments following the increase in fluorescence and the production of H_2O_2 during maturation showed there was a lag between H_2O_2 production and fluorescence maturation [65]. The loss of 2 Da from accumulated intermediates, determined from MALDI-MS of tryptic peptides, confirmed oxidation as being the second step in chromophore maturation [65]. This implies oxidation takes place before dehydration resulting in full conjugation of the chromophore, giving rise to fluorescence. It is highly probable that the chromophore matures by both mechanisms but under highly aerobic conditions the predominant mechanism follows the cyclisation-oxidation-dehydration route.

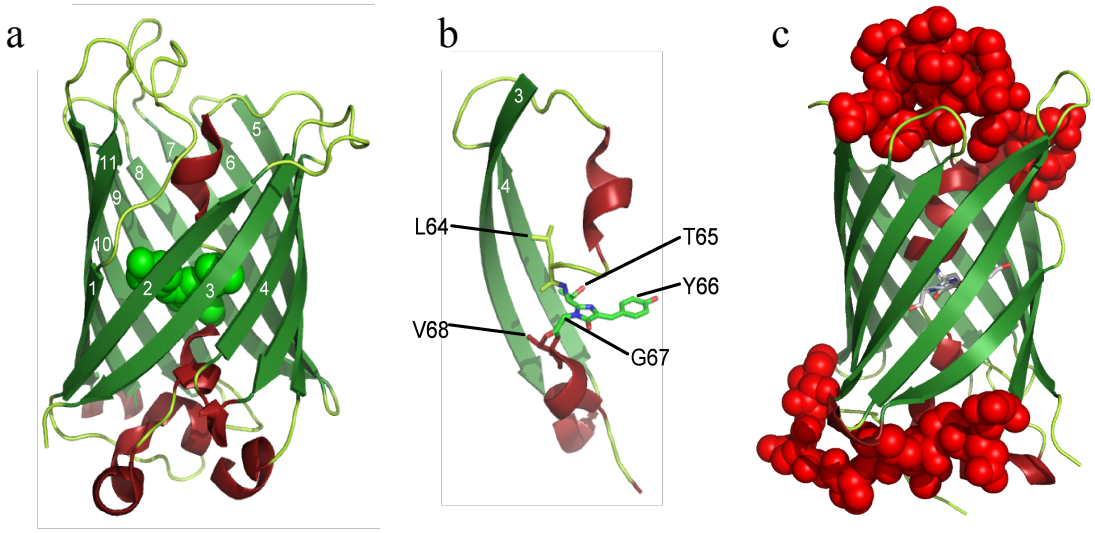


Fig 1.11 Green fluorescent protein structure. a, β-barrel fold of S65T-GFP (pdb:1EMA) comprising of 11 β-strands (dark green) with the chromophore (light green spheres) in the middle of the β-barrel. **b,** Cartoon representation of β-strands 3 and 4 (green) leading into the central α-helix (red) containing the chromophore (green sticks). The rest of the β-barrel has been cut away to reveal the chromophore. **c,** The two longest loops in GFP (red spacefill) form caps over the two ends of the β-barrel protecting the chromophore from the external environment.

1.4.3 GFP tertiary structure

GFP forms a β-barrel (or β-can) structure comprising 11 antiparallel β-strands (Fig 1.11 a) with a distorted α-helix running through the core. The chromophore forming residues reside on the distorted α-helix protected from the external environment by the β-barrel (Fig 1.11 b). The majority of the β-strands forming the β-barrel are connected by short loops between 3 – 6 residues long. There are two longer loops that span the ends of the β-barrel acting as caps, further protecting the chromophore from the external environment (Fig 1.11 c). The chromophore of GFP is therefore encapsulated by the proteins β-barrel structure and is sensitive to the external environment with fluorescence being quenched upon exposure to the external solvent [68]. The chromophore therefore acts as a sensitive probe for the state of the protein. This has led to the exploitation of GFPs to sense changes in their local environment through engineering of efficient biosensors. The chromophore sensitivity to its local environment makes it a model protein for engineering studies as the effect of mutations can be easily observed.

The GFP chromophore forms an extensive hydrogen-bonding network with residues and structured water molecules, in particular Q69, R96, N146, H148, R168, T203, S205 and E222 (Fig 1.12 a) [62]. The hydrogen-bonding network promotes a protonated form of the chromophore (Fig 1.12 a). The hydroxyl group of S65 is capable of donating a hydrogen bond to the carboxylate group of E222 which in turn participates in a hydrogen bond to the hydroxyl group of S205. The hydroxyl group of S205 forms a hyrdogen bond to the phenyl group of the chromophore through a conserved water molecule which is also coordinated by the mainchain carbonyl of N146. In this conformation E222 carries a negative charge and due to electrostatic repulsion results in the chromophore being protonated.

In the S65T mutant, the additional methyl group on the side chain of T65 results in the T65 hydroxyl being in a different position to that of the corresponding hydroxyl group of S65 in wt GFP, due to steric reasons. This results in the T65 hydroxyl group donating a hydrogen bond to the main chain carbonyl of V61. This in turn results in the E222 carboxylate group being protonated and therefore neutral in the ground state. Other residues solvating the chromophore (H148, T203, S205) promote the ionization of the chromophore tyrosyl group, giving rise to altered

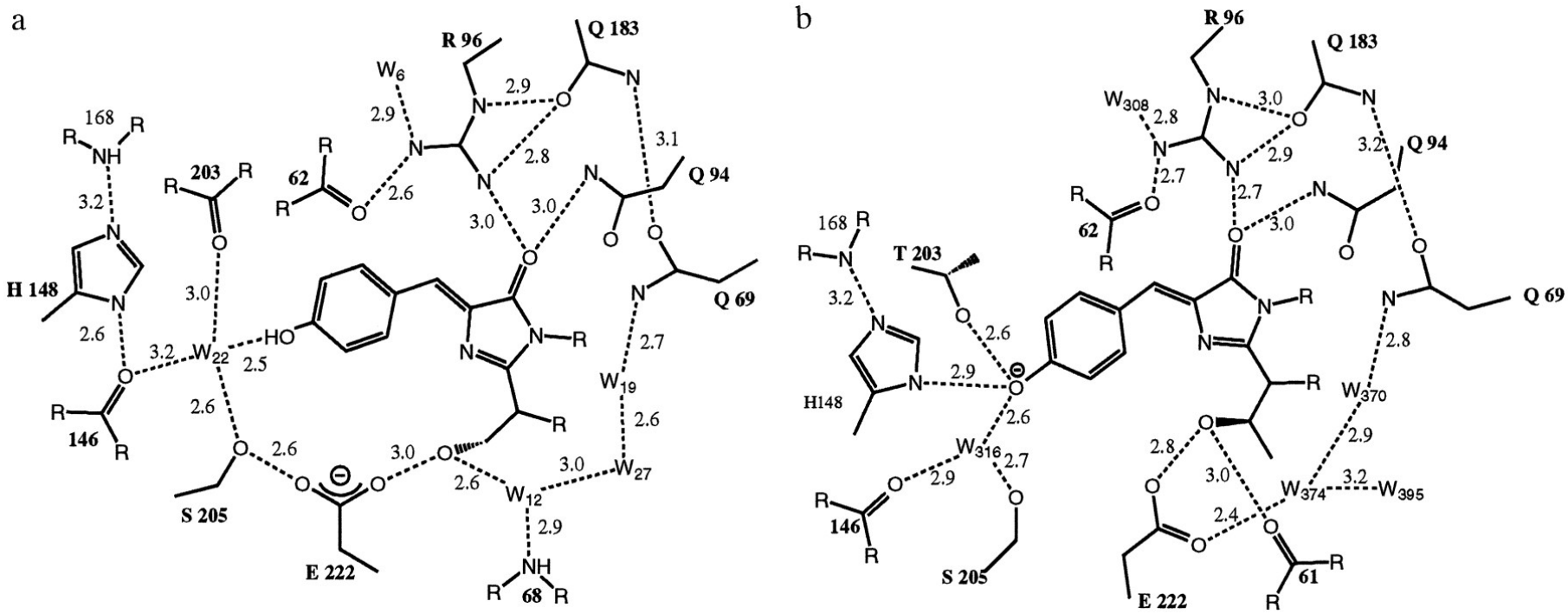


Fig 1.12 Hydrogen bonding network in the chromophore local environment. There are extensive hydrogen bonding networks between the mature chromophore and surrounding residues in both **a,** wild type GFP and **b,** S65T-GFP. Hydrogen bonds are shown as dashed lines with the length of the bonds shown in Å. Figure taken from [69]

spectral properties (Fig 1.12 b). Due to the close proximity (~3.7Å) of the E222 side chain to the chromophore, electrostatic repulsion would forbid both species being in an anionic form simultaneously.

The wt GFP and S65T-GFP chromophores have different spectral characteristics being excited at either ~395 nm or ~490 nm respectively. However both chromophores emit fluorescence at ~509 nm. This is due to an anionic form of the chromophore in the excited state for both wt GFP and S65T-GFP. Excitation at 395 nm results in the deprotonation of the chromophore in the excited state. This is why regardless of whether the chromophore is protonated (Fig 1.11 a) or deprotonated (Fig 1.12 b) in the ground state the emission is always at ~509 nm due to a deprotonated form of thc chromophore in the excited state.

1.4.4 Engineering GFP

As previously mentioned site directed mutagenesis of GFP has generated many colour variants with enhanced fluorescence and stability [23]. Substitution of residue Tyr66 in GFP to either histidine or tryptophan gives rise to blue or cyan fluorescence respectively [14], while substitution of Thr203 for a tyrosine results in π-stacking with the GFP chromophore consequently red shifting emission to give yellow fluorescence [70]. Further substitution mutations have also yielded variants with improved folding characteristics and stability resulting in improved brightness [71].

The two main problems with wild-type GFP were decreased folding efficiency when expressed above room temperature and its excitation by light in the UV range, which is potentially harmful to living cells. Given the obvious interest in fluorescent proteins that can efficiently fold at higher temperatures, GFP has been extensively engineered to identify mutations that confer improved folding and brightness [62]. One of the first advances in the use of GFP was the generation of enhanced GFP (EGFP). EGFP exhibits increased folding efficiency at 37°C and is excited by ~490 nm light. It is still one of the most widely used GFP variants.

EGFP has two substitution mutations compared to the wt GFP; F64L and S65T. F64L increases the rate of folding and stability at 37°C, whilst S65T alters the spectral characteristics [24] and increases the oxidation rate during chromophore maturation ~4-fold [72]. S65T suppresses the major excitation peak at 375 nm

observed for wt GFP and increases the minor excitation peak at 475 nm ~5-fold with a red shift to ~488 nm [62]. The altered spectral properties conferred by the S65T mutation are due to promotion of the anionic form of the chromophore (Section 1.4.2, Fig 1.12) [69].

1.5 Cytochrome b_{562}

Cytochrome b_{562} (Cyt b_{562}) is a 106 amino acid 4 helix bundle protein, found in the periplasm of *Eschericia coli*. It binds a single haem molecule non-covalently, with the S_δ atom of a methionine (Met 7) and the N_ε atom of a histidine (His 102) coordinating the iron moiety [73-75] (Fig 1.13). The antiparallel α-helices form a left-handed bundle separated by 3 interconnecting loops [74]. The haem moiety is situated towards the terminal end of the bundle in a hydrophobic pocket formed by the folded protein [74]. Cyt b_{562} undergoes a major change in conformation on binding haem most notably at the C-terminal helix, which changes from a dynamic unordered structure to a structurally ordered α-helix [73] (Fig 1.13). The apo-cyt b_{562} is essentially a partially folded protein and the binding of haem to cyt b_{562} more than doubles the stability of the protein, increasing the free energy of denaturation by over 3 kcal mol^{-1} [76]. Haem also binds to cyt b_{562} in a redox dependent manor with tighter binding of reduced haem (~10 pM) than oxidized haem (~10 nM). Bound haem has also been shown to be sensitive to oxidative modification [77].

In the holo-cyt b_{562} protein the N- and C-termini are relatively close to one another (~17 Å). In the majority of naturally occurring discontinuous multi-domain proteins the insert domain has its N and C-termini in close proximity (~8 Å) so as not to disrupt the structure and therefore the function of the parent domain [24]. This therefore makes cyt b_{562} a potentially suitable candidate protein for domain insertion.

1.5.1 Cytochrome b_{562} as a sensing domain.

The haem binding properties of cyt b_{562} opens up the possibility of its use as a potential sensor domain. Haem is a biologically important small molecule that acts as a cofactor to many different proteins encompassing a wide range of roles such as; oxygen transport [78], catalysis [79], electron transfer [80] and sensing. It also acts as an efficient fluorescence quencher [81], including that of EGFP [33, 82].

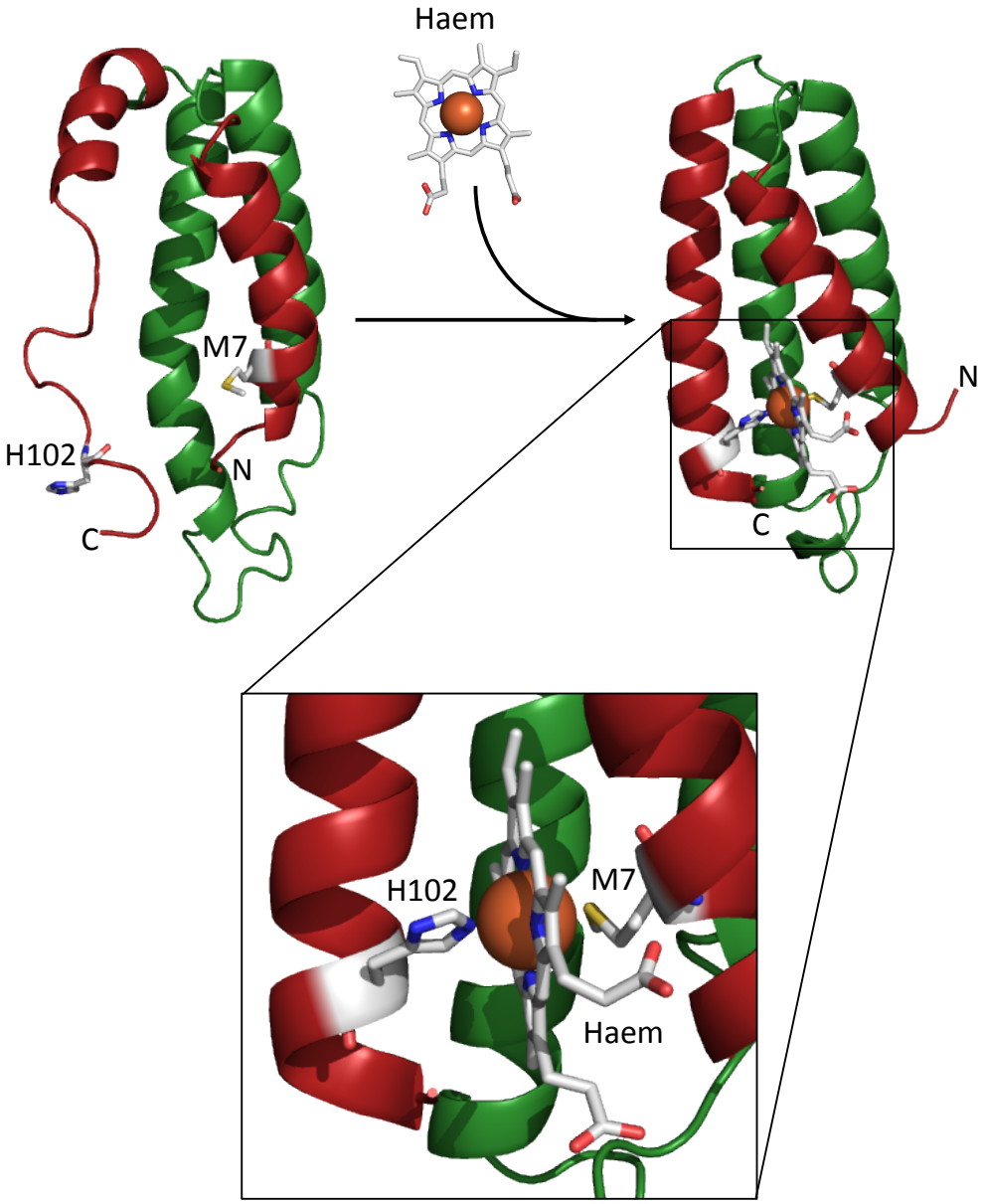


Fig 1.13 Structure of apo and holo-cyt b_{562}. In the apo form the C-terminal region of cyt b_{562} (pdb:1APC) is unstructured and dynamic. Binding of haem non covalently between its axial ligands, a methionine (M7) and histidine (H102), results in a large conformational rearrangement of the C-terminal region into a structured α-helix. **Inset,** close up view of haem coordinated between its axial ligands

Haem when free in solution can be severely toxic with high levels of free haem leading to organ, tissue and cellular injury [83]. Free haem can lead to the oxidation, covalent crosslinking and aggregation of proteins and has been implicated in many pathological states [83]. Therefore construction of a haem sensor would be extremely useful for characterizing levels of free haem in disease states.

The redox-dependent binding of haem, conformational changes upon haem binding and the fact that the N- and C-termini are in close proximity to one another makes cyt b_{562} an attractive candidate as a sensing insert domain. Cyt b_{562} has been used in previous work to link the redox dependent haem binding events to other disparate functions such as DNA binding [84], antibiotic resistance or even as a bio/nanoelectronics component (Fig 1.14) [31, 37].

Attachment of the DNA-binding basic helix region of the leucine zipper protein GCN4 to the N- and C-termini of cyt b_{562} ($NC_{BHR}b_{562}$) conferred site-specific DNA binding activity (Fig 1.14 a) [84]. Further mutagenesis of the loop connecting α-helices 2 and 3 of cyt b_{562} in the $NC_{BHR}b_{562}$ constructs resulted in haem-dependent DNA binding affinity [84].

The use of cyt b_{562} as an insert domain has also been demonstrated [31, 37]. Insertion of the cyt b_{562} domain into various positions throughout TEM-1 β-lactamase resulted in integral fusion scaffolds with haem dependent antibiotic resistance (Fig 1.14 b). The redox properties of cyt b_{562} are also being exploited for potential use in molecular electronics, with electrochemical gating providing a route to modulate current flow through the protein (Fig 1.14 c) [85, 86].

The presence of haem bound to cyt b_{562} in close proximity to the chromophore of EGFP has been shown to quench fluorescence through energy transfer [33]. The quenching of fluorescence by haem proximity to the fluorophore, in conjunction with the conformational changes on redox dependent haem binding may act synergistically to modulate GFP fluorescence.

1.6 Scope of the present project

This thesis focuses on the use of a recently developed transposon based directed evolution approach to sample mutational events in proteins not currently used in the available protein engineering toolbox; single amino acid mutagenesis and

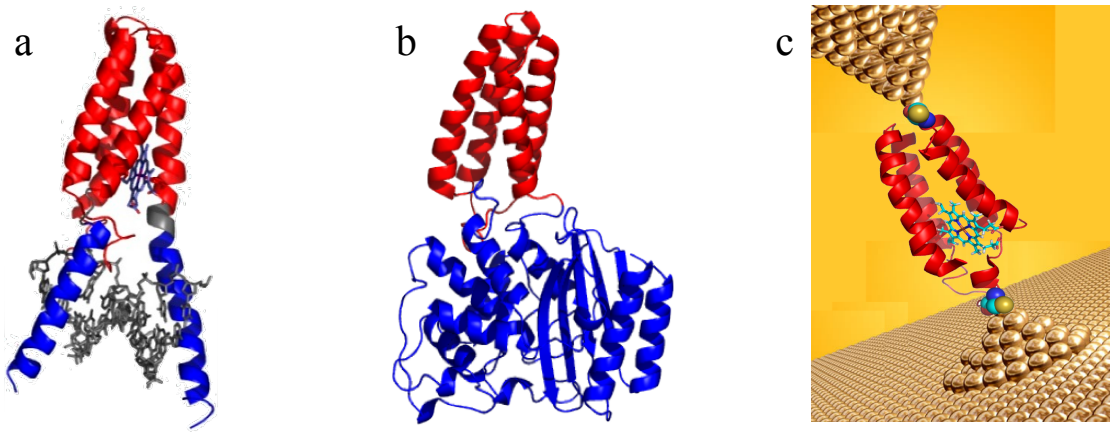


Fig 1.14 Models of engineered proteins involving cyt b_{562}. a, An engineered DNA binding cyt b_{562} with haem dependent DNA affinity. **b,** An integral domain fusion scaffold of cyt b_{562} (red) inserted within TEM-1 β-lactamase (blue) conferring haem dependent antibiotic resistance. **c,** Gold-thiol interactions of an engineered cyt b_{562} for the use as a protein based single molecule transistor.

domain insertion. The thesis will then explore the tolerance and impact of these mutational events, and how novel proteins with useful properties can be generated.

As described in Section 1.2.1 substitution mutations are extensively used to engineer new or improved function into proteins. However, single amino acid deletion mutagenesis is rarely used as a protein-engineering tool due to the dogma that mutational events affecting the protein backbone can be detrimental to structure and therefore function. It is also difficult to predict the structural outcome of an amino acid deletion event therefore requires a directed evolution approach to survey for tolerated mutation positions, which as of now there has not been a suitable technique to introduce deletion mutations randomly throughout a target gene.

The practice of domain insertion as a protein-engineering tool to develop novel protein scaffolds to act as artificial biomolecular switches is also largely underused. Both of these mutational techniques share a common theme in that they both alter the protein backbone by either shortening it or by fragmenting the continuity of the polypeptide chain.

Due to the nature of the mutations being introduced, altering the peptide backbone, it is very difficult to rationally design positions within a target protein that will be tolerant to the effects on the local and global structure. This problem is confounded by a lack of detailed structural information concerning the effects of amino acid deletion mutagenesis or domain insertion. Therefore to identify sites within a target protein that are tolerant to alterations of the backbone structure a transposon based directed evolution approach has been taken to develop libraries of mutant proteins of which can be screened for desired characteristics.

In this study the genetically encoded auto fluorescent protein EGFP has been used as a target protein to study the effects of single amino acid deletion mutagenesis and to assess its tolerance to cyt b_{562} domain insertion, for the construction of potential novel energy transfer scaffolds.

In the first experimental chapter of this thesis (Chapter 3), work is presented on the construction of a transposon insertion library exhibiting insertions randomly positioned throughout the *egfp* gene. The processes behind the construction and subsequent screening of triplet nucleotide deletion (TND) and domain insertion sub libraries have also been described. DNA sequence analysis of isolated variants from both of the sub libraries will help towards identifying a possible target site consensus sequence for the engineered Mu transposon, MuDel, which as of yet has not been

investigated in detail. The outcome of this work was the generation of a diverse transposon insertion library within *egfp*, with the construction of two sub libraries for the sampling of single amino acid deletion mutations or cyt b_{562} domain insertions in EGFP. Sequence analysis implied MuDel has very low target site specificity with no obvious target site consensus sequence.

Chapter 4 investigates the tolerance and impact of EGFP to single amino acid deletion mutagenesis. Variants with altered or improved properties (e.g. fluorescence characteristics, stability, cellular production) that were identified from the TND library were further characterised. The crystal structure for EGFP has also been solved, which during the course of this thesis had not been done before, and will help to describe the effects of single amino acid deletions at the molecular level. High-resolution structure determination of EGFP highlighted the effects of the F64L and S65T mutations at the molecular level. Variants were also identified from the library with improved folding and stability characteristics, resulting in increased cellular fluorescence.

Chapter 5 investigates the tolerance and effect of EGFP to cyt b_{562} domain insertion. A novel variant identified from the library was characterized in much greater detail and provides the basis for Chapter 6. The work in this chapter highlighted the importance of domain insertion position and the magnitude of functional coupling between the two domains. Two variants were also identified that have altered spectral properties and lead to the development of ratiometric fluorescent redox sensors providing the basis for Chapter 7.

Chapter 6 describes the crystal structure determination of a novel cyt b_{562}-EGFP integral fusion scaffold, identified in Chapter 5, with interesting haem mediated quenching characteristics to ascertain at the molecular level reasons behind the observed properties. This will further aid our understanding of integral domain fusions at a molecular level for future scaffold design. Further biophysical characterization of this variant has also been described to identify how whole domain insertion has affected the structure and stability of EGFP.

The final chapter concerns using a rational approach to design a cyt b_{562}-EGFP chimera, identified in Chapter 5, to further develop ratiometric fluorescent redox sensing properties. The redox characteristics for the resulting mutants have been investigated by redox buffer titration and determination of redox kinetics. These experiments examined the influence of the introduced mutations on the redox sensing

capabilities of a cyt b_{562}-EGFP scaffold (CG15). A CG15 double cysteine mutant was identified with the most reducing midpoint described to date and the fastest response rates to the natural cellular reactive oxygen species (ROS) H_2O_2.

The combination of directed evolution, rational design, *in silico* design and X-ray crystal structure determination performed in this thesis can be used to help build a picture of the structural effects of mutations that alter the backbone of a protein. This will in turn aid future design processes for the tailoring of more specific protein constructs.

Chapter 2: Materials and Methods

2.1 Materials

2.1.1 Chemicals

Deionised or MilliQ^{TM} (MQ) water were used throughout. Ampicillin (Melford, Ipswich, Suffolk), kanamycin or chloramphenicol (Duchefa Biochemie, Melford, Ipswich, Suffolk) were made as a 100, 25 or 40 mg/ml stock solution respectively and filter sterilized with a 0.22 μm filter unit (Thermo Fisher Scientific, Loughborough, Leicestershire, UK). The stock solutions were used to supplement bacterial growth media by dilution to an appropriate working concentration of 100 μg/ml ampicillin, 20 or 40 μg/ml chloramphenicol or 25 μg/ml kanamycin. Isopropyl β-D-1-thiogalactopyranoside (IPTG) (Melford, Ipswich, Suffolk) was made as a 100 mM stock solution and filter sterilized with a 0.22 μm filter unit (Thermo Fisher Scientific, Loughborough, Leicestershire, UK). The stock solution was used to supplement bacterial growth media by dilution to an appropriate final concentration.

Haem (Sigma-Aldrich, Dorset, UK) was prepared as a 10 mM stock solution by suspension into 0.5 M NaOH (Sigma-Aldrich, Dorset, UK) and diluted into dH_2O to working concentrations as required. NaOH is used to ionize haem in aqueous solution to prevent aggregation.

KNO_3 (BDH AnalaR, VWR International Ltd., Poole, UK) was prepared as a 100 mM stock solution in dH_2O and diluted into a protein sample to an appropriate working concentration of 1 mM. Ascorbic acid (Sigma-Aldrich, Dorset, UK) was prepared as a 100 mM stock solution in dH_2O and diluted into a protein sample to an appropriate working concentration of 1 mM. KNO_3 and ascorbic acid are used to generate oxidizing and reducing conditions respectively.

H_2O_2 (30% (v/v)) or NaHClO (Sigma-Aldrich, Dorset, UK) was diluted into dH_2O as necessary. Dithiothreitol (reduced DTT) (Melford, Ipswich, Suffolk) and trans-4,5-dihydroxy-1,2-dithione (oxidized DTT) (Sigma-Aldrich, Dorset, UK) stock solutions of 10 mM were prepared in dH_2O and diluted as necessary to make redox buffers for redox midpoint determination.

Agarose (Melford, Ipswich, Suffolk) was of molecular biology grade and prepared by boiling 1.0-2.0 % (w/v) in TAE (Tris Acetate EDTA). Acrylamide and N,N'-methylene bis-acrylamide at a 37.5:1 (w/w) ratio in a (40% w/v) solution (Sigma-Aldrich, Dorset, UK) was used for poly acrylamide electrophoresis (PAGE)

(Section 2.5.1). SDS, APS (Sigma-Aldrich, Dorset, UK) and β-mercaptoethanol (Melford, Ipswich, Suffolk) were used for SDS-PAGE at concentrations stated in Table 2.3. Ammonium sulphate and guanidine hydrochloride (Melford, Ipswich, Suffolk) were used for protein precipitation or equilibrium unfolding analysis respectively.

2.1.2 Chromatographic columns

All chromatographic columns used were supplied by GE Healthcare and used in conjunction with an ÄKTApurifier FPLC. Resource™ Q, Mono™ Q and HiLoad™ Superdex™ 75 pg (preparative grade) columns were used for protein purification whilst Superdex™ 75 GL and Superdex™ 200 GL columns were used for analytical size exclusion chromatography.

2.1.3 Bacterial cell strains

All DNA libraries were grown in *E. coli* NovaBlue GigaSingles™ cells (Merck), *E. coli* DH5α cells (New England Biolabs) were used for carrying and amplifying DNA generated from recombinant methods and all protein expression work was conducted within *E. coli* BL21-Gold (DE3) cells (Stratagene) or *E. coli* TUNER™ (DE3) cells (Novagen) as specified (Table 2.1).

Table 2.1. Species, strains and genotype of bacteria used within different techniques.

Species	Strain	Genotype
E. coli	NovaBlue GigaSingles™	endA1 hsdR17 (rK12-mK12+) supE44 thi-1 recA1 gyrA96 relA1 lac F'[proA+B+ lacIq ZΔM15::Tn10 (TcR)]
E. coli	DH5α	F^- endA1 glnV44 thi-1 recA1 relA1 gyrA96 deoR nupG Φ80d*lacZ*ΔM15 Δ(*lacZYA-argF*)U169, hsdR17(r_K^- m_K^+), λ–
E. coli	BL21-Gold (DE3)	F– ompT hsdS(r – m –) dcm+ Tetr gal λ(DE3) endA Hte
E. coli	TUNER™ (DE3)	F^- *ompT hsdS*$_B$ (r_B^- m_B^-) *gal dcm lacY1*(DE3)

2.1.4 Bacterial growth media

Luaria Bertani (LB) broth medium and LB Agar plates were prepared by dissolving 20 g of granulated LB-Broth (Melford, Ipswich, Suffolk, UK) or 35 g

powdered LB Agar (Sigma Aldrich, Dorset, UK) respectively in 1 L of ultra pure water. M9 minimal medium was comprised of 48 mM Na_2HPO_4, 22 mM KH_2PO_4, 8.5 mM NaCl, 1.8 mM NH_4Cl, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$ and 22 mM glucose. M9 minimal media agar plates were prepared as above supplemented with 1.5% agar (Sigma-Aldrich, Dorset, UK). SOC medium was comprised of 2% (w/v) Bacto Tryptone, 0.5% (w/v) Bacto Yeast Extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$. After sterilisation using an autoclave the media was supplemented with 20 mM glucose. All media were sterilized by autoclave at 121 °C.

2.1.5 Molecular weight standard markers

DNA molecular standard markers (100 bp, 1 kb) and broad range prestained protein standard marker were obtained from New England Biolabs (NEB, Hertfordshire, UK) and stored at -20 °C. Protein standards were boiled for 5 mins before use.

2.2 Molecular biology and recombinant DNA methods

2.2.1 Purification of DNA

When using DNA purification kits, the manufactures guidelines were followed. DNA concentrations and 260/280 nm ratios after purification were estimated using a NanoDrop® ND-1000 UV-Vis spectrophotometer (Thermo Fisher Scientific, Loughborough, Leicestershire, UK)

2.2.1.1 From bacterial cell cultures

All large scale preparations (midipreps) of plasmid DNA, using 50 ml of bacterial culture, were carried out using Qiagen midiprep kits. Small scale preparations (minipreps), using < 5 ml of bacterial culture were carried out using Qiagen miniprep kits. Both methods involve alkaline/SDS lysis of pelleted cells followed by neutralization with a 1-3 M solution of sodium acetate, pH <5.5. The DNA was bound to a silica membrane in the presence of high salt, washed and eluted off the membrane with 10 mM Tris-HCl buffer, pH 8.5.

2.2.1.2 From agarose gel

Purification of DNA from agarose gel slices was performed using the QIAquick® gel extraction kit (Qiagen). Guanidine thiocyanate acts as the chaotropic salt and together with a 50 °C incubation disrupts the agarose gel matrix leading to solubilisation of the agarose. The DNA liberated from the agarose binds to a silica membrane when at a pH < 7.5, is washed then eluted with 10 mM Tris-HCl buffer, pH 8.5.

2.2.1.3 From PCR reactions

Purification of double stranded DNA products from PCR reactions was performed using the QIAquick® PCR purification kit (Qiagen). Guanidine hydrochloride and isopropanol denature polymerases in the PCR reaction mix and allow the efficient binding of DNA >100 bp to a silica membrane when at a pH<7.5. Bound DNA is washed and eluted with 10 mM Tris-HCl buffer, pH 8.5.

2.2.1.4 From other enzymatic reactions

Purification of DNA products from other recombinant DNA methods was performed using a MinElute® reaction cleanup kit (Qiagen). Optimum binding of DNA fragments between 70 bp – 4 kb to a silica membrane is achieved at a pH <7.5. Elution in a small volume (10 µl) of 10 mM Tris-HCl buffer, pH 8.5, results in concentrated pure DNA samples ready for subsequent reactions.

2.2.2 Agarose gel electrophoresis

Analysis and separation of DNA fragments was carried out by agarose gel electrophoresis. Agarose (0.7 – 2% (w/v)) was suspended in TAE buffer (40 mM Tris-acetate pH 9.5 and 1 mM EDTA) with ethidium bromide (0.5 µg/ml). A 1/6 dilution of a stock solution loading buffer (Tris pH 8.0, 40% sucrose (w/v), 0.01% (w/v) bromophenol blue) was added to DNA samples prior to loading onto the gel. Electrophoresis was then performed at 100 V for 30 - 50 mins. DNA bands were visualised using an UV-transilluminator (GelDoc-It Imaging System, Ultra-Violet Products Ltd, Cambridge, UK). Approximate molecular weights were determined using DNA molecular size standards (Section 2.1.5) by comparing the distance

travelled on the gel of the DNA fragments in the sample with the fragments in the marker.

2.2.3 PCR with GoTaq polymerase

DNA amplification with GoTaq DNA polymerase (1.25 U) (Promega Ltd, Southampton, UK) was carried out in 50 mM KCl, 10 mM Tris-HCL pH 9.0, 0.1% (v/v) Triton®X-100 and 1.5 mM $MgCl_2$ with 0.2 mM dNTPs, 2 μM of each primer, template DNA (< 20 ng) made to a total volume of 50 μl. Reactions were placed in a TC-412 thermo cycler (Techne) and raised to 95 °C for 10 min. The following cycle was repeated 25 times: the temperature was taken to 95 °C for 30 sec for double stranded DNA to denature: The temperature was dropped to 55-65 °C for 30 sec to allow the primers to anneal to the template DNA: The temperature was then taken to 72 °C for 1 min/kb of template DNA to allow DNA chain elongation by the polymerase. After 25 cycles the temperature was held at 72 °C for 5 min to complete elongation before being held at 4 °C until removed from the thermo cycler. The reaction (2 μl) was analysed by agarose gel electrophoresis (1.0-2.0% (w/v)) (see 2.2.2). Colony screening by PCR follows the above method with a selected colony suspended within the PCR reaction mix in place of template DNA.

2.2.4 PCR with Phusion polymerase

DNA amplification with Phusion High Fidelity DNA polymerase (1 U) (Finnzymes, Braintree, Essex, UK) was carried out in 10 mM Tris-HCL, pH 8.8, 50 mM $MgCl_2$ and 0.1% (v/v) Triton®X-100 with 0.2 mM dNTPs, 2 μM of each primer, template DNA (< 20 ng) made to a total volume of 50 μl. Reactions were placed in a TC-412 thermo cycler (Techne) and raised to 98 °C for 30 sec. The following cycle was repeated 30 times: the temperature was taken to 98 °C for 15 sec for double stranded DNA to denature: The temperature was dropped to 55-65 °C for 15 sec to allow the primers to anneal to the template DNA: The temperature was then taken to 72 °C for 30 sec/kb of template DNA to allow DNA chain elongation by the polymerase. After 30 cycles the temperature was held at 72 °C for 5 min to complete elongation before being held at 4 °C until removed from the thermo cycler. The reaction (2 μl) was analysed by agarose gel electrophoresis (1.0-2.0% (w/v)) (see 2.2.2).

2.2.5 Oligonucleotides for PCR

The sequence of oligonucleotides (synthesisied by Integrated DNA Technologies, www.IDTDNA.com) used as part of this work are given in Table 2.2.

2.2.6 Restriction digestion

*Nde*I (1 U/µg DNA), *Mly*I (1 U/µg DNA) and *Xho*I (1 U/µg DNA) (New England Biolabs) restriction digests were performed in 1 x NEBuffer 4 (50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate, 1 mM DTT, pH 7.9) supplemented with 0.1 mg/ml bovine serum albumin (BSA) (New England Biolabs) in a total reaction volume of 50 µl , and incubated at 37°C (1 hr/µg DNA). *Nde*I and *Xho*I were heat inactivated at 65 °C for 20 mins. *Pst*I (1 U/µg DNA) (Promega) restriction digests were performed in 1 x Buffer H (90 mM Tris-HCL pH 7.5, 50 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT) supplemented with a final concentration of 0.1 mg/ml acetylated BSA (Promega) made to a total reaction volume of 50 µl with sterile distilled water and incubated at 37°C (1 hr/µg DNA). The *Pst*I was heat denatured at 65 °C for 15 mins.

2.2.7 Phosphorylation and dephosphorylation

Phosphorylation of the 5' ends of linear DNA was performed using T4 polynucleotide kinase (10 U/µl) (NEB). Linear DNA (usually from PCR reactions) was supplemented with 1 µl of T4 polynucleotide kinase, 1 x T4 polynucleotide kinase reaction buffer (7 mM Tris-HCl, 1 mM $MgCl_2$, 0.5 mM DTT pH 7.6) (New England Biolabs) and ATP to a final reaction concentration of 1 mM. The reaction was incubated at 37 °C for 1 hr then heat inactivated at 65 °C for 20 mins.

Dephosphorylation of the 5' ends of linear DNA was performed by the addition of Apex™ Heat-Labile Alkaline Phosphatase (1 µl/µg DNA) (Epicentre) to restriction endonuclease digestion reactions (see 2.2.6) and incubated at 37 °C for the duration of the restriction digest. The phosphatase was heat inactivated by incubating the reaction mixture at 70 °C for 5 mins.

Table 2.2. The name and DNA sequence of oligonucleotides used for PCR based applications.

Name	Sequence (5' to 3')[a]
AJBgfp009	AGCAGACATATGGTGAGCAAGGGCGAGGAGC
AJBgfp010	ATACTCGAGTTACTTGTACAGCTCGTCCATGCCG
DDJdi029	CTGACTCTTATACACATCTAATACCTGTGACGGAAGATC
DDJdi030	CTGACTCTTATACACATCTCTCAGGCATTTGAGAAGCACAC
AS1Primer	CTGACTCTTATACACAAGTCGCGAAAGCGTTTCACGATA
DDJdi023	GGCGGTAGCGCAGATCTTGAAGACAATATGGA
DDJdi024	GCTGCCACCCCTATACTTCTGGTGATAGGCGT
JAJA003	NGGTGGGAGCGCAGATCTTGAAGACAATATGGA
JAJA004	NNGCTCCCACCCCTATACTTCTGGTGATAGGCGT
JAJA005	NNGGTGGGAGCGCAGATCTTGAAGACAATATGGA
JAJA006	NGCTCCCACCCCTATACTTCTGGTGATAGGCGT
WREcbf005	NNSGCAGATCTTGAAGACAATATGGAAACCC
WREcbf006	SNNACGATACTTCTGGTGATAGGCGT
WREcbf007	SNNSGATCTTGAAGACAATATGGAAACCC
WREcbf008	NNSNNACGATACTTCTGGTGATAGGCGT
WREcbf009	NSNNSGATCTTGAAGACAATATGGAAACCC
WREcbf010	NSNNACGATACTTCTGGTGATAGGCGT
pEXP-F	TGCTCACATGTGCGTAGAGG
DDJdi013	GTCTCATGAGCGGA
AJBEGFP119-F	GGTCCTGTGCTGCTGCCC
AJBEGFP120-R	GCCGATTGGGGTGTTCTG
AJBEGFP121-F	GGGATCACTCTCGGCATGGAC
AJBEGFP122-R	GGCGGTCACGAACTCCAGCA
JAJA052	GTAGTTGTACTCCAGCTTGTGCCCCAG
JAJA053	TGCGGTGGGAGCGCAGATCTTGAAG
JAJA054	GAAGTTCACCTTGATGCCGTTCTTCT
JAJA055	TGCATCCGCCACAACATCGAGGAC
JAJA056	GATCTTGAAGTTCACCTTGATGCCGTT
JAJA057	TGCCACAACATCGAGGACGGCAG
JAJA058	GGTGCTCAGGTAGTGGTTGTCGGGCA
JAJA059	TGCTCCGCCCTGAGCAAAGACC
JAJA065	GTCGCCCATATGGTGAGCAAGGAGGAGCTGTT

[a]N refers to A, G, C or T nucleotides, S refers to C or G nucleotides.

2.2.8 Ligation

DNA ligations were carried out using Quick T4 DNA ligase (1 μl/reaction) (New England Biolabs) in 1 x quick ligation reaction buffer (66mM Tris-HCL, 10 mM $MgCl_2$, 1 mM DTT, 1 mM ATP, 15% (w/v) PEG 6000, pH 7.6) (New England Biolabs). Ligation of an insert gene into a plasmid was performed using 50 ng of vector DNA with a 3-fold molar excess of insert in a total reaction volume of 20 μl. Recircularisation (intramolecular) ligation of linear DNA was performed with 50 ng of the linear plasmid. Ligation reactions were incubated at 25°C for 20 mins then the DNA purified using a MinElute kit (Qiagen) (see 2.2.1.4) to remove any residual PEG that may reduce efficiency of the subsequent transformation by electroporation.

2.2.9 Preparation of electro competent cells

An LB broth (10 ml) (see 2.1.4) overnight culture was prepared from a single *E. coli* (Table 2.1) colony and incubated in a shaking incubator (200 rpm) at 37 °C overnight. The 10 ml overnight culture was diluted into two 500 ml cultures of LB broth (see 2.1.4) and grown to an A_{600} of 0.4-0.8. The cells are harvested by centrifugation (1500 x g for 15 mins at 4 °C). The pellet was resuspended in 1l of ice-cold sterile water and harvested by centrifugation (1500 x g for 20 mins at 4 °C). The pellet was resuspended in 500 ml of ice cold sterile water and the previous harvesting step was repeated. The pellet was resuspended in 250 ml of ice-cold sterile water and harvested according to the last step. The pellet was resuspended in ~100 ml ice-cold sterile 10% glycerol and harvested as before. The pellet size was estimated and resuspended in an equal volume of 10% glycerol. The cells were divided into 40 μl aliquots, snap frozen in liquid nitrogen and stored at -80 °C.

2.2.10 Transformation by electroporation

Electrocompetent cells stored at -80°C were thawed on ice. DNA (10 ng – 20 ng) was added to an aliquot of thawed cells (40 μl), mixed and transferred to a pre-chilled, sterile electroporation cuvette and subjected to a 4.5 – 5 ms electrical pulse at 12.5 $kV.cm^{-1}$ field strength using a gene pulser (Bio-Rad laboratories Ltd., UK) with capacitance and resistance set to 25 μF and 200 Ω, respectively. The cells were recovered by the addition of 960 μl room temperature SOC (see 2.1.4) in a sterile tube and incubated (37 °C at 200 rpm) for 1 hr. After recovery the cells were plated on LB

agar (see 2.1.4) supplemented with suitable antibiotics and IPTG if required (see 2.1.1).

2.2.11 Transformation by heat shock

E. coli NovaBlue Giga Singles cells (see 2.1.3) stored at -80 °C were thawed on ice. The cells were gently flicked to resuspend evenly. DNA (10 ng – 20 ng) was added to the cells (50 μl), gently mixed and stored on ice for 5 mins. The cells were heated in a water bath for 30 sec at 42 °C then put back on ice for 2 mins. The cells were recovered by the addition of 950 μl room temperature SOC (see 2.1.4) in a sterile tube and incubated (37 °C at 200 rpm) for 1 hr. After recovery the cells are plated on LB agar (see 2.1.4) supplemented with suitable antibiotics (see 2.1.1).

2.2.12 MuDel, OE-Tn5 and ME-4A-Tn5 preparation.

Construction of MuDel by PCR and cloning into pENT was previously described [21]. To prepare MuDel for *in vitro* transposition it was liberated from the plasmid pENT-MuDel by *Bgl*II restriction digestion, kindly supplied by Dr Wayne Edwards. OE-Tn5 was constructed by amplification of MuDel as template DNA by phusion PCR (see 2.2.4) with primer AS1primer (Table 2.2). ME-4A-Tn5 was constructed by amplification of MuDel as template DNA by phusion PCR (see 2.2.4) with primers DDJdi029 and DDJdi030 (Table 2.2). After amplification both OE-Tn5 and ME-4A-Tn5 were purified from their respective PCR reactions using a QIAquick® PCR purification kit (see 2.2.1.3). Both OE-Tn5 and ME-4A-Tn5 transposons were phosphorylated (see 2.2.7) before use and the reaction cleaned up with a MinElute® reaction cleanup kit (see 2.2.1.4).

2.2.13 *In Vitro* MuDel transposition

In vitro transposition reactions using the MuDel transposon were performed using HyperMu^{TM} MuA Transposase (1 μl/ reaction) (Epicentre) in a 1 x reaction buffer (0.15 M potassium acetate, 0.05 M Tris-acetate pH 7.5, 0.01 M magnesium acetate and 4 mM spermidine) with 300 ng of target DNA (pNOM-XP3-*EGFP*, see 2.2.15) and 40 ng MuDel transposon (see 2.2.12) made to a total reaction volume of 20 μl with dH_2O. The reaction was incubated at 37 °C for 3 hrs. The reaction was terminated by the addition of 1 x stop solution (0.1% SDS) followed by incubation at

70 °C for 10 mins. *E. coli* NovaBlue Giga Singles cells (Table 2.1) were transformed (see 2.2.11) with 1 μl of transposition reaction mixture.

2.2.14 *In Vitro* Tn5 transposition

In vitro transposition reactions using the Tn5 transposons (see 2.2.12) using EZ-Tn5^{TM} Transposase (1 μl/reaction) (Epicentre) was performed in a 1 x reaction buffer (0.15 M potassium acetate, 0.05 M Tris-acetate pH 7.5, 0.01 M magnesium acetate and 4 mM spermidine) with 200 ng of target DNA and a molar equivalent of Tn5 transposon, made up to a final reaction volume of 20 μl. The reaction was incubated at 37 °C for 3 hrs. The reaction was terminated by the addition of 1 x stop solution (0.1% SDS) followed by incubation at 70 °C for 10 mins. *E. coli* NovaBlue Giga Singles cells (Table 2.1) were transformed (see 2.2.11) with 1 μl of transposition reaction mixture.

2.2.15 Cloning of *EGFP* into pNOM-XP3

The gene encoding EGFP (*EGFP*) from plasmid pEGFP-N3 (Clontech) was amplified by PCR with GoTaq polymerase (see 2.2.3) at an annealing temperature of 60 °C using primers AJBgfp009, which incorporates a *Nde*I restriction site to the 5' end of *EGFP*, and AJBgfp010 (Table 2.1), which incorporates a *Xho*I site to the 3' end of *EGFP* (N-*EGFP*-X). N-*EGFP*-X was digested with *Nde*I and *Xho*I (see 2.2.6) and purified using a QIAquick PCR purification kit (Qiagen) (see 2.2.1.3). Vector pNOM-XP3 was digested with *Nde*I and *Xho*I (see 2.2.6) and separated on a 1.0% (w/v) agarose gel (see 2.2.2). The DNA band at an expected size of 2143 bp was cut from the gel and extracted using a QIAquick gel extraction kit (Qiagen) (2.2.1.2). Digested N-*EGFP*-X and pNOM-XP3 were ligated (see 2.2.8) together to produce vector pNOM-XP3-*EGFP* used to transform electrocompetent (see 2.2.10) *E. coli* $DH5_{\alpha}$ cells (Table 2.1), plated on LB agar (see 2.1.4) supplemented with ampicillin (see 2.1.1) and incubated at 37 °C overnight. Colony screening by PCR (see 2.2.3) using primers pEXP-F and DDJdi013 (Table 2.2) was carried out to assess whether *EGFP* cloning into pNOM-XP3 was successful. *Nde*I and *Xho*I restriction digestion (see 2.2.6) of pNOM-XP3-*EGFP* was performed and separated on a 1.0% (w/v) agarose gel (see 2.2.2) to confirm the restriction sites had been maintained, as they are required for a later library construction step (see 2.3.2).

2.2.16 DNA Sequencing

DNA samples were sequenced at the Cardiff University Molecular Biology support units DNA sequencing core. Sample preparation and submission was carried out as to the specifications found online at http://probe.biosi.cf.ac.uk/seq/.

2.3 Library construction

The entire library construction process has been outlined in Figure 2.1 with reference to materials and methods sections at each stage.

2.3.1 Transposon insertion library construction

In vitro transposition into pNOM-XP3-*EGFP* with MuDel (see 2.2.13) or Tn5 (see 2.2.14) transposons was performed and used to transform NovaBlue GigaSingles^{TM} (transformation efficiency > 10^9 cfu/μg DNA) (see 2.2.11). A single LB agar plate (see 2.1.4) supplemented with 20 μg/ml chloramphenicol (see 2.1.1) was prepared for each transformed transposition, a percentage of cells were plated and the rest used to inoculate 1l LB broth (see 2.1.4) supplemented with 20 μg/ml chloramphenicol (see 2.1.1) and incubated at 37°C overnight shaking at 200 rpm. Pooled library DNA was purified from the liquid cell cultures (see 2.2.4) and stored at -20 °C. The pooled DNA libraries are named based on the target gene and transposon used to generate them: pNOM-XP3-*EGFP*Δ^{MuDel}, pNOM-XP3-*EGFP*$\Delta^{OE\text{-}Tn5}$ or pNOM-XP3-*EGFP*$\Delta^{ME\text{-}4A\text{-}Tn5}$.

2.3.2 Isolation of variants with transposons within *egfp*

Restriction digestion of library DNA (3 μg) with *Nde*I and *Xho*I (see 2.2.6) was separated by agarose gel electrophoresis (see 2.2.2). Restriction digest produces 4 DNA fragments: vector with or without a transposon and *EGFP* with or without a transposon. The two bands corresponding to vector without a transposon and *EGFP* with a transposon were isolated and gel extracted (see 2.2.1.2). The two purified fragments were ligated (see 2.2.8) together and used to transform (see 2.2.11) *E. coli* NovaBlue Giga Singles cells (Table 2.1). A percentage of the transformed cells were grown on LB agar (see 2.1.4) supplemented with 20 μg/ml chloramphenicol (see 2.1.1) and the rest were used to inoculate 1L LB broth (see 2.1.4) supplemented with 20 μg/ml chloramphenicol (see 2.1.1) and incubated as described in section 2.3.1.

Pooled library DNA was purified from the liquid cell cultures (see 2.2.1.1) and stored at -20 °C. Libraries with transposons within *EGFP* were named, as before, after the target gene and the transposon *EGFP*Δ^{MuDel} and *EGFP*Δ$^{OE\text{-}Tn5}$ (Fig 7).

2.3.3 Triplet nucleotide deletion (TND) library construction

*Mly*I restriction digestion (see 2.2.6) was performed on *EGFP*Δ^{MuDel} DNA (3 μg) to remove MuDel from the pooled plasmid library and analysed by 1.0% (w/v) agarose gel electrophoresis (see 2.2.2). The linear library DNA was purified from the agarose gel using a QIAquick® gel purification kit (see 2.2.1.2). The purified linear library DNA was recircularised by intramolecular ligation with Quick T4 DNA ligase (see 2.2.8) and the reaction cleaned up with a MinElute reaction cleanup kit (see 2.2.1.4). The ligation reaction mixture (1 μl) was used to transform (see 2.2.10) *E. coli* BL21-Gold (DE3) cells (Table 2.1). The transformed cells were grown on LB agar plates (see 2.1.4) supplemented with 100 μg/ml ampicillin and 150 μM IPTG (see 2.1.1) and incubated at 37 °C for 24 hrs then stored at 4 °C. Colonies presenting a green colour phenotype upon illumination on a UV transilluminator and colonies with no colour phenotype were selected for a colony PCR screen (see 2.2.3) with primers pEXP-F and DDJ013 (Table 2.2). The PCR products produced (2 μl) were analysed by agarose gel electrophoresis (see 2.2.2) and the rest purified using a QIAquick PCR purification kit (see 2.2.1.3) for DNA sequence analysis, to identify the nature of the triplet nucleotide deletions.

2.3.4 Production of cybC cassettes

A cyt b_{562} DNA cassette (*cybC)* was constructed in previous work [6] that contains a kanamycin resistance gene (*kan*^{r}) flanked by *Pst*I restriction sites in its open reading frame (*cybC-kan*^{r}). This allows for selection of successful cloning of the DNA cassette into the insertion site library. *CybC-kan*^{r} cassettes were produced by PCR (see 2.2.4) with oligonucleotides coding for GGS or random single amino acid (X) linkers attached to the 5' and 3' ends of the cassette (*cybC-kan*^{r}*-GGS*, *cybC-kan*^{r}*-X*). Vector pWRE-22 containing the *cybC-kan*^{r} cassette was used as a template at an annealing temperature of 62 °C. Additional random nucleotides were also added to the 5' or 3' end of some of the cassettes to be able to sample out-of-frame insertion positions and therefore utilize all three reading frames (*cybC-kan*^{r}*-GGS1, cybC-kan*^{r}*-*

GGS2, cybC-kan^r-GGS3, cybC-kan^r-X1, cybC-kan^r-X2, cybC-kan^r-X3). Primer combinations and generated cassettes produced are depicted in Table 3.2, Chapter 3 with oligonucleotide sequences in Table 2.2. All six *cybC* cassettes were phosphorylated (see 2.2.7) and subsequently purified with a MinElute reaction cleanup kit (see 2.2.1.4).

2.3.5 Domain insertion library construction

The *EGFP*Δ^{MuDel} library digested with *Mly*I (see 2.3.3) was dephosphorylated (see 2.2.7), to avoid intramolecular ligation, and separated by agarose gel electrophoresis (1.0% (w/v)) (see 2.2.2). The dephosphorylated linear library DNA was purified from the agarose gel using a QIAquick® gel purification kit (see 2.2.1.2). *CybC-kan^r-GGS* and *cybC-kan^r-X* cassettes in all three reading frames (see 2.3.4) were ligated (see 2.2.8) with the previously prepared linear library DNA in 6 separate ligation reactions.

The ligation products were used to transform *E. coli* BL21-Gold (DE3) cells (Table 2.1), by electroporation (see 2.2.10), a percentage of which were grown on LB agar (see 2.1.4) supplemented with 25 μg/ml kanamycin (see 2.1.1). The rest of the transformed cells were used to inoculate separate 50 ml LB broth (see 2.1.4) cultures supplemented with 25 μg/ml kanamycin (see 2.1.1), incubated at 37 °C shaking (200 rpm) overnight. The DNA from all six *cybC* libraries (GGS1, GGS2, GGS3, X1, X2, X3) was purified from the cell cultures (see 2.2.1.1) and stored at -20 °C. A proportional quantity of library DNA from each reading frame was pooled dependent on the number of variants obtained from each liquid culture. This way no particular reading frame is over represented within the pooled libraries (GGS_{pool} or X_{pool}).

GGS_{pool} or X_{pool} (500 ng) were restriction digested with *Pst*I (see 2.2.6) to remove the kanamycin resistance gene from the libraries. The restriction digests were separated by agarose gel electrophoresis (1.0% (w/v)) (see 2.2.2) and the DNA bands corresponding to linear library DNA isolated and purified by gel extraction (see 2.2.1.2). The purified library DNA was recircularised by intramolecular ligation (see 2.2.8) producing *cybC-EGFP-GGS* and *cybC-EGFP-X* libraries, used to transform *E. coli* TUNER^{TM} (DE3) cells (Table 2.1) by electroporation (see 2.2.10). The transformed cells were grown on LB agar (see 2.1.4) supplemented with 100 μg/ml

ampicillin and 150 μM IPTG (see 2.1.1), and incubated for 24 hrs at 37 °C then stored at 4 °C.

2.3.6 Library screening

After the 37 °C overnight incubation colonies presenting a green colour phenotype was visualized by excitation with a UV-transilluminator. Fluorescent colonies were noted and the plates stored at 4 °C. This process was repeated two more times after 24 hrs and 72 hrs at 4 °C. Colonies that presented a green colour phenotype after the initial 37 °C incubation and colonies that gained a green colour phenotype after 24 hrs and 72 hrs at 4 °C were screened by colony PCR (see 2.2.3). A plasmid specific primer, pEXP-F, and a *cybC* specific primer, WREcbr012, (Table 2.2) were used at an annealing temperature of 60 °C. PCR products between 456 bp and 1167 bp can be produced depending on the position of insertion within *EGFP* (see Fig 3.9, Chapter 3). DNA for sequencing the chimeric genes, to identify the nature of the insertions, was generated by PCR (see 2.2.3) with primers pEXP-F and DDJdi013 (Table 2.2) at an annealing temperature of 60 °C.

2.4 Rational design by site-directed mutagenesis

2.4.1 Cloning of *eyfp* into pNOM-XP3 and construction of the single amino acid deletion mutant eYFPΔG4

The gene coding for enhanced yellow fluorescent protein (*eyfp*) was amplified by PCR with GoTaq polymerase (see 2.2.3) using the plasmid pEYFP-N1 (Clontech) as template DNA with primers AJBgfp009 and AJBgfp010 (Table 2.2). The primers add *Nde*I and *Xho*I restriction sites to the 5' and 3' end of *eyfp* respectively. The DNA product was purified from the PCR reaction using a QIAquick PCR purification kit (see 2.2.1.3). The purified DNA fragment was then digested with *Nde*I and *Xho*I restriction endonucleases (see 2.2.6) and purified using a MinElute reaction cleanup kit (see 2.2.1.4). pNOM-XP3 linearised by *Nde*I and *Xho*I restriction digestion (see 2.2.6 and 2.2.15) was used in an intermolecular ligation (see 2.2.8) with the digested *eyfp* DNA fragment to produce pNOM-XP3-*eyfp*. *E. coli* BL21-Gold (DE3) (Table 2.1) were transformed, by electroporation (see 2.2.10), with 1 μl of the ligation reaction and plated on LB agar plates (see 2.1.4) supplemented with 100 μg/ml

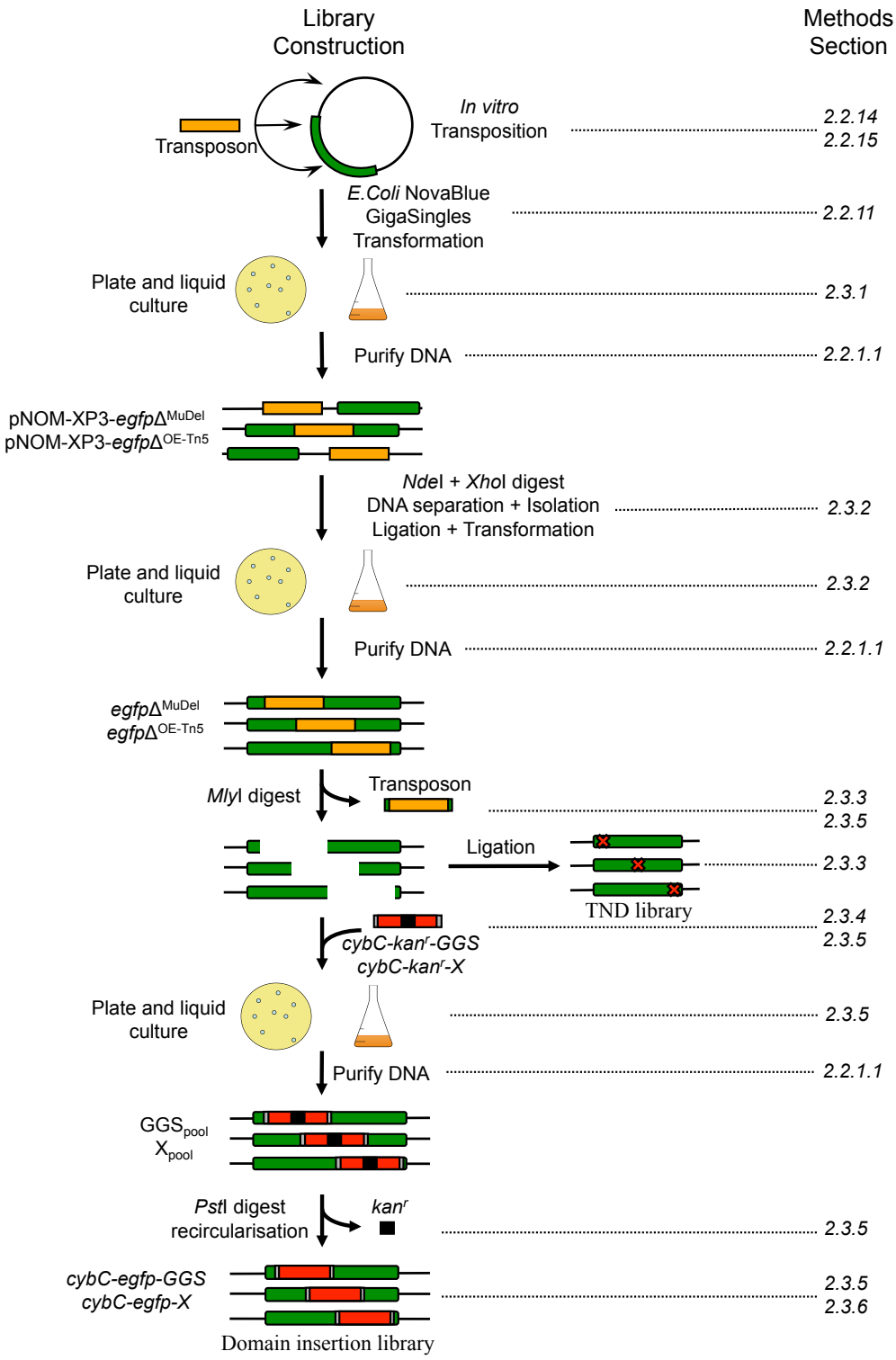


Fig 2.1. Schematic of TND and domain insertion library construction with references to method sections. Vector (black), transposons (orange), *EGFP* (green), *cybC* (red), kanamycin resistance gene (*kan*^{r}) (black), nucleotides coding for GGS or random single amino acid linkers (grey).

ampicillin and 150 μM IPTG (see 2.1.1). A positive clone was selected by its yellow colour phenotype, after illumination with a UV transilluminator, and the presence of *eyfp* within pNOM-XP3 confirmed by DNA sequence analysis (see 2.2.16).

Site directed mutagenesis was used to remove the triplet nucleotide coding for glycine 4 (G4) in eYFP (enhanced yellow fluorescent protein) to assess the transferability of this mutation to a different fluorescent protein. PCR was carried out using GoTaq polymerase (see 2.2.3) with pEYFP-N1 (Clontech) as template DNA and primers JAJA065 and AJBgfp010 (Table 2.2). The primers incorporate *Nde*I and *Xho*I restriction sites to the 5' and 3' end respectively while removing the triplet nucleotide coding for G4 (JAJA065). The PCR fragment was then cloned into pNOM-XP3 as described above to produce pNOM-XP3-*eyfp*-G4Δ.

2.4.2 Construction of double TND mutations

The double amino acid deletion variant EGFP G4Δ D190Δ was constructed using the Phusion site directed mutagenesis kit (New England Biolabs) (see 2.2.4) with primers AJBEGFP119-F and AJBEGFP120-F (Table 2.2) and pNOM-XP3-*EGFP*-G4Δ as template DNA (Fig 2.3). PCR products were analysed by 1.0 % (w/v) agarose gel electrophoresis (see 2.2.2) with subsequent extraction of the DNA product using a QIAquick® gel extraction kit (see 2.2.1.2) (Fig 2.2).

The DNA fragment was phosphorylated (see 2.2.7) and recircularized by intramolecular ligation with Quick T4 DNA ligase (see 2.2.8). The double amino acid deletion variant EGFP G4Δ A227Δ was constructed as described above but with primers AJBEGFP121-F and AJBEGFP122-R (Table 2.2). The ligation reactions (1 μl) were used to transform *E. coli* BL21-Gold (DE3) cells (Table 2.1) by electroporation (see 2.2.10).

The transformed cells were grown on LB agar plates (see 2.1.4) supplemented with 100 μg/ml ampicillin and 150 μM IPTG (see 2.1.1) and incubated overnight at 37 °C. Colonies presenting a green colour phenotype were selected for a colony PCR screen (see 2.2.4) with primers pEXP-F and DDJ013 to produce a DNA fragment for sequence analysis (see 2.2.16) to confirm the presence of the desired mutations.

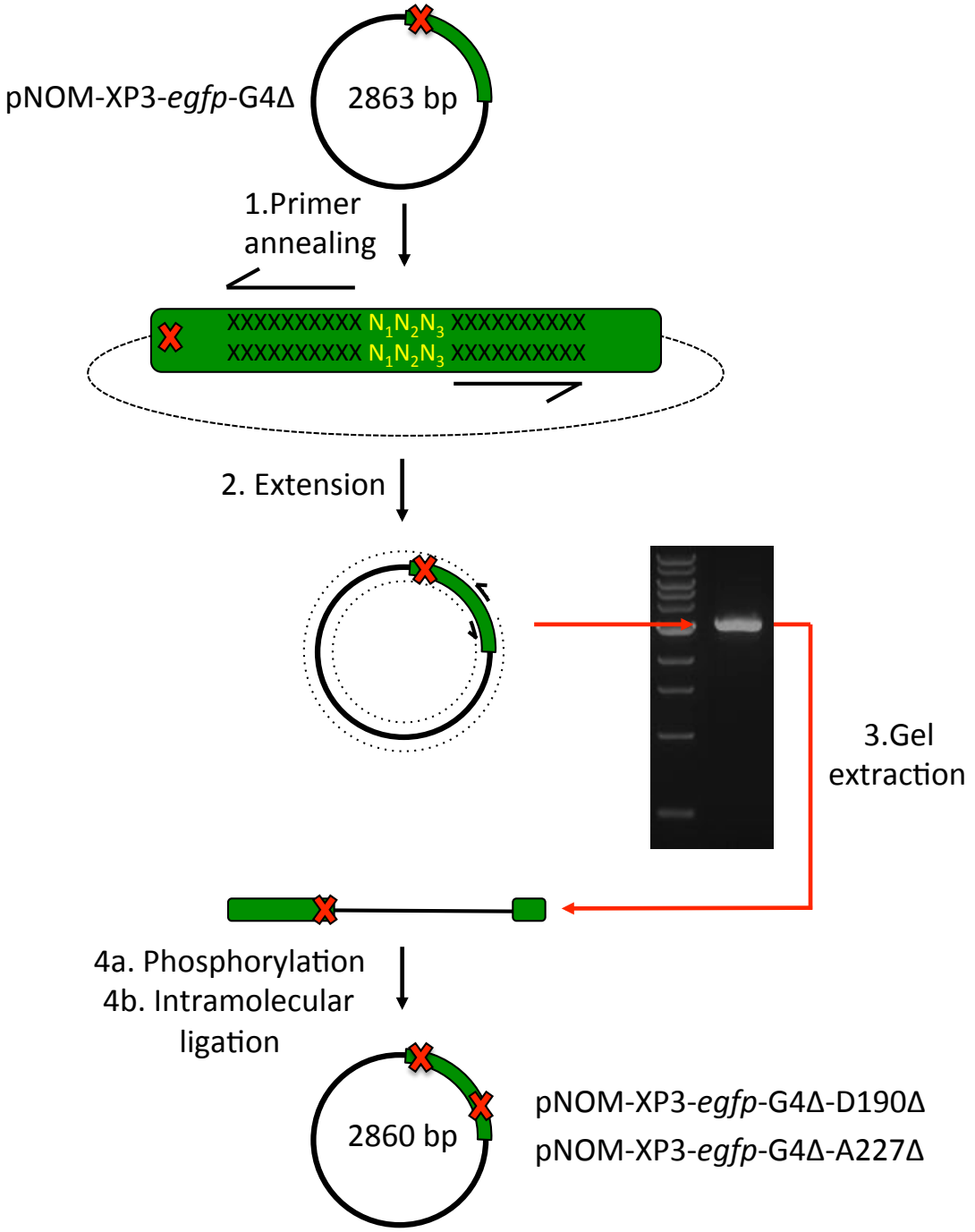


Fig 2.2. Site directed mutagenesis strategy for construction of double TND variants. ***Stage 1*****,** Primers (black arrows) anneal to the target gene (pNOM-XP3-*EGFP*-G4Δ: green, TND:red cross), back to back, with the desired triplet nucleotide to be removed, not coded for by either primer. ***Stage 2*****,** Extension by Phusion polymerase (see 2.2.4) (black dotted line) produces a linear DNA product missing the triplet nucleotide. ***Stage 3*****,** The PCR product is separated by 1.0% (w/v) agarose gel electrophoresis (see 2.2.2) and purified from the agarose gel with a QIAquick gel extraction kit (see 2.2.1.2). ***Stage 4a*****,** The Purified linear DNA fragment is phosphorylated (see 2.2.7). ***Stage 4b,*** The purified linear DNA fragment is recircularised by intramolecular ligation with Quick T4 DNA ligase (see 2.2.8) resulting in a double triplet nucleotide deletion (red crosses) variant of *EGFP*.

2.4.3 Constuction of CG15 double cysteine mutants

The cyt b_{562}-EGFP chimera, CG15, identified by library screening (see 2.3.6), has a single cysteine residue at position 147 introduced as a secondary point mutation of S147 due to an out of frame *cybC* cassette insertion into the *EGFP* insertion library (see 2.3.5). Site directed mutagenesis was used to introduce secondary cysteine point mutations to construct four different CG15 double cysteine mutants ($CG15^{CC}$) (Fig 2.3).

The phusion site directed mutagenesis kit was used (see 2.2.4) with pNOM-XP3-CG15 as template DNA with primers JAJA052 and JAJA053 for N146C mutation, JAJA054 and JAJA055 for K166C mutation, JAJA056 and JAJA057 for R168C mutation or JAJA058 and JAJA059 for Q204C mutation (Table 2.2). PCR products were separated by agarose gel electrophoresis (1.0% (w/v)) (see 2.2.2) and purified from the agarose gel with a QIAquick gel extraction kit (see 2.2.1.2). The DNA fragments were phosphorylated (2.2.7) and used in intramolecular ligation reactions (see 2.2.8) to recircularise the plasmids.

The ligation reactions (1 µl) were used to transform *E. coli* BL21-Gold (DE3) cells (Table 2.1) by electroporation (see 2.2.10). The transformed cells were grown on LB agar plates (see 2.1.4) supplemented with 100 µg/ml ampicillin and 150 µM IPTG (see 2.1.1) and incubated overnight at 37 °C. Colonies presenting a green colour phenotype were selected for a colony PCR screen (see 2.2.4) with primers pEXP-F and DDJ013 to produce a DNA fragment for sequence analysis (see 2.2.16) to confirm the presence of the desired mutations.

2.5 Methods for protein production, purification and analysis

2.5.1 Sodium dodecylsulphate polyacrylamide gel electrophoresis (SDS-PAGE)

SDS-PAGE was performed using the mini-PROTEAN 3 electrophoresis system (Bio-RAD, Hertfordshire, UK). The resolving and stacking gels were comprised of the components listed in Table 2.3. Sample loading buffer (6x) contained 2% (w/v) SDS, 0.2 M Tris-HCL pH 6.8, 0.04% (w/v) bromophenol blue, 8% (w/v) glycerol and 10% (v/v) β-mercaptoethanol. The electrophoresis buffer contained 28.8 g glycine, 8.0 g Tris Base made up to 2 L with distilled water and SDS to 0.1% (w/v). Protein samples (usually 0.1-0.25 µg) or whole cell extracts (see 2.5.2) were dissolved in sample buffer (1X final concentration), heated at 95 °C for 10 mins

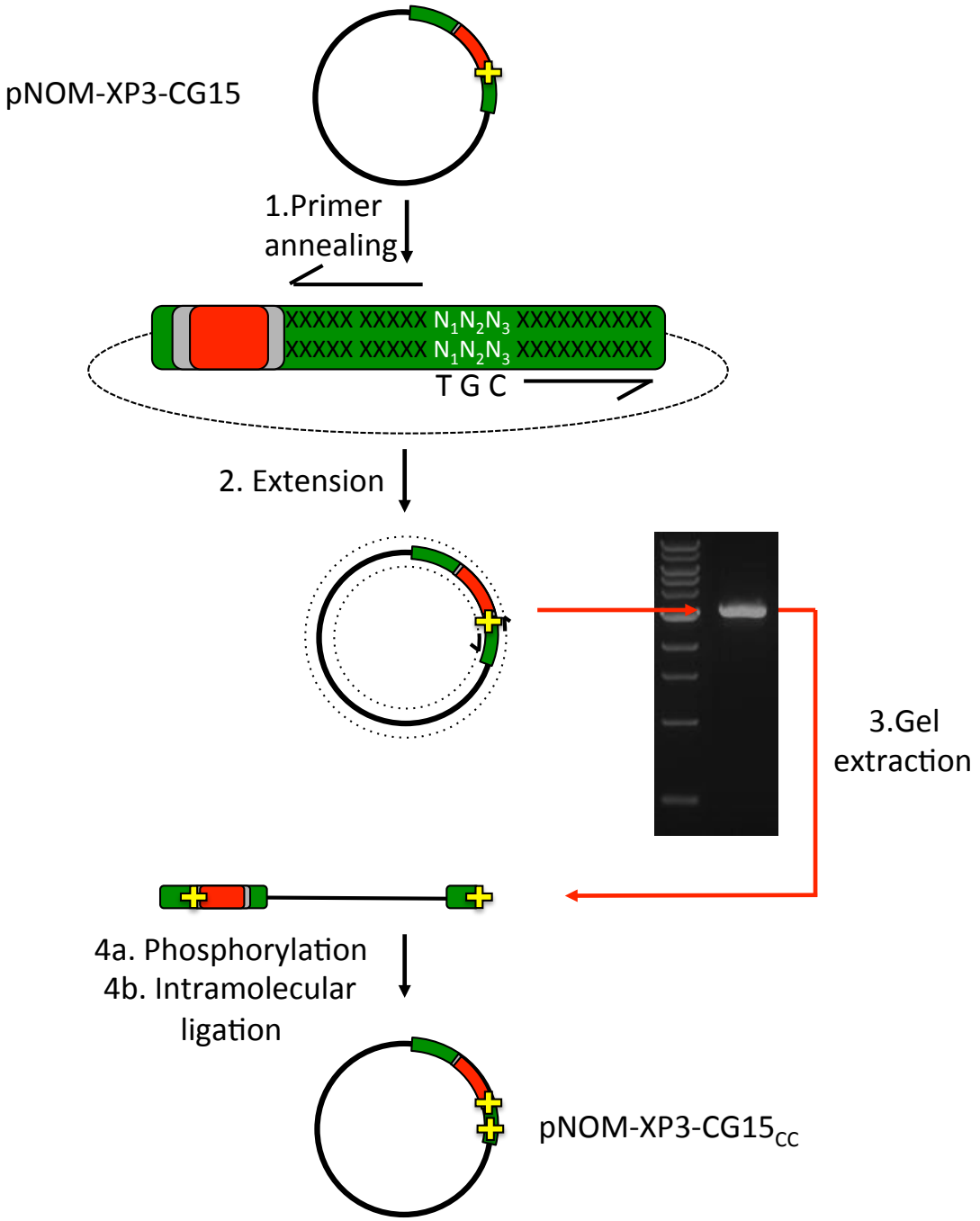


Fig 2.3. Site directed mutagenesis strategy for construction of CG15 double cysteine variants. ***Stage 1*****,** Primers (black arrow, and black arrow with TGC codon) anneal to the target gene (pNOM-XP3-CG15: green/red/grey, not to scale, TGC codon:yellow plus sign), back to back, with the desired TGC codon coded by one primer. ***Stage 2*****,** Extension by Phusion polymerase (see 2.2.4) (black dotted line) produces a linear DNA product with the TGC point mutation. ***Stage 3*****,** The PCR product is separated by 1.0% (w/v) agarose gel electrophoresis (see 2.2.2) and purified from the agarose gel with a QIAquick gel extraction kit (see 2.2.1.2). ***Stage 4a*****,** The Purified linear DNA fragment is phosphorylated (see 2.2.7). ***Stage 4b,*** The purified linear DNA fragment is recircularised by intramolecular ligation with Quick T4 DNA ligase (see 2.2.8) resulting in a CG15 gene coding for a double cysteine (yellow plus signs) variant.

and applied to the gels. The gels were run according to the manufactures recommendations.

The gels were stained using 50% (v/v) methanol, 10% (v/v) acetic acid with R250 Coomassie blue (0.1% (w/v)) and destained using 20% (v/v) methanol, 10% (v/v) acetic acid. Protein standards (Fig 3.1c) were run on every gel for molecular size comparison.

Table 2.3. Concentrations for components of SDS-PAGE resolving and stacking gels.

Component	Resolving Gel	Stacking Gel
Acrylamide/bis-Acrylamide[a]	12.5% (w/v)	5% (w/v)
Tris-HCL	0.375 mM, pH8.8	65 mM, pH6.8
SDS	0.1% (w/v)	0.2% (w/v)
APS	0.05% (w/v)	0.1% (w/v)
TEMED	0.02% (w/v)	0.02% (w/v)

[a]Acrylamide and N,N'-methylene bis acrylamide (40% w/v) solution in a ratio of 37.5:1 (w/w).

2.5.2 Preparation of culture aliquots for SDS-PAGE analysis

Cell aliquots (1 ml) were harvested in a microfuge at 13000 rpm for 2 min and the supernatant discarded. The pellets were resuspended in 1 x SDS sample loading buffer (Section 2.5.1) to an apparent $O.D._{600}$ of 10, to standardize the number of cells between samples, and heated to 95 °C for 10 mins.

2.5.3 Expression of EGFP and single amino acid deletion mutants

LB Broth (15 ml) (see 2.1.4) supplemented with 100 μg/ml ampicillin (see 2.1.1) was inoculated with a single *E. coli* BL21-Gold (DE3) colony (Table 2.1) containing a relevant plasmid (pNOM-XP3 containing the *EGFP* gene or TND *EGFP genes* (see 2.3.3)) and incubated overnight (37 °C shaking at 200 rpm). Some of the overnight culture (5 ml) was used to inoculate 1 l of LB broth (see 2.1.4) supplemented with 100 μg/ml ampicillin (see 2.1.1), which was incubated (37 °C shaking at 200 rpm) until an $O.D._{600}$ of 0.4-0.8 was reached. Sterile IPTG (1 mM, final concentration) (see 2.1.1) was added, to induce protein expression, and incubated for 24 hrs at 37 °C. After induction the 1 l culture was harvested by centrifugation (3000 x g for 20 mins) and the pellet resuspended in 25 ml 50 mM

Tris-HCl, pH 8.0 at 25 °C, 1 mM PMSF and 1 mM EDTA ready for the cells to be lysed by French Press (see 2.5.5).

2.5.4 Expression of cyt b_{562}-EGFP chimera proteins

LB Broth (15 ml) (see 2.1.4) supplemented with 100 μg/ml ampicillin (see 2.1.1) was inoculated with a single *E. coli* TUNER^{TM} (DE3) colony (Table 2.1) containing a relevant plasmid (pNOM-XP3-*EGFP*, pNOM-XP3-CG1 through to pNOM-CG12, pNOM-XP3-CG15 or pNOM-XP3-CG15^{CC}) (see 2.3.5 and 2.4.3) and incubated for 6 hrs (37 °C shaking at 200 rpm). A 1/100 dilution of the starter culture (10 ml) was used to inoculate 1 l M9 minimal media (see 2.1.4) supplemented with 100 μg/ml ampicillin (see 2.1.1) and grown (20 °C at 200 rpm) overnight until an optical density of A_{600} = 0.4 was achieved. The overnight culture was supplemented with 22mM glycerol and protein expression was induced by the addition of IPTG to a final concentration of 300 μM (see 2.1.1) and incubated for 48 hrs (20 °C at 200 rpm). After induction the 1 l culture was harvested by centrifugation (3000 x g for 20 mins) and the cell pellet resuspended in 25 ml 50 mM Tris-HCl pH 8.0 at 25 °C, 1 mM PMSF and 1 mM EDTA ready for the cells to be lysed by French Press (see 2.5.5).

2.5.5 Cell lysis by French Press

E. coli BL21-Gold (DE3) or TUNER^{TM} (DE3) cells (Table 2.1) were lysed using a French pressure cell press. A cell suspension (see 2.5.3 and 2.5.4) was put in a chilled pressure cell. Pressure (1010 psi) was applied to the pressure cell and the cell suspension was released through an aperture at a flow rate of ~1 drop/sec and collected. This process was repeated to assure complete cell lysis. The lysate was then centrifuged, at 10,000 rpm in a Beckman JA-20 rotor for 20 min, to pellet any cell debris and the supernatant was decanted and stored at 4 °C.

2.5.6 Buffer exchange by dialysis.

Any ammonium sulphate remaining in a sample after protein precipitation (see 2.5.8) was removed from the sample by dialysis. Excess salt left in the protein sample inhibited binding to ion exchange chromatography columns. Protein samples (5 ml) were buffer exchanged by overnight incubation in dialysis membrane (MWCO: 12-14

kDa, Spectra Por) suspended in 4 l of 50 mM Tris-HCl, pH 8.0 at 25 °C (Buffer A) under constant stirring.

2.5.7 Protein sample concentration and buffer exchange

Protein samples were concentrated and buffer exchanged using 15 ml Amicon® Ultra Centrifugal concentrators (MWCO: 10 kDa, Merck Millipore) in a T41 spin out rotor, for an IEC CL30R centrifuge (Thermo Fisher Scientific), at 4100 rpm. The buffer used for buffer exchange was dependent on the intended application of the protein sample.

2.5.8 Protein purification

The supernatant from cell lysate (see 2.5.5) was subjected to fractionation with ammonium sulphate precipitation. Ammonium sulphate was added to 45% saturation at 4 °C with constant stirring, using a magnet flea, for 30 min. The precipitate was removed by centrifugation at 10,000 rpm in a Beckman JA-20 rotor for 40 min. The supernatant was then taken up to 75% saturation at 4 °C with constant stirring for 30 min. The suspension was centrifuged at 10,000 rpm in a Beckman JA-20 rotor for 40 min and the resulting pellet resuspended in Buffer A (5 ml). The sample was buffer exchanged into fresh Buffer A by dialysis (see 2.5.6) to remove any remaining ammonium sulphate. A precipitate formed during dialysis and was removed by centrifugation at 10,000 rpm in a Beckman JA-20 rotor for 20 min.

The supernatant was applied to a Resource Q anion exchange column (5 ml bed volume, flow rate 2 ml/min) equilibrated with Buffer A. Target proteins were eluted using a gradient from 0 mM to 500 mM NaCl in Buffer A over five column volumes. Fractions were monitored for absorbance at 280 nm and 488 nm, and analysed by SDS-PAGE (see 2.5.1). Fractions containing the target protein were pooled and concentrated using Amicon® Ultra centrifugal concentrators (see 2.5.7).

The concentrated sample was made up to 2 ml with Buffer A supplemented with 150 mM NaCl (Buffer B) and applied to a HiLoad^{TM} Superdex^{TM} 75 pg gel filtration column (120 ml bed volume, flow rate 0.5 ml/min) (see 2.1.2). Fractions were monitored for absorbance at 280 nm and 488 nm, and analysed by SDS-PAGE (see 2.5.1).

Wild type EGFP and single amino acid deletion mutants were deemed to be >95% pure at this stage and fractions containing the target proteins were pooled and buffer exchanged into Buffer A (see 2.5.7). Fractions containing chimeric proteins after gel filtration were pooled, concentrated and buffer-exchanged into Buffer A (2 ml) (see 2.5.7).

The samples were applied to a Mono Q ion exchange column (2 ml bed volume, flow rate 1 ml/min) (see 2.1.2) equilibrated with Buffer A. Chimeric proteins wee eluted with a gradient from 0 mM to 500 mM NaCl in Buffer A over ten column volumes. Fractions were monitored for absorbance at 280 nm and 488 nm, and analysed by SDS-PAGE (see 2.5.1). Fractions containing pure protein were pooled, concentrated and buffer-exchanged into Buffer A (2 ml) (see 2.5.7).

2.5.7 Determination of protein concentration: colorimetric assay

Protein concentration was determined with the DC Protein assay kit (Bio-Rad) using bovine serum albumin (BSA) (NEB) as a protein standard. The assay was performed as to the manufactures guidelines for use in a microplate assay.

2.6 Methods for the analysis of EGFP, single amino acid deletion mutants of EGFP and cyt b_{562}-EGFP chimeric proteins

2.6.1 Spectroscopy techniques

2.6.1.1 Fluorescent spectral scans

All fluorescence studies were performed using a Cary Eclipse fluorescence spectrophotometer (Varian). Excitation and emission spectra of cell lysates (Section 2.5.5) were measured in a cuvette of dimensions 5 x 5 mm, 10 nm excitation and emission band pass at a scan rate of 600 nm/min. Excitation scans were measured by monitoring emission at 511 nm and emission was measured after excitation at 488 nm. Samples were prepared by dilution into Buffer B to give a final emission intensity of 500 a.u. at 511 nm.

Whole cell fluorescence spectroscopy was performed on *E. coli* BL21-Gold (DE3) (Table 2.1) cell cultures after expression of EGFP or single amino acid deletion variants of EGFP. Expression cultures were harvested by centrifugation (1500 x g for 10 mins) and all supernatant removed and discarded. The cell pellet was resuspended in 50 mM Tris-HCl, pH 8.0 at 25 °C, 150 mM NaCl and 10% (v/v) glycerol to an $O.D._{600}$ = 0.1 in a 1 cm pathlength cuvette. The resuspended cells were transferred to

a cuvette with 5 x 5 mm dimensions and excitation and emission spectra measured as described above.

2.6.1.2 Haem titration fluorescent quenching analysis

Crude cell lysates (see 2.5.5) were diluted into Buffer B supplemented with 1 mM ascorbic acid or KNO_3 (Section 2.1.1) to give reducing or oxidising conditions respectively, with starting fluorescence emission intensities of 100 a.u. at 511 nm. Haem solutions (Section 2.1.1) were titrated into the protein sample as necessary. Single wavelength measurements were taken with excitation at 488 nm monitoring emission at 511 nm with excitation and emission band pass of 10 nm in a cuvette of dimensions 5 x 5 mm. After addition of haem, the samples were mixed and incubated at room temperature (25 °C) for 10 mins before measurements were taken.

2.6.1.3 CG1 and CG6 redox mediated fluorescent switching

For fluorescence switching from oxidizing to reducing conditions or from reducing to oxidizing conditions crude cell lysates (Section 2.5.5) were measured in a cuvette with 5 x 5 mm dimensions, excitation and emission band passes of 10 nm monitoring fluorescence at 511 nm after excitation at 488 nm. For oxidizing to reducing conditions crude cell lysates were diluted into Buffer B with 1 mM KNO_3 (Section 2.1.1) to give a starting fluorescence emission of 100 a.u. at 511 nm, after excitation at 488 nm. Haem (Section 2.1.1) was titrated into the solution so that protein and haem concentration in the sample were equal (determined to be 20 nM). Ascorbic acid (10 mM) was added to the sample and the fluorescence emission at 511 nm monitored for 6 hrs.

Switching from reducing to oxidising conditions was monitored in the same way as above except that 1 mM ascorbic acid (see 2.1.4) was added to the initial sample instead of KNO_3. Haem (20 nM) was titrated into the sample and after 10 min incubation H_2O_2 (between 0.0-0.02 % (w/v)) (see 2.1.4) was added. Fluorescence emission at 511 nm was monitored at 25 °C for 20 hrs after excitation at 488 nm.

Switching from reducing to oxidisng conditions was also performed with purified CG6 protein samples (20 nM) (see 2.5.8) in Buffer B with different oxidizing agents. The experiment was run as before but with the addition of either 0.02% (w/v) H_2O_2, 0.02% (v/v) NaOCl, or 10 mM KNO_3 to induce oxidizing conditions. Switching from reducing to oxidizing conditions was also performed with purified

CG1 (20 nM), as with CG6 above, with 0.02% (w/v) H_2O_2 to induce oxidizing conditions.

Kinetic rate constants were determined by fitting the fluorescence intensity data to a single exponential function (Equation 17)

2.6.1.4 Monitoring UV-Vis absorption spectra for H_2O_2 induced CG6 switching

UV-Vis absorption spectra were measured alongside the fluorescence measurements on CG6 switching from reducing to oxidizing conditions induced by H_2O_2 (see 2.6.1.3). Purified CG6 (8 μM) in Buffer B was measured in a 1 cm pathlength cuvette with a Hewlett Packard diode array spectrophotometer (Agilent). UV-Vis absorption spectra was measured of apo CG6, 10 min after the addition of a slight excess of haem (10 μM) and 2 hrs, 4 hrs and 20 hrs after the addition of 0.02% (w/v) H_2O_2.

2.6.1.5 Absorbance extinction coefficient determination

Absorbance extinction coefficients were calculated by two methods with averages of all the values taken (minimum of three measurements). UV-Vis absorbance spectra were recorded with either a Hewlett Packard diode array spectrophotometer (Agilent) or a Jasco V-660 spectrophotometer (Jasco), both with a 1 cm path length quartz cuvette. Protein samples were diluted into Buffer B to a final concentration of 5 μM and spectra measured between 200-800 nm. The absorbance value at λ_{max} was used in Equation 1 to give an extinction coefficient with $M^{-1}cm^{-1}$ units, where ε is extinction coefficient, A is the absorbance value, c is the concentration and l is the pathlength. Alternatively protein samples were diluted in Buffer B to an A_{488} of 0.5 and the sample quantified by DC protein assay (see 2.5.7). The concentration value was then substituted into Equation 1 to give extinction coefficients with $M^{-1}cm^{-1}$.

$$\varepsilon = \frac{A}{c \times l} \qquad \text{Equation 1}$$

2.6.1.6 Holo cyt b_{562}-EGFP chimera λ_{max} determination

Apo cyt b_{562}-EGFP chimeras were diluted in Buffer B to a final concentration of 5 μM with an equilmolar concentration of haem under either reducing conditions (1 mM ascorbic acid) or oxidizing conditions (1 mM KNO_3) and incubated for 30 mins at 4 °C. UV-Vis absorbance spectra were measured in a 1 cm pathlength cuvette with a Jasco V-660 spectrophotometer.

2.6.1.7 Quantum yield determination

Quantum yields for EGFP and apo-cyt b_{562}-EGFP chimeras were calculated using fluorescein as a reference. Fluorescein (in 0.1 M NaOH) and the chimeras (in Buffer A) were prepared to an A_{488} of 0.05 in a 1 cm pathlength cuvette. Emission spectra were measured after excitation at 488 nm using a 5 x 5 mm dimension cuvette with excitation and emission band passes of 2.5 nm at 30 nm/min scan rate. Integrated emission intensity between 500 and 650 nm was calculated and used in Equation 2 to generate quantum yield values

$$\phi_x = \phi_{st} . \frac{Area_x}{Area_{st}} . \frac{\eta_x^2}{\eta_{st}^2} \qquad \text{Equation 2}$$

where ϕ_x and ϕ_{st} refer to the fluorescence quantum yield of the sample and fluorescein standard, respectively. $Area_x$ and $Area_{st}$ are the integrated emission intensities for the sample and fluorescein standard, respectively. η_x and η_{st} are the refractive index of the solvent for the sample and fluorescein standard, respectively. The refractive index correction here was negligible as 0.1 M NaOH and aqueous buffers differ in refractive index by <1%.

2.6.1.8 Fluorescence lifetime determination

Fluorescence lifetimes were determined by a time correlated single photon counting (TCSPC) technique using an FLS-920 fluorometer (Edinbugh Instruments) with a hydrogen nF900 nanosecond flashlamp excitation source, under 0.4 bar pressure, and a PMT-Hamamatsu R2658P detector. Samples were measured across 512 channels (0.039 ns/channel) until one channel reached a maximum photon count

of 500 after excitation at 488 nm and emission monitored at 511 nm. An instrument response function (IRF) was measured with every sample measurement for reconvolution of the data for exponential curve fitting. The IRF is a lifetime recorded for the instrument with no sample present. Protein samples were diluted into Buffer B to a final concentration of 6 μM in a 1 cm pathlength cuvette for measurement. The fluorescence lifetime data were fit to a single exponential decay (Equation 3) after IRF reconvolution of the data using Edinburgh instruments F900 software.

$$N(t) = A + N_0 . e^{(-t/\tau)} \qquad \text{Equation 3}$$

Where N(t) is the photon counts at time t, A is the value (photon counts) at which the decay plateaus, N_0 is the photon counts at t = 0, t is the time (ns) and τ is the fluorescence lifetime.

2.6.1.9 CD spectroscopy

CD spectroscopy measurements were performed using a Chirascan CD spectrometer (Applied Photophysics), measuring CD spectra between 190 and 250 nm at a scan rate of 1nm/sec in a quartz cuvette with a 1 mm pathlength. Protein samples were buffer exchanged (see 2.5.7) into 10 mM sodium phosphate buffer at pH 8.0 and diluted to a final concentration of 5 μM. Fresh ascorbic acid was titrated into the sample to a final concentration of 1 mM and CD spectra measured, with an equivalent buffer containing no protein sample measured for baseline subtraction. Haem was then titrated into the sample to a final concentration of 5 μM and CD spectra measured.

2.6.1.10 Redox midpoint determination for CG15 double cysteine mutants

Protein samples (1 uM) were diluted into degassed Buffer B supplemented with 1 mM EDTA and 1 mM DTT redox buffers. Concentrations of oxidised and reduced DTT were reciprocally varied from 0:1 to 1:0 mM over ~27 concentration increments. The samples (200 μl) were incubated in 96 well plates under each redox condition for 4 hrs at 30 °C in an airtight container under anaerobic conditions, produced by Anaerocult A (Merck), to reach equilibrium.

Excitation spectra were measured using the 96 well plate attachment for the Cary Eclipse fluorescence spectrophotometer (Varian) monitoring fluorescence at 530 nm with excitation and emission band passes of 20 nm. Given that the excitation spectra for CG15 double cysteine mutants ($CG15^{CC}$) is strongly and reversibly dependent on the redox state and that two excitation maxima varied with a well defined isosbestic point, the reactions between $CG15^{CC}$ variants and reduced or oxidized DTT could be analysed using a two state model. The equilibrium for the oxidation of reduced $CG15^{CC}$ variants and the corresponding K_{eq} are given by reaction 1 and Equation 4.

$$CG15^{CC}_{red} + DTT_{ox} \overset{K_{eq}}{\longleftrightarrow} CG15^{CC}_{ox} + DTT_{red} \qquad \text{Reaction 1}$$

$$K_{eq} = \frac{[CG15^{CC}_{ox}].[DTT_{red}]}{[CG15^{CC}_{red}].[DTT_{ox}]} \qquad \text{Equation 4}$$

After equilibration between DTT redox buffers and a $CG15^{CC}$ variant the fractional amount of $CG15^{CC}$ reduced (R) could be calculated using Equation 5.

$$\mathrm{R} = \frac{(\mathrm{F}-\mathrm{F_{ox}})}{(\mathrm{F_{red}}-\mathrm{F_{ox}})} \qquad \text{Equation 5}$$

Where F is the ratio between excitation intensity, at 488 nm and the isosbestic point, at a given DTT redox buffer ratio, F_{ox} is the fluorescence ratio in 1 mM oxidized DTT and F_{red} is the fluorescence ratio in 1 mM reduced DTT. Using Equations 5 and 6 the actual concentrations of DTT_{red} and DTT_{ox} , after equilibration with a $CG15^{CC}$ variant, can be determined. Where $[DTT_{red\ 0}]$ and $[DTT_{ox\ 0}]$ are the starting concentrations and R is the fraction of reduced $CG15^{CC}$ variant calculated from Equation 5

$$[DTT_{red}] = [DTT_{red\ 0}] - R.[CG15^{CC}] \qquad \text{Equation 6}$$

$$[DTT_{ox}] = [DTT_{ox\ 0}] + R.[CG15^{CC}] \qquad \text{Equation 7}$$

By plotting the R values against the log [DTTred]/[DTTox] ratios and fitting the data to a titration curve determined by Equation 8, K_{eq} for the CG15^{CC} variant with DTT could be found.

$$R = \frac{[DTT_{red}]/[DTT_{ox}]}{Keq + [DTT_{red}]/[DTT_{ox}]} \qquad \text{Equation 8}$$

The K_{eq} value could then be used in the Nernst equation (Equation 9) to determine the redox midpoint potential for the CG15^{CC} variant.

$$E'_{0(CG15cc)} = E'_{0(DTT)} - \frac{RT}{nF} . lnK_{eq} \qquad \text{Equation 9}$$

Where $E'_{0(DTT)}$ is the standard reduction potential for the DTT_{red}/DTT_{ox} couple (-0.323 V at pH 7.0 and 30 °C), R is the gas constant (8.315 J K^{-1} mol^{-1}), T is the absolute temperature (303.15 K), n is the number of transferred electrons (2) and F is the Faraday constant (9.649×10^4 C mol^{-1}).

2.6.1.11 Redox Kinetics

In vitro rates for the reaction between CG15^{CC} variants with either DTT or H_2O_2 were determined by the change in fluorescence excitation ratio over time after the addition of excess oxidizing or reducing agent. Excitation maxima (400/495 nm) intensities were measured over time while monitoring fluorescence emission at 530 nm, with a 5 nm excitation band pass and 10 nm emission band pass, in a 5 x 5 mm dimension cuvette, using a Cary Eclipse fluorescence spectrophotometer (Varian).

Protein samples (10 μM) were incubated in buffer B supplemented with 1 mM EDTA and either 1 mM DTT or H_2O_2 for 1 hr at 30 °C to fully reduce or oxidize the sample. For reduction kinetics the protein sample incubated in H_2O_2 was diluted to a final concentration of 0.5 μM in buffer B with 1 mM EDTA. Fluorescence excitation intensities were measured for 20 min after the addition of 1 mM DTT.

For oxidation kinetics the protein sample incubated with DTT was diluted to a final concentration of 0.5 μM in Buffer B with 1 mM EDTA. Fluorescence excitation intensities were measured for 20 mins after the addition of 1 mM H_2O_2. The fraction of reduced CG15^{CC} variants (R) was calculated from the excitation maxima ratios (see

Equation 5) and plot against time, with the pseudo first order rate constants calculated using a curve fit to Equation 10.

$$R = A.(e^{-k.t}) + B \qquad \text{Equation 10}$$

2.6.1.12 pH sensitivity

To assess the effect of pH on $CG15^{CC}$ variant excitation peak ratio, samples were incubated in 50 mM phosphate buffers ranging from pH 4 to pH 9 for 24 hrs at 25 °C and excitation spectra measured. Protein samples were diluted to a final concentration of 0.5 μM, in a 96 well plate (200 ul total volume), in a redox buffer supplemented with 1 mM EDTA and either 1 mM DTT or 1 mM H_2O_2, to fully reduce or oxidize the samples respectively. Fluorescent excitation spectra were measured while monitoring fluorescence emission at 530 nm with 20 nm excitation and emission band passes using a 96 well plate attachment for the Cary eclipse fluorescence spectrophotometer (Varian). The excitation peak ratio of the samples under reducing and oxidizing conditions were plot against pH to assess the effect of pH.

2.6.1.13 Calculating haem binding affinity

The haem binding affinity under oxidising conditions was calculated from the Haem-mediated fluorescence quenching data. The value in the absence of haem (initial data point) was equal to 0 haem bound (apo-protein); the data point at the highest saturating haem concentration where no further decrease in fluorescence was observed was scaled to 1 (all binding sites occupied). The scaled data was plotted against haem concentration and least squares non-linear regression was used to fit the data to the following binding Equation:

$$Holo_n = (Holo_{max}[Haem])/(K_d + [Haem])$$

where $Holo_n$ is the proportion of holo-protein derived from fluorescence quenching measurement (0-1), $Holo_{max}$ holo-protein fully occupied with haem.

2.6.2 Methods for biophysical characterization

2.6.2.1 Size exclusion chromatography (SEC)

To determine if EGFP, single amino acid deletion variants of EGFP, cyt b_{562}-EGFP chimeras or CG15^{CC} variants form quaternary structures SEC analysis was performed. Biorad gel filtration standards (Biorad, Table 2.4) were applied to a Superdex^{TM} 75 or Superdex^{TM} 200 column (20 ml bed volume, flow rate 0.5 ml/min), as per the manufacturers guidelines, with protein elution monitored at 280 nm. Using Equation 11, K_{av} can be determined for each standard and plotted against LogMw to produce a standard curve, where V_e is the elution volume, V_t is the total volume and V_o is the void volume. The void volume was determined from the elution volume of protein aggregates (Table 2.4) and the total volume was determined from the elution volume of Vitamin B_{12}.

Protein samples were diluted in Buffer B to final concentrations of 25, 50 or 100 μM (in 200 ul total volume) and applied to a Superdex 75 (EGFP and single amino acid deletion mutants of EGFP) or Superdex 200 column (EGFP or cyt b_{562}-EGFP chimeras) (20 ml bed volume, flow rate 0.5 ml/min). Elution volumes (V_e) were determined for each sample by the peak absorbance at 488 nm and K_{av} values calculated. Using the standard curve, estimated molecular weights could be determined for each protein sample.

$$K_{av} = \frac{V_e - V_o}{V_t - V_o} \qquad \text{Equation 11}$$

Table 2.4. Gel filtration standards and molecular weights

	Protein	Molecular weight (Mw, Da)
Void volume (V_o)	Protein aggregates	NA
Elution volume (V_e)	Thyroglobulin (Bovine)	670, 000
Elution volume (V_e)	γ-Globulin (Bovine)	158, 000
Elution volume (V_e)	Ovalbumin (Chicken)	44, 000
Elution volume (V_e)	Myoglobin (Horse)	17, 000
Total volume (V_t)	Vitamin B_{12}	1,350

2.6.2.2 TND guanidine hydrochloride equilibrium unfolding

Wild type EGFP and single amino acid deletion variants of EGFP were diluted to a final concentration of 1 μM in 50 mM Tris-HCl, pH 8.0 at 25 °C supplemented with 150 mM NaCl, 10% (v/v) glycerol (TNG Buffer) and guanidine hydrochloride (0-6 M) (GdmCl), across 32 concentration increments. All samples were set up in

triplicate. Protein samples (200 µl) were incubated in 96 well plates with the required guanidine hydrochloride concentration for up to 250 hrs at 25 °C. Protein unfolding was monitored by fluorescence at 520 nm after excitation at 480 nm using a FLUOstar Omega 96 well plate reader (BMG Labtech).

To estimate the apparent [GdmCl] at which 50% of the protein is folded and 50% of the protein is unfolded (transition midpoint) from data measured at different incubation time points a five parameter asymmetric curve fitting function was used in the GraphPad Prism software (Equation 12). A two-state model (see below) could not be used to fit the data at the first three incubation time points for all protein samples measured due to a lack of data points in the transition region.

$$F = F_D + \frac{(F_N - F_D)}{(1+10^{\left(\left(\left([D]_{50\%} + \left(\frac{1}{Z}\right)\log(2) - 1\right) - [D]\right) \times Z\right)})} \qquad \text{Equation 12}$$

Where F is the fraction of folded protein, F_D is the fraction of protein in the denatured state, F_N is the fraction of protein in the native state, $[D]_{50\%}$ is the midpoint of the unfolding transition and represents the concentration of denaturant at which 50% of the protein is folded and 50% is unfolded, Z is a unitless slope factor or Hill slope.

Equilibrium unfolding data measured after 250 hr incubation were fit to a two state model equation as follows although this was not used to generate $\Delta G_{D-N}^{H_2O}$ values.

$$F = (F_N + A_N[D]) - \left((F_N + A_N[D]) - (F_D + B_D[D])\right) \frac{\exp^{\frac{m([D]-[D]_{50\%})}{RT}}}{1 + exp^{\frac{m([D]-[D]_{50\%})}{RT}}}$$

Where F is the fraction of folded protein, F_N and F_D are the fraction of protein folded in the native (1) and denatured state (0) respectively, A_N and B_D are the slope of the native and denatured baselines respectively, m is a constant related to the dependence of ΔG on GdmCl concentration and [D] is the concentration of denaturant. $[D]_{50\%}$ is the midpoint of the unfolding transition and represents the concentration of denaturant at which 50% of the protein is folded and 50% is unfolded.

The linear extrapolation method (LEM) was used to calculate free energies of unfolding from the equilibrium unfolding data. Using Equation 13 the fraction of

denatured protein (F_D) could be calculated from the fluorescence intensities at different GdmCl concentrations.

$$F_D = \left(\frac{Y - Y_N}{Y_D - Y_N}\right)$$ Equation 13

Where Y is the fluorescence intensity at a given GdmCl concentration, Y_N is the fluorescence intensity of native protein (0 M GdmCl) and Y_D is the fluorescence intensity of denatured protein (6 M GdmCl). By substituting Equation 12 into the expression for the equilibrium constant of reaction 2, K_{eq} values can be calculated at each GdmCl concentration with Equation 14:

$$[N] \underset{\rightharpoonup}{\overset{K_{eq}}{\leftharpoondown}} [D]$$ Reaction 2

$$K_{eq} = \frac{[D]_{eq}}{[N]_{eq}} = \frac{F_D}{1 - F_D}$$ Equation 14

The K_{eq} values can then be used to calculate ΔG^0_{D-N} values, at each GdmCl concentration, using Equation 15.

$$\Delta G^0_{D-N} = -RTln(K_{eq})$$ Equation 15

Where R is the gas constant (1.9858775 $cal.K^{-1}.mol^{-1}$) and T is the absolute temperature (310 K). As ΔG^0_{D-N} values show a linear dependence on GdmCl concentrations the free energy of unfolding for a given protein ($\Delta G^{H_2O}_{D-N}$) can be calculated by plotting ΔG^0_{D-N} values against GdmCl concentration and fitting the data to Equation 16.

$$\Delta G^0_{D-N} = \Delta G^{H_2O}_{D-N} - m_{D-N}.[GdmCl]$$ Equation 16

Where $\Delta G^{H_2O}_{D-N}$ is the free energy of denaturation at 0 M GdmCl and m_{D-N} is the proportionality constant, with units of $cal.mol^{-1}$ or $cal.mol^{-1}.M^{-1}$ respectively.

2.6.2.3 Cyt b_{562}-EGFP chimera guanidine hydrochloride equilibrium unfolding

Wild-type EGFP and the cyt b_{562}-EGFP chimeras, CG1 and CG6, were diluted to a final concentration of 4 μM in TNG Buffer and GdmCl (0-6 M), across 32 concentration increments in triplicate. Two equilibrium-unfolding experiments were performed for each protein sample under apo and holo conditions (equimolar concentration of haem) to see if haem bound to the chimeric proteins had an effect on stability.

Haem was also added to the EGFP sample as a control to make sure that if any changes in stability were seen it was due to haem bound to the chimeras and not due to free haem in solution. Protein unfolding was monitored by absorbance at 488 nm for the EGFP domains and at 420 nm for the holo-cytochrome domains. Absorbance was measured instead of fluorescence, as haem quenches fluorescence when bound to the chimeric proteins, using a FLUOstar Omega 96 well plate reader (BMG Labtech). Data analysis and $\Delta G_{D-N}^{H_2O}$ values were calculated as described above (see 2.6.2.2).

2.6.2.4 Protein unfolding and refolding kinetics

Protein unfolding kinetics for EGFP and single amino acid deletion variants of EGFP were determined by rapid dilution, to a final concentration of 1 μM, into 6.4 M GdmCl in TNG Buffer and 1 mM DTT. Fluorescence was monitored for 20 min at 510 nm after excitation at 488 nm in a 5 x 5 mm dimension cuvette with an excitation band pass of 2.5 nm and emission band pass of 5 nm using a Cary Eclipse fluorescence spectrophotometer (Varian). Protein samples were also rapidly diluted into TNG Buffer with no GdmCl and fluorescence monitored for data normalization and to take into account any photobleaching effects. To determine refolding kinetics, samples from the unfolding reaction were rapidly diluted (1/10) into TNG Buffer containing no GdmCl so that the final protein concentration was 100 nM and GdmCl was 0.64 M. Fluorescence was monitored for 20 min as before but with excitation and emission band passes of 5 nm. Unfolding experiments were carried out in duplicate except for the single amino acid deletion variant A227Δ, which was only carried out once, with data fit to Equation 17. All refolding experiments were carried out in triplicate and data fit to Equation 18.

$$Y = (Y_0 - P) \times e^{-kt} + P \qquad \text{Equation 17}$$

$$Y = Y_0 + \left(F_1 \times (1 - e^{-K_1 t})\right) + (F_2 \times (1 - e^{-K_2 t})) \qquad \text{Equation 18}$$

Where Y_0 is the Y value when t = 0, P is the Y value at infinite time, F_1 is a proportional value for the first rate constant, k_1, and F_2 is the proportional value for the second rate constant and k_2.

2.6.2.5 Thermal denaturation

Melting temperatures (T_m) of EGFP, single amino acid deletion variants of EGFP and cyt b_{562}-EGFP chimeras were determined by monitoring fluorescence with a Opticon 2 qPCR thermal cycler (MJ Research) while ramping the temperature from 25-98 °C. Protein samples were diluted to a final concentration of 1 μM in 50 mM sodium phosphate buffer pH 8.0 (total volume 50 μl) and the temperature ramped at 1°C/min. MJ Research Software supplied with the qPCR machine determined the melting temperature of the samples based on the collected fluorescence data.

2.7 X-ray crystallography and structure determination

2.7.1 Protein Crystalisation

The EGFP, EGFPΔ^{D190}, EGFPΔ^{A227} (10 mg/ml in 50 mM Tris-HCl, pH 8.0 and 150 mM NaCl) and CG6 protein samples (10 mg/ml in 50 mM Tris-HCl, pH 8.0, 150 mM NaCl and 256 μM haem) were screened for crystal formation by the sitting drop vapour diffusion method with incubation at 4 °C. Drops were set up with equal volumes of protein and precipitant solutions (0.5 μl each). The crystal of EGFP was obtained from 0.1 M MES/NaOH, pH 6.5, 200 mM calcium acetate and 20% (w/v) PEG 8000. A crystal was transferred to mother liquor supplemented with 13% (w/v) PEG 200 as a cryoprotectant and vitrified. The crystal of EGFPΔ^{D190} was obtained from 0.1 M Hepes/NaOH, pH 6.6, 200 mM ammonium sulphate and 2 M (K/Na)-phosphate. A crystal was transferred to mother liquor supplemented with 16% (v/v) glycerol as a cryoprotectant and vitrified. The crystal of EGFPΔ^{A227} was obtained from 0.05 M MES/NaOH, pH 5.6, 100 mM ammonium sulphate, 10 mM magnesium chloride hexahydrate and 20% (w/v) PEG 8000. A crystal was transferred to mother liquor supplemented with 13% (w/v) PEG 200 as a cryoprotectant and vitrified. The crystal of holo-CG6 was obtained from 0.1 M MES/NaOH, pH 6.4, 200 mM

Magnesium acetate and 20% (w/v) PEG 8000. A crystal was transferred to mother liquor supplemented with 16% (v/v) glycerol as a cryoprotectant and vitrified. Data were collected on the Diamond Light Source beamline I02.

2.7.2 Structure determination

Initial molecular replacement for the EGFP structure was performed using a previously determined GFP structure (PDB:2HQZ) as the search model, using the CCP4 program MOLREP [87]. The structure for EGFP was adjusted manually using COOT [88] and refinement of the completed molecule was carried out using the REFMAC program [89].

Structural determination of the CG6 cyt b_{562}-EGFP integral fusion domain scaffold was performed by Dr Matthias Bochtler and Dr Honorata Czapinska as follows. Holo-CG6 fusion protein crystallized in a nearly orthorhombic lattice. Reprocessing the data for a monoclinic lattice indicated cell constants of 64.76 Å x 125.22 Å x 89.26 Å and β=90.37. The extinctions on the 0l0 reciprocal space axis further predicted that the correct space group was P1 2(1) 1. Assuming that the insertion of the cytochrome domain into GFP would leave the β-barrel structure of GFP intact, molecular replacement was attempted with a single GFP molecule (pdb-accession code 2HQZ) as the search model, using the CCP4 program MOLREP. Then the cytochrome models were placed using MOLREP with fixed GFP molecules already in place. In order to pair correct GFP and cytochrome fragments, crystallographic symmetry mates were displayed. Pairings were supported by connecting density, as well as similar relative domain orientations. The structure with two fusion proteins was adjusted manually using COOT. The presence of a third molecule was suggested by a gap in the packing arrangement, and by the strong clustering of residual density, which was of approximately the right size and shape. Inspection indicated that the density for the GFP domain was poorer than for the cytochrome domain. Therefore, the cytochrome domain alone was used as a search model using the EPMR [90] and PHASER [91] programs. The resulting placement was identical and confirmed further by the anomalous signal for iron, with a peak at the expected site in the anomalous difference Fourier map. The GFP domain of the third molecule was located using a difference Fourier map calculated with phases for the already placed models. Connection of the GFP and cytochrome domains revealed that their relative orientation was slightly different from the other two conformations,

and that the GFP domain was less ordered than the cytochrome domain, explaining the earlier observations. Refinement of the completed molecule was carried out using the REFMAC program, with an adjusted library for the haem to enforce expected geometry and particularly planarity. Refinement was done with separate TLS parameters for the six domains in the asymmetric unit, with separate (standard weight) NCS restraints for GFP and cytochrome domains.

2.7.3 Small angle X-ray scattering

Small angle X-ray scattering data were collected at beamline X33 at the DORIS ring, DESY, Hamburg. Data were processed and background corrected in the usual manner. Small angle X-ray scattering data were recorded at various protein concentrations. As the dependence of relative intensities on the scattering vector was independent of protein concentration, all further analysis was done with the scattering data for the highest concentration. Scattering for given atomic coordinates were calculated and fitted to the experimental data using the CRYSOL program [92]. *Ab Initio* shape determination was done with the DAMMIF program [93]. All SAXS data analysis was performed by Dr Michal Gajda at EMBL.

Chapter 3: GFP library generation: random trinucleotide deletion and domain insertion

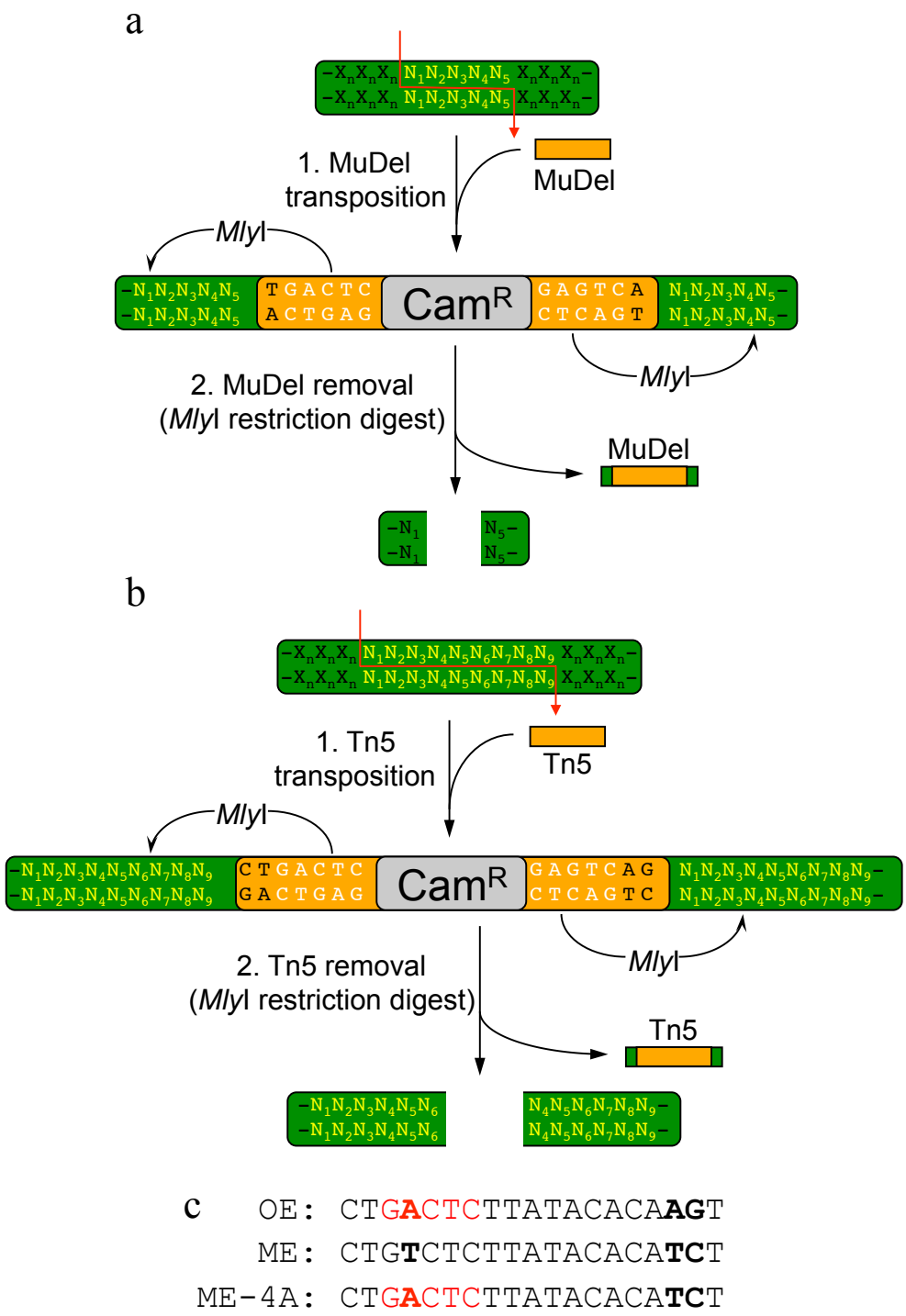
3.1 Introduction

With an ever-expanding protein engineering toolbox it is possible to construct and tailor proteins with desired or improved properties. The two main branches are rational site directed mutagenesis and directed evolution [94]. The limitation being rational mutagenesis approaches often require knowledge of the structure and function of the protein in order to aid the engineering process [94]. Recent development of *in silico* design approaches should greatly aid *de novo* rational protein design [95] but is still limited and requires "tuning" through random mutagenesis approaches.

The advent of directed evolution approaches in conjunction with improved screening methods allows for libraries of variants with mutations scattered throughout the target protein to be produced and those with desired characteristics identified for further characterisation [5, 38, 39]. Here we describe a transposon based directed evolution approach which produces a library of random breaks throughout a target gene [19, 21, 31, 37, 49, 50] which can in turn be used to sample different mutagenesis events not normally sampled by traditional approaches: triplet nucleotide deletion (TND) [19, 21], domain insertion [31, 37] and triplet nucleotide exchange (TriNEx) [49] for unnatural amino acid incorporation [50].

This method utilizes the bacteriophage transposon, Mu, and transposase, MuA, which has very low target site preference and can efficiently insert Mu transposons randomly throughout target genes [40, 43]. An engineered Mu transposon, termed MuDel, has been generated for the purposes of sampling various downstream mutagenesis events [21]. MuDel contains *Mly*I restriction sites 1 bp from its 5' and 3' termini for quick and easy removal of the transposon from the target DNA after transposition (Fig 3.1 a). Transposition with MuDel results in the duplication of 5 bp of the target gene around the insertion site due to the staggered nature with which the transposase cleaves the target DNA substrate (Fig 3.1 a) [43].

*Mly*I restriction digestion cuts 5 bp outside of its recognition sequence resulting in 3 bp being deleted from the target gene (Fig 3.1 a). The resulting sub-library has random breaks throughout the target gene and can be recircularised to produce a TND library, have DNA cassettes encoding whole protein domains



c

OE:	CTGACTCTTATACACA**AG**T
ME:	CTG**T**CTCTTATACACA**TC**T
ME-4A:	CTGACTCTTATACACA**TC**T

Fig 3.1. Schematic mechanism of transposon insertion and removal by MlyI restriction digestion. a, The MuDel system. ***Step 1***, Transposition with MuDel (orange), containing a chloramphenicol resistance gene (Cam^R), by MuA transposase generates a 5 bp duplication of the target DNA (green) due to the staggered nature of target gene cleavage (red arrow). ***Step 2,*** MlyI restriction digestion cuts 5 bp outside its recognition sequence (white) excising MuDel and deleting a triplet nucleotide from the target gene. **b,** Tn5 system. ***Step 1,*** Transposition with a Tn5 transposon (orange), containing Cam^R, by the Tn5 transposase generates a 9 bp duplication of the target DNA (green). ***Step 2,*** MlyI restriction digestion excises the Tn5 transposon, leaving a triplet nucleotide duplication of the target gene. **c,** Tn5 Outward End (OE) and Mosaic End (ME) transposase recognition sequences differ at position 3, 17 and 18 indicated in bold. An MlyI recognition sequence (red) is generated in ME-4A by mutating the thiamine at position 4 in the ME recognition sequence.

ligated into the random breaks or can have a triplet nucleotide replaced (TriNEx) (Fig 1.8, Section 1.3.3.3)

Another transposon to be investigated is the bacterial Tn5 transposon from the insertion sequence 4 (IS4) family [96]. This transposon generates a 9 bp duplication upon insertion into a target gene [97] (Fig 3.1b). There are two different Tn5 transposase recognition sequences (TREs) that can be used, the wild type outward end (OE) recognition sequence [96] or the hyperactive mosaic end (ME) sequence (Fig 3.1c) [97]. While the ME recognition site confers ~8 fold higher transposition efficiency [98], the OE recognition sequence already has a critical *Mly*I recognition sequence 2 bp from its 5' and 3' termini (Fig 3.1 c). The positioning of this *Mly*I site allows for a triplet nucleotide duplication to be introduced into a target gene as outlined in Figure 3.1b. This transposon method could therefore be used to potentially sample single amino acid duplications. Additionally a DNA cassette can be inserted to generate an alternate route to sample domain insertion variants.

Mutation of a single nucleotide in the ME recognition sequence of Tn5 (ME-4A Tn5) introduces the critical *Mly*I recognition sequence 2 bp from its 5' and 3' termini (Fig 3.1c). This allows for the two different Tn5 recognition sequences to be tested for transposition efficiency and their ability to produce diverse insertion libraries within a target gene.

The target gene used in this study was *egfp,* encoding for enhanced green fluorescent protein (EGFP) [99]. Fluorescent proteins are used extensively across many different research disciplines with the engineering and development of improved or altered fluorescent variants always in high demand. EGFP is a desirable target protein for engineering as it is genetically encoded, requires no cofactors for fluorescence to mature and the fluorophore is sensitive to changes in its local environment [62]. The effects of mutational studies can therefore be easily screened for by monitoring changes in fluorescent characteristics.

In this Chapter, the process of library construction will be described. The MuDel and Tn5 transposition efficiencies and insertion position diversity within the *egfp* target gene will be examined and compared. The sub-library of transposon inserted randomly within *egfp* will then form the basis for generating the main libraries of EGFP sampling single amino acid deletions, codon replacements and domain insertions.

3.2 Results

3.2.1 Construction of pNOM-XP3-egfp

A major step in the success of the transposon based directed evolution approach requires the removal of the transposon from a plasmid library by *Mly*I restriction digestion. It is therefore imperative that no other *Mly*I sites lie in the host plasmid or gene to prevent undesired side reactions. A plasmid previously described, pNOM [21], which has had all *Mly*I sites removed by site directed mutagenesis was converted from a simple cloning plasmid to an expression plasmid by inserting a modified T7 expression cassette (promoter, multiple cloning site (MCS) and terminator) sequence (from pET24c) in place of the pNOM MCS. The T7 promoter region was mutagenised to remove a *Mly*I site without loss of promoter function. This expression plasmid was constructed and supplied by Dr Wayne Edwards and was named pNOM-XP3 (Fig 3.2 a).

The *egfp* gene was successfully cloned between the *Nde*I and *Xho*I restriction sites of pNOM-XP3 (Section 2.2.15), as confirmed by PCR analysis of colonies containing the pNOM-XP3-*egfp* plasmid (data not shown). An expected PCR product of 1146 bp was generated indicating successful cloning. Retention of the *Nde*I and *Xho*I restriction sites in pNOM-XP3-*egfp* critical for later steps in library construction was also confirmed by restriction analysis (Section 2.2.6). The restriction analysis produced the two expected DNA bands corresponding to *egfp* (723 bp) and pNOM-XP3 vector (2143 bp) confirming that the restriction sites were retained (Fig 3.2 b).

3.2.2 Transposition efficiency of ME-4A-Tn5, OE-Tn5 and MuDel transposons into pNOM-XP3-*egfp*

The efficiency with which a transposase can incorporate a transposon into a target gene is of upmost importance in order to produce a library that samples a diverse number of insertion positions within a target gene. The transposition efficiency of the three transposons, ME-4A-Tn5, OE Tn5 and MuDel were investigated to assess the suitability of each transposon for further library production.

In vitro transposition with MuDel, ME-4A-Tn5 and OE-Tn5 transposons (Section 2.2.13, 2.2.14) into pNOM-XP3-*egfp* was performed in three separate reactions. Highly competent (2.08×10^8 cfu/µg DNA) *E.* coli NovaBlue GigaSingles^{TM} cells were transformed with the individual transposition reactions,

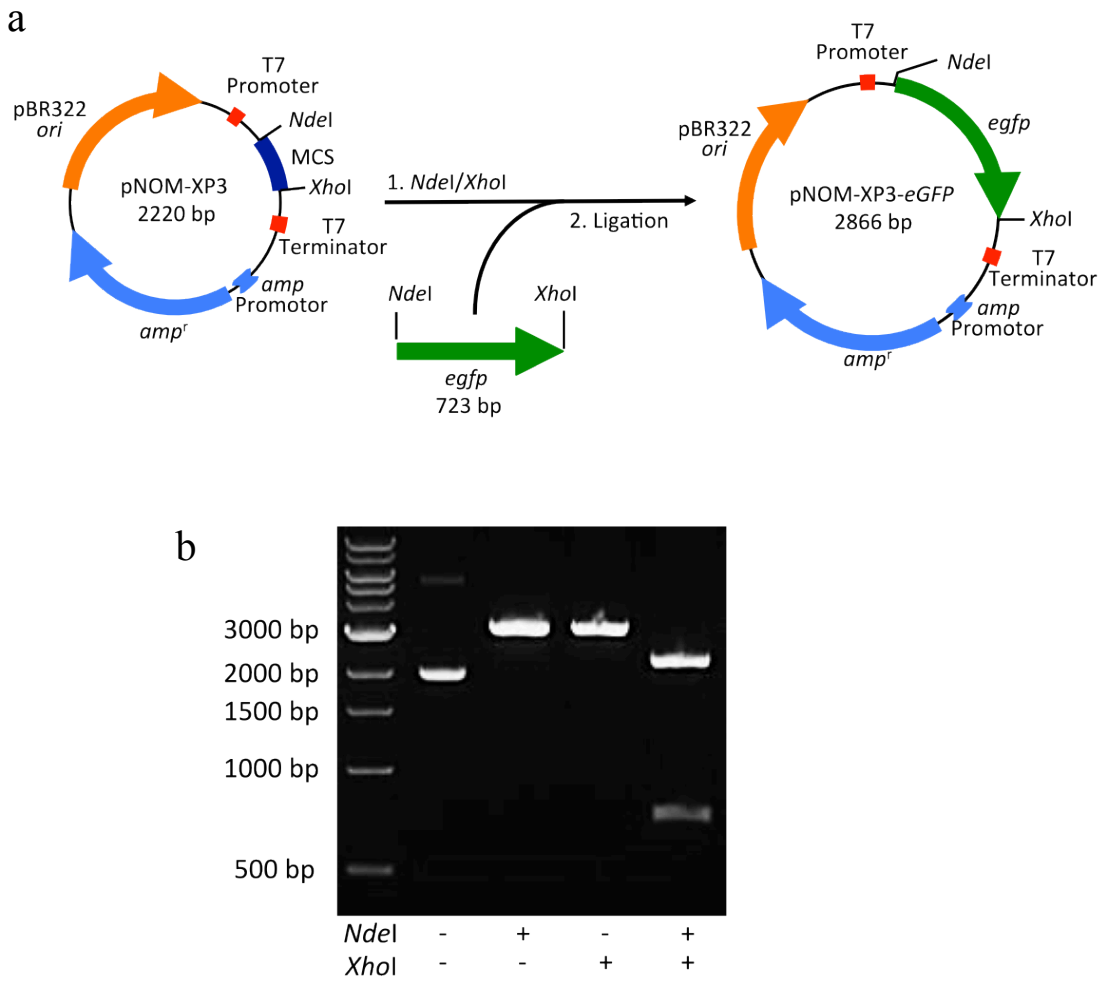


Fig 3.2. pNOM-XP3-*egfp* cloning strategy and restriction digest analysis a. The *egfp* gene (green) with *Nde*I and *Xho*I restriction sites introduced by PCR to the 5' and 3' end of the gene respectively was inserted into the pNOM-XP3 vector between the *Nde*I and *Xho*I sites within the multiple cloning site (MCS) resulting in the expression plasmid pNOM-XP3-*egfp*. **b.** Restriction digest of pNOM-XP3-*egfp* with either *Nde*I or *Xho*I gave single products of 2866 bp. Restriction with both *Nde*I and *Xho*I gave the two expected products of 2143 bp or 723 bp.

with positive clones selected for on LB Agar plates supplemented with 40 μg/ml chloramphenicol. The three libraries produced 135,882, 18,500 and 103,815 cfu/μg DNA respectively. MuDel and OE-Tn5 transposition were the most efficient but ME-4A-Tn5 transposition was up to 7 fold less efficient despite the wild type ME recognition sequence showing an 8-fold increase in transposition efficiency. All further work with the Tn5 transposon system was therefore performed with OE-Tn5 due to the inefficiency of the ME-4A-Tn5 system. The transposon insertion libraries are named by the target DNA used in transposition and the transposon used to generate that library; pNOM-XP3-*egfp*Δ^{MuDel} and pNOM-XP3-*egfp*$\Delta^{\text{OE-Tn5}}$.

3.2.3 Transposon library size and diversity

When producing transposon libraries within a target gene it is not only important to have enough insertion events, to theoretically cover every possible insertion position in the target gene, but also to have extensive diversity. It was therefore necessary to assess the number of different insertion events within the MuDel and OE-Tn5 libraries and the diversity of insertion positions throughout the target gene, pNOM-XP3-*egfp*.

3.2.3.1 Determining library size

The number of variants in the libraries was identified by plating a proportion of the *E. coli* cells transformed with a transposition reaction on LB Agar supplemented with 20 μg/ml of chloramphenicol, with the remaining cells used to inoculate LB broth supplemented with 20 μg/ml chloramphenicol. After overnight incubation at 37 °C the number of colonies on the LB Agar plates was used to infer the number of variants in the library isolated from the cell culture. The number of variants in the pNOM-XP3-*egfp*Δ^{MuDel} and pNOM-XP3-*egfp*$\Delta^{\text{OE-Tn5}}$ libraries is discussed in Section 3.2.3.3 and 3.2.3.4 respectively.

3.2.3.2 Determining library diversity

The diversity of transposon insertions in the target plasmid pNOM-XP3-*egfp* was assessed using a restriction digestion method outlined in Figure 3.3 a. Briefly; the diversity of transposon insertion positions within the libraries was assessed by restriction digest analysis using *Xho*I and *Mly*I (Section 2.2.6). Firstly the plasmid library containing transposon insertions throughout pNOM-XP3-*egfp* was linearised

by an *Xho*I restriction digest so that subsequent *Mly*I digestion, to liberate the transposon from the library, would produce fragments ranging in sizes from 0 bp to ~2860 bp (Fig 3.3 a). Analysis by agarose gel electrophoresis to visualize the fragmented libraries should produce an even DNA smear across the size range indicating extensive diversity. Increased band intensity in the DNA smear would therefore imply bias for particular transposon insertion sites within the target DNA. The pNOM-XP3-*egfp*Δ^{MuDel} and pNOM-XP3-*egfp*$\Delta^{\text{OE-Tn5}}$ library diversities are discussed in Sections 3.2.3.3 and 3.2.3.4 respectively.

3.2.3.3 MuDel library size and diversity

The *in vitro* transposition of MuDel into the target gene pNOM-XP3-*egfp* (Section 2.3.1) produced 110 colonies from 1.5 % of the transformed cells when plated on LB agar supplemented with 20 μg/ml chloramphenicol. From these figures, the remaining 98.5% of transformed cells, which encompasses the library as a whole, can be said to contain about 7,590 variants. This provides enough variants to sample a wide range of transposon insertions throughout pNOM-XP3-*egfp* as the number of variants is over twice that of different potential insertion sites within the 2866 bp plasmid.

The diversity of the pNOM-XP3-*egfp*Δ^{MuDel} library was assessed by *Xho*I and *Mly*I restriction digest and analysed by agarose gel electrophoresis (Fig 3.3 b). Restriction digestion of the pNOM-XP3-*egfp*Δ^{MuDel} library with *Xho*I alone produced a single fragment of ~4173 bp indicating that only a single transposition event had taken place per molecule of DNA (Fig 3.3 b) (double transposition events would produce a fragment of ~5480 bp).

Restriction digestion with *Mly*I produced two fragments corresponding to excised MuDel (1310 bp) and the linearized library DNA (2863 bp) (Fig 3.3 b). Restriction digestion with both *Xho*I and *Mly*I resulted in DNA fragments ranging from ~0 bp to ~2863 bp, indicative of transposon insertions throughout the plasmid, and a 1310 bp fragment corresponding to MuDel (Fig 3.3 b). There are however a much higher proportion of fragments in the range of ~2000 bp to ~2863 bp indicated by increased DNA band intensity in this region of the agarose gel.

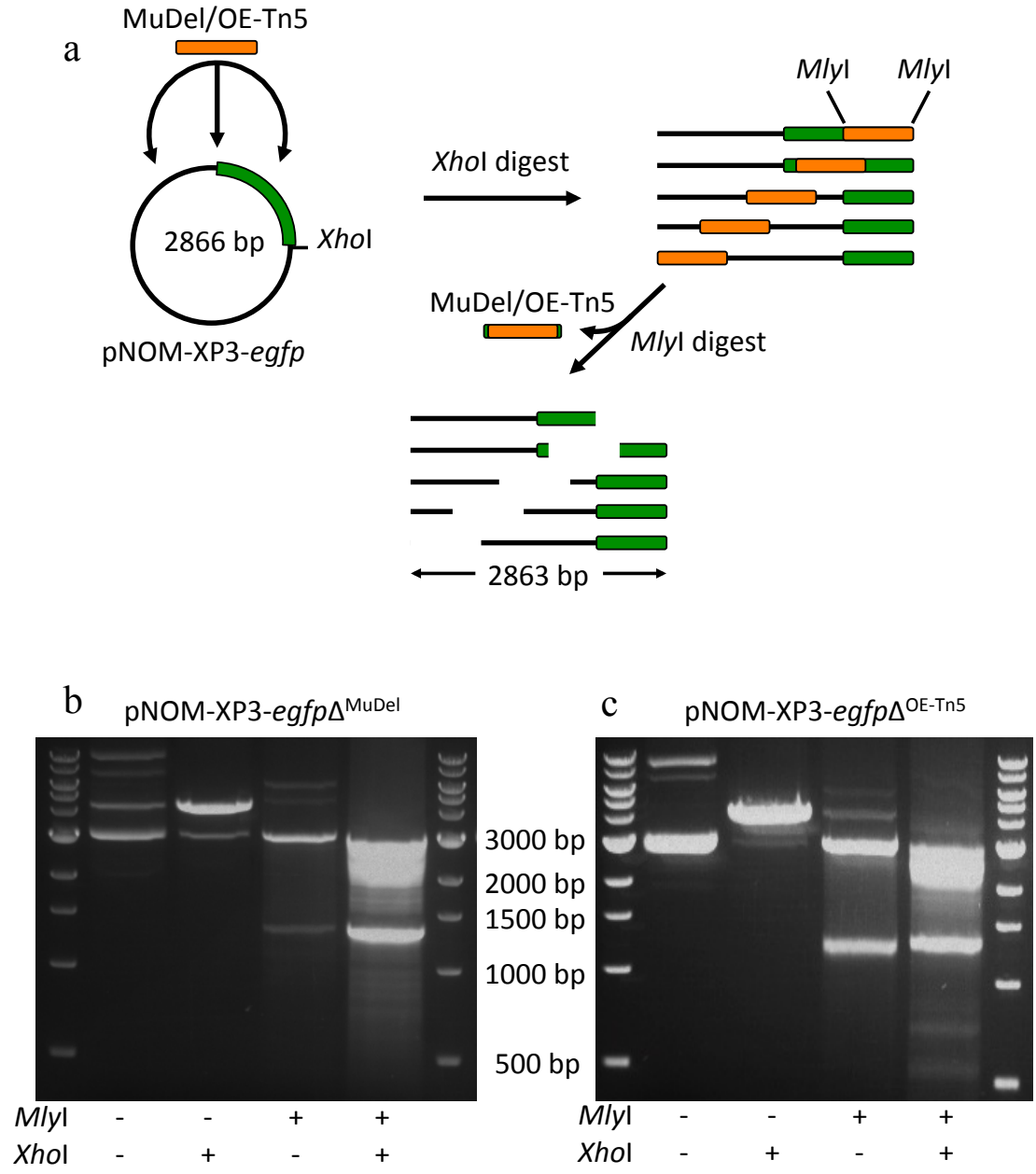


Fig 3.3. Transposon insertion diversity analysis. a. Schematic of restriction digest analysis. ***Step 1*****,** XhoI restriction digest linearizes the transposon plasmid libraries (vector: black, *e*gfp: green, transposon: orange). ***Step 2*****,** MlyI restriction digest liberates the transposon (1310 bp) leaving fragments ranging from 0 bp to ~2860 bp. Agarose gel electrophoresis analysis of **b,** pNOM-XP3-*egfp*Δ^{MuDel} and **c,** pNOM-XP3-*egfp*$\Delta^{\text{OE-Tn5}}$ library restriction digest with MlyI and/or XhoI as indicated in the figure.

The increased number of fragments at these sizes implies that there is a much higher rate of transposon insertion into the target gene, *egfp,* and the DNA immediately flanking it, probably due to non-tolerated transposon insertions into the ampicillin selectable marker or origin of replication.

A PCR screen of *E. coli* colonies (Section 2.2.3) from the MuDel library was performed using *egfp* specific primers AJB009 and AJB010 (Section 2.2.5) to assess the percentage of variants that had transposon insertions within the target gene, *egfp* (Fig 3.4). Of the 48 colonies screened 16 (33%) contained transposons within the target gene evident by the presence of a 2045 bp PCR product as opposed to a 738 bp fragment equivalent to the *egfp* gene alone (Fig 3.4). Given that the pNOM-XP3-*egfp*Δ^{MuDel} library had 7590 variants with the transposon inserted throughout the whole plasmid and that 33% contained a transposon within *egfp*, it can be said that the *egfp*Δ^{MuDel} library (Section 3.2.4) has 2504 total variants with transposon insertions within *egfp*.

This provides enough variants to sample a wide range of transposon insertions throughout *egfp* as the number of variants is over three times that of different potential insertion sites within the 720 bp gene.

3.2.3.4 OE-Tn5 library size and diversity

Transformation of *E.coli* NovaBlue GigaSingles^{TM} cells with the OE-Tn5 *in vitro* transposition into pNOM-XP3-*egfp* (Section 2.3.1) produced 347 colonies from 2.5% of the transformed cells plated on LB agar supplemented with 20 μg/ml chloramphenicol. From these figures, the remaining 97.5% of transformed cells, which encompasses the library as a whole, can be said to contain about 13,533 variants. Again, this library (pNOM-XP3-*egfp*$\Delta^{\text{OE-Tn5}}$) will provide enough variants to sample a wide range of transposon insertions within pNOM-XP3-*egfp*.

The diversity of the pNOM-XP3-*egfp*$\Delta^{\text{OE-Tn5}}$ library was also assessed by *Xho*I and *Mly*I restriction digest and analysed by agarose gel electrophoresis (Fig 3.3 c). Restriction digestion with *Xho*I alone produced a single fragment of ~4173 bp indicating that only a single transposition event had taken place per molecule of DNA (Fig 3.3 c). Restriction digestion with *Mly*I produced two fragments corresponding to excised OE-Tn5 (1308 bp) and the linearized library DNA (2869 bp) (Fig 3.3 c).

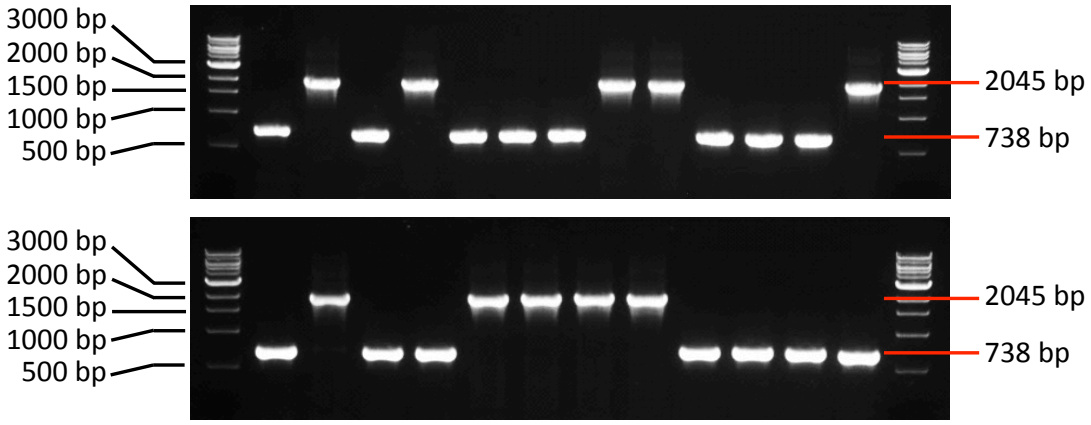


Fig 3.4. Analysis of bacterial colonies for the presence of MuDel within *egfp*. Bacterial colonies formed after transformation of E. coli with the pNOM-XP3-*egfp*Δ^{MuDel} library were screened by PCR (Section 2.2.3) with GoTaq polymerase and *egfp* specific primers to identify the proportion of MuDel insertions within the target gene *egfp*. Fragments of 2045 bp indicate a MuDel insertion has taken place within *egfp*, fragments of 738 bp indicate no insertion within *egfp*.

Restriction digestion with both *Xho*I and *Mly*I resulted in DNA fragments ranging from ~0 bp to ~2869 bp, indicative of transposon insertions throughout the plasmid, and a 1308 bp fragment corresponding to OE-Tn5 (Fig 3.3 c). As with the pNOM-XP3-*egfp*Δ^{MuDel} library the pNOM-XP3-*egfp*$\Delta^{OE\text{-}Tn5}$ library exhibited a higher proportion of transposon insertions within the target gene, *egfp*, and the flanking DNA indicated by an increase in band intensity of fragments in the size range of ~2000 bp and ~2869 bp (Fig 3.3 c). This again is probably due to non-tolerated transposon insertions into the ampicillin selectable marker or origin of replication.

The pNOM-XP3-*egfp*$\Delta^{OE\text{-}Tn5}$ library also displayed more intense bands on the agarose gel analysis of the *Mly*I and *Xho*I double digest, corresponding to fragments of ~500 bp and ~700 bp. The increased number of DNA fragments of these two sizes implies the OE-Tn5 transposon has higher target site specificity than MuDel.

Due to downstream library diversity analysis indicating that this library had potential bias in it (Section 3.2.5.2, Fig 3.6) and therefore would probably not be used, the number of variants in the library with transposon insertions within the target gene, *egfp*, was not determined.

3.2.4 Isolation and cloning of transposon containing *egfp*

The pNOM-XP3-*egfp*Δ^{MuDel} and pNOM-XP3-*egfp*$\Delta^{OE\text{-}Tn5}$ libraries contain transposon insertions throughout the plasmid but insertion sites only within *egfp* are desired. Using a restriction digest method outlined in Fig 3.5 a, all the variants with MuDel and OE-Tn5 insertions within *egfp* can be extracted (Section 2.3.2).

An *Xho*I and *Nde*I double digest was performed on the pNOM-XP3-*egfp*Δ^{MuDel} and pNOM-XP3-*egfp*$\Delta^{OE\text{-}Tn5}$ libraries to produce four distinct fragments; vector sequence with and without a transposon, and *egfp* with and without a transposon (Fig 3.5 a). The four fragments were separated by agarose gel electrophoresis (Fig 3.5 b) and the two fragments required, vector without a transposon and egfp containing a transposon were extracted (Section 2.2.1.2) and ligated together (Section 2.2.8). The ligation products were used to transform high efficiency competent cells (4 x 10^8 cfu/μg DNA); a percentage (1.6% or 5% for MuDel or OE-Tn5 respectively) were grown on LB agar plates supplemented with 20 μg/ml chloramphenicol and the remaining cells were used to inoculate LB broth supplemented with 20 μg/ml chloramphenicol, so generating libraries with transposon

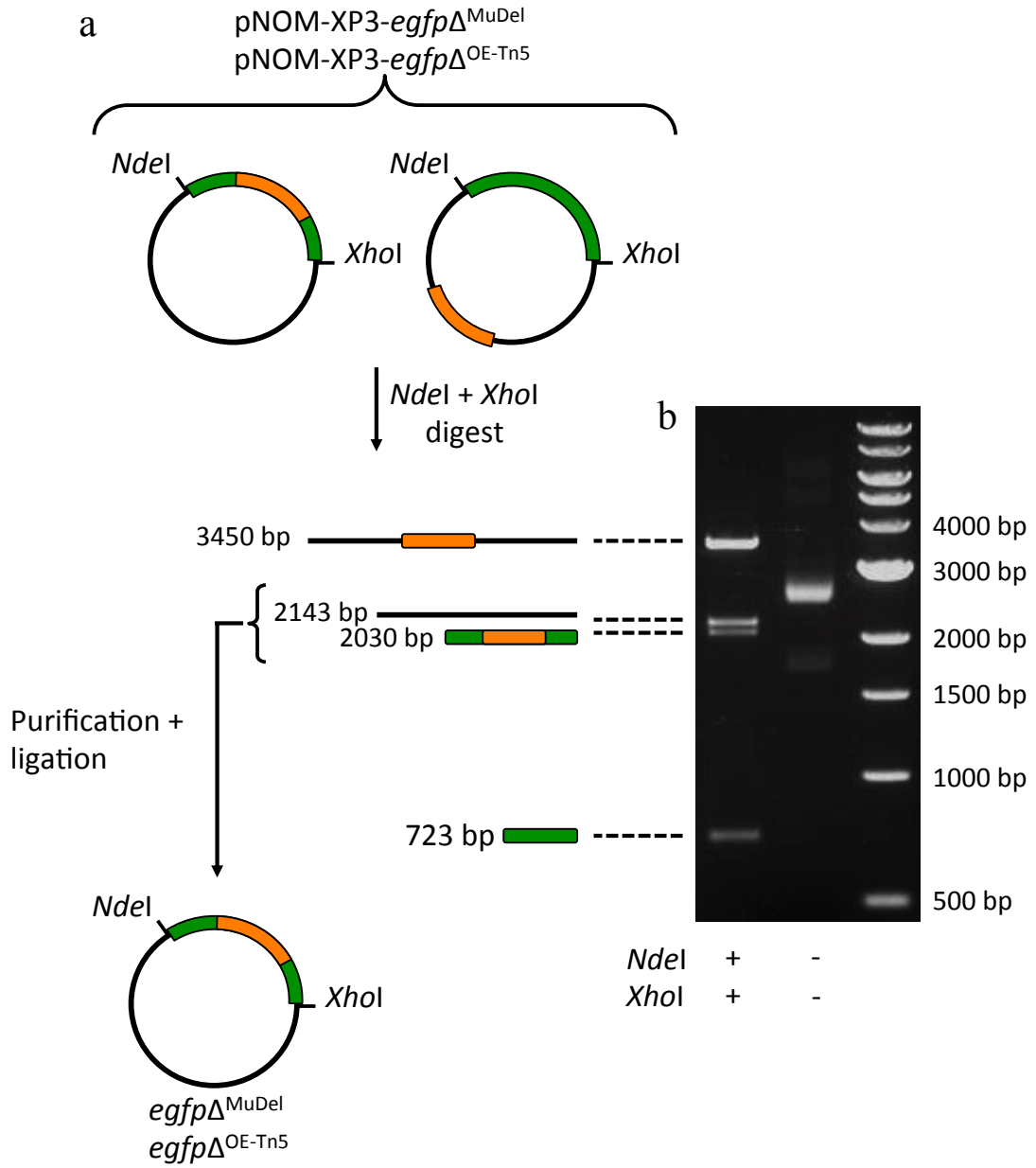


Fig 3.5. Approach for isolation of variants containing a transposon insertion into *egfp*. a. Schematic representation for isolation of desired fragments. **Step 1.** pNOM-XP3-*egfp*Δ^{MuDel} and pNOM-XP3-*egfp*Δ^{OE-Tn5} (vector: black, gfp: green, transposon: orange) are digested with NdeI and XhoI producing four different sized DNA fragments: vector or *egfp* with and without a transposon insertion. **Step 2.** The two DNA fragments representing *egfp* containing a transposon and the plasmid backbone are purified from an agarose gel and ligated together to produce *egfp*Δ^{MuDel} and *egfp*Δ^{OE-Tn5} libraries. **b.** Restriction digest analysis of pNOM-XP3-*egfp*Δ^{MuDel} using *Nde*I and *Xho*I in combinations as indicated in the figure.

insertions just within *egfp* (*egfp*Δ^{MuDel} and *egfp*Δ$^{OE\text{-}Tn5}$) (Fig 3.5 a).

The number of colonies observed on each LB agar plate was 450 and 424 for the MuDel and OE-Tn5 library ligation respectively. This implies that there are 27,675 and 8,056 cells respectively that contain *egfp*-transposon variants in the liquid cultures. The library ligation producing the *egfp*Δ^{MuDel} library resulted in ~10 fold more cells than number of variants in the library (2,504) giving confidence that all possible variants have been sampled.

3.2.5 Diversity of *egfp*Δ^{MuDel} and *egfp*Δ$^{OE\text{-}Tn5}$ libraries

After the isolation of variants containing transposon insertions just within *egfp* it was necessary to check that no diversity was lost from the libraries and that no bias had been introduced into the library. This is important for future screening steps (Section 3.2.6 and 3.7) so that a diverse number of sites are selected and that particular insertion positions do not dominate within the libraries.

*egfp*Δ^{MuDel} and *egfp*Δ$^{OE\text{-}Tn5}$ library diversity was determined by *Xho*I and *Mly*I restriction digestion, as before (Section 3.2.3.2), and analysed by agarose gel electrophoresis (Fig 3.6). The *egfp*Δ^{MuDel} and *egfp*Δ$^{OE\text{-}Tn5}$ libraries only have transposon insertions within *egfp* therefore double restriction digest produces DNA fragments between two distinct size ranges, 0-713 bp and 2147-2860 bp, as outlined in Fig 3.6 a.

3.2.5.1 *egfp*Δ^{MuDel} library diversity

Restriction digestion of the *egfp*Δ^{MuDel} library with *Xho*I produced a single fragment of 4163 bp as expected showing that there is only a single transposon per molecule of DNA (Fig 3.6 b). Restriction digestion with *Mly*I resulted in the two expected fragments corresponding to linear library DNA (2863 bp) and excised MuDel (1310 bp) (Fig 3.6 b). The double restriction digest with *Mly*I and *Xho*I produced fragments between the two expected size ranges (0-713 bp and 2147-2860 bp).

The DNA smears generated on separation by agarose gel electrophoresis are smooth with no regions of greater intensity suggesting there is no obvious bias for a particular transposon insertion position within *egfp*.

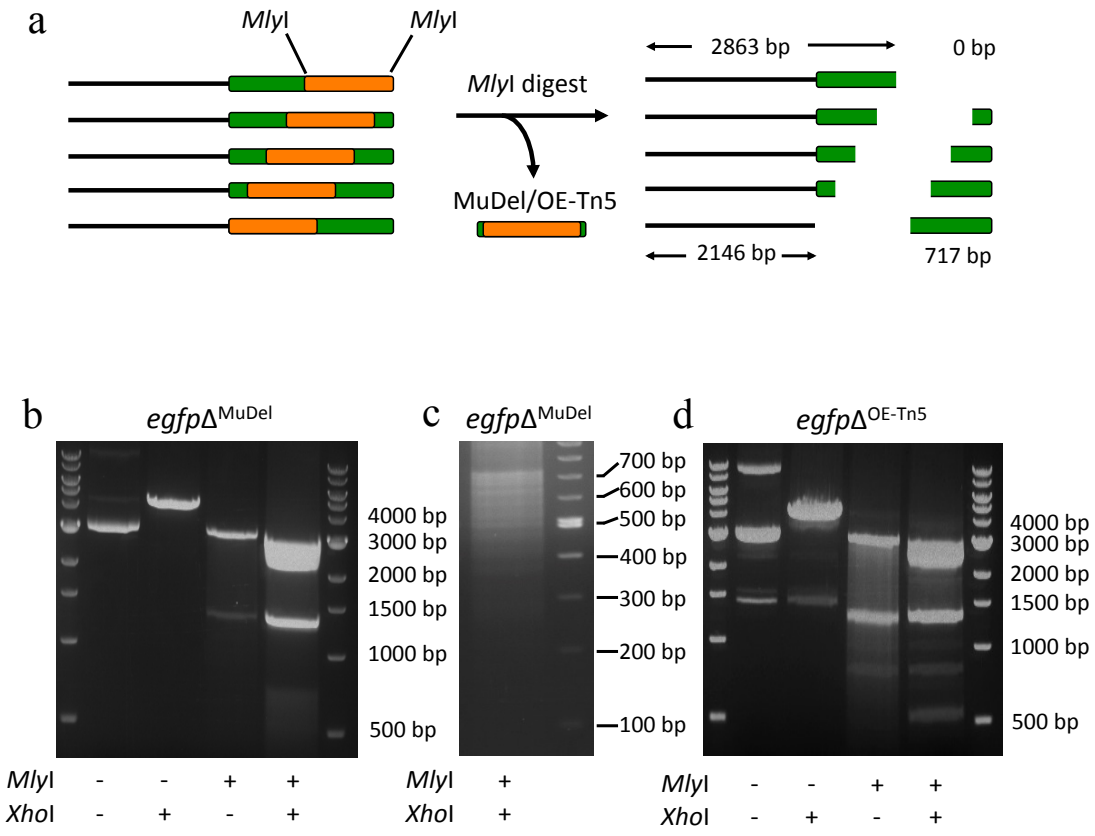


Fig 3.6. Analysis of transposon insertion diversity for *egfp*Δ$^{\mathbf{MuDel}}$ and *egfp*Δ$^{\mathbf{OE\text{-}Tn5}}$.
a. Schematic representation of the diversity analysis process. When linearized *egfp*Δ$^{\mathrm{MuDel}}$ and *egfp*Δ$^{\mathrm{OE\text{-}Tn5}}$ (vector: black, *e*gfp: green, transposon: orange) are digested with *Mly*I the transposon (~1300 bp) is liberated leaving different DNA fragments between two distinct size ranges, 0-713 bp or 2147-2860 bp. Restriction digest analysis of the *egfp*Δ$^{\mathrm{MuDel}}$ library by **b,** 1.0% (w/v) agarose **c,** 2.0% (w/v) agarose and **d,** the *egfp*Δ$^{\mathrm{OE\text{-}Tn5}}$ library by 1.0% (w/v) agarose, with *Mly*I and/or *Xho*I as indicated in the figure.

Analysis of the double restriction digest by 2.0 % (w/v) agarose gel electrophoresis was performed to give better resolution of the smaller DNA fragments produced. The DNA smear only appears to span ~720 bp to ~200 bp and fragments smaller than this can't be seen in the agarose gel image (Fig 3.6 c). This is due to the fragments being smaller and therefore not binding as much ethidium bromide (used to stain DNA, Section 2.2.2) as larger fragments, giving decreased band intensity.

3.2.5.2 *egfp*Δ$^{OE\text{-}Tn5}$ library diversity

The undigested OE-Tn5 library, when analysed by agarose gel electrophoresis (Fig 3.6 c), appeared to consist of three plasmid forms presented by three band intensities at ~9000 bp, 3000 bp and ~1500 bp. Restriction digest by *Xho*I alone linearized the plasmids presenting at ~9000 bp and ~3000 bp in the undigested library however did not linearize the ~1500 bp form (Fig 3.6 c). The plasmid species giving rise to the band intensity at ~9000 bp is probably nicked plasmids in an open conformation rather than supercoiled. The plasmid species giving rise to the band intensity at ~1500 bp could be denatured supercoiled DNA, which can be resistant to some restriction digestion, often formed by an extended alkaline lysis step during plasmid isolation.

Restriction digestion by *Mly*I excised the OE-Tn5 transposon (1308 bp) from the library DNA (4173 bp) as expected (Fig 3.6 c). The double restriction digest of the *egfp*Δ$^{OE\text{-}Tn5}$ library with *Mly*I and *Xho*I produced one of the expected DNA smears (2147–2860 bp) when separated by agarose gel electrophoresis, however the DNA smear representing fragments within the size range of 0-713 bp appeared to be concentrated into three DNA fragments with sizes of ~500, ~800 and ~1000 bp (Fig 3.6 c). This DNA band pattern was not expected and implies the library is biased for particular insertion positions. Due to the fact that the *egfp*Δ$^{OE\text{-}Tn5}$ library had limited diversity all further library work was conducted with the MuDel library.

3.2.6 Creation and analysis of a triplet nucleotide deletion library in *egfp*.

To study the effects of single amino acid deletions on the structure and function of EGFP, the *egfp*Δ^{MuDel} library was used to generate a triplet nucleotide deletion (TND) sub-library. Removal of MuDel from the *egfp*Δ^{MuDel} library by *Mly*I

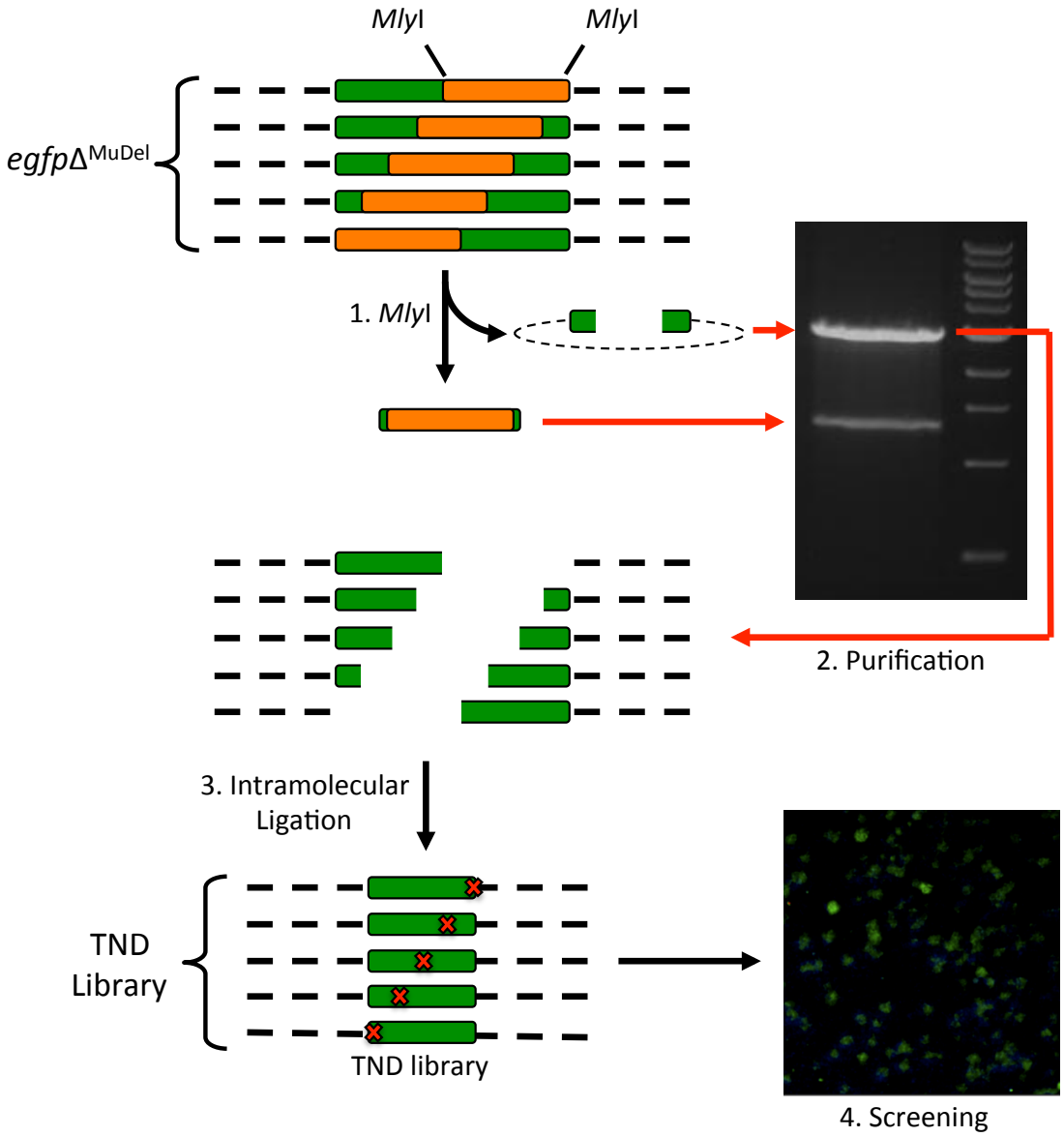


Fig 3.7 Schematic representation for the construction and screening of an egfp triplet nucleotide deletion library. ***Step 1*****,** *Mly*I restriction digestion removes MuDel (orange) and a triplet nucleotide from the pooled plasmid library, *egfp*Δ^{MuDel}. ***Step 2*****,** Separation of the library DNA from MuDel is performed by agarose gel electrophoresis with subsequent extraction and purification (Section 2.2.1.2) of the linear library. ***Step 3*****,** Intramolecular ligation, with NEB Quick ligase (Section 2.2.8), recircularizes the plasmid library resulting in the loss of 3 bp (red cross) from *egfp* (green). ***Step 4*****,** Bacterial colonies transformed with the TND library are screened by their fluorescence. The bacterial colonies expressing an EGFP variant are excited by illumination on a UV transilluminator (Section 2.3.3)

restriction digestion, subsequent purification and intramolecular ligation of the linearized library (Fig 3.7) resulted in the deletion of contiguous triplet nucleotides at random positions in *egfp* (Section 3.1, Fig 3.1).

Transformation of *E. coli* BL21 Gold (DE3) cells (Table 2.1) with the TND library and subsequent plating on M9 minimal media agar plates supplemented with 100 μg/ml ampicillin and 150 μM IPTG produced colonies after 24 hr at 37 °C, 2.5% of which exhibited fluorescence when illuminated on a UV-transilluminater. The plates were then incubated for a further 2 weeks at 4 °C with the number of fluorescent colonies increasing to 10%.

A total of 153 colonies were chosen based on their observable colour phenotype and the *egfp* gene isolated and sequenced: 88 fluorescent and 65 non-fluorescent colonies. Of the 88 fluorescent variants selected 42 unique TNDs were identified and from the 65 non-fluorescent variants 45 were unique TNDs (87 total unique TNDs). A full list of these variants is shown in Table 4.3 and 4.4 in Chapter 4.

MuDel can randomly insert itself into the target gene therefore the TND produced, after *Mly*I restriction and removal of the transposon, is not always in frame. This means that a TND can give rise to the deletion of an amino acid and cause a point mutation (Table 3.1). Out of frame TND can therefore give rise to a stop codon being introduced prematurely into *egfp* depending on the sequence surrounding the TND site (Table 3.1).

From the 153 variants identified and the sequence determined only three variants resulted in an out-of-frame TND and subsequent stop codon being introduced. Sequence analysis indicated that from the 87 unique TNDs 53 (60%) were in the second and third open reading frame (ORF) with 29 of these TNDs producing secondary mutations. Analysis of all TND positions showed they were spread randomly throughout the target gene (Fig 3.8). There is an apparent clustering of TNDs within *egfp* that give rise to fluorescent or non-fluorescent versions of EGFP suggesting that some sites/regions are more or less tolerable to single amino acid deletion respectively.

A detailed characterisation of fluorescent single amino acid deletion variants of EGFP selected from the TND library is described in Chapter 4.

Table 3.1. The possible DNA and protein mutations caused by TND with respect to MuDel insertion position within a codon.

Wild type sequence[a]	Deletion Sequence[b]	Amino Acid Mutation[c]	ORF[d]
$_{15}$GAG GAG CTG$_{25}$	$_{15}$GAG CTG$_{22}$	E6Δ	1
$_{24}$TTC ACC GGG$_{34}$	$_{24}$TTC AGG$_{31}$	T9Δ G10R	2
$_{678}$GCC GGG ATC$_{688}$	$_{678}$GCG ATC$_{685}$	G228Δ	3
$_{546}$TAC CAG CAG$_{556}$	$_{546}$TAG CAG$_{553}$	Y182STOP Q183Δ	3

[a]Wild type egfp sequence with the 5bp duplication caused by MuDel transposition highlighted in bold, with the triplet nucleotide to be deleted underlined.
[b]The resulting DNA sequence after TND, stop codon highlighted in red.
[c]Amino acid mutation caused by TND, a Δ after a residue number indicates that residue has been deleted.
[d] Open reading frame (ORF)

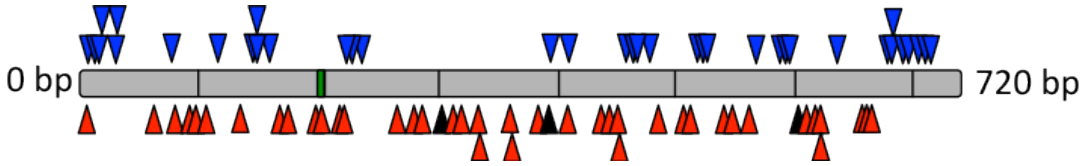


Fig 3.8. Triplet nucleotide deletion (TND) positions within egfp. Sequence analysis of fluorescent (blue triangles) and non-fluorescent variants (red triangles) selected during the screening process identified the position of the triplet nucleotide deleted from *egfp* (grey bar). Non-Fluorescent variants due to a TND and subsequent introduction of a premature stop codon are highlighted by black triangles.

3.2.7. Creation of libraries encoding domain insertion chimeras

The random breaks produced within *egfp* after removal of MuDel by *Mly*I restriction digestion can have DNA cassettes, encoding another protein domain, ligated into the library. The resulting constructs encode integral domain fusion proteins with one protein domain (insert domain) randomly inserted throughout another (parent domain). Integral fusion proteins have the potential to act as biomolecular switches with the input domain modulating the function of the parent domain. The gene encoding cytochrome b_{562} (*CybC*) has been used here as a DNA cassette for insertion into the single breaks within *egfp* to encode cyt b_{562}-EGFP integral fusion chimeras.

MuDel was removed from the *egfp*Δ^{MuDel} library by *Mly*I restriction digestion (Section 3.2.6) and separated from the library DNA by agarose gel electrophoresis (Fig 3.7). The linear *egfp*Δ library was purified from the agarose gel ready for the insertion of the DNA cassette *cybC-Kan*^{r} and production of domain insertion libraries.

The cybC-Kan^r cassette comprises a kanamycin resistance gene, for selection of clones containing an insertion, straddled by two *Pst*I restriction sites within a fragment of *cybC* encoding the cytochrome b_{562} (cyt b_{562}) domain without its N-terminal periplasmic location signal sequence (Fig 3.9) (supplied courtesy of Dr Wayne Edwards) [31].

Previous studies of integral domain fusion proteins has shown that the length and flexibility of the linking peptides, connecting the two domains, has a strong influence on the insert domains ability to modulate the parent domains function. Two *cybC-kan*^{r} cassettes were therefore generated by PCR (Section 2.3.4): one with DNA sequences coding for flexible Gly-Gly-Ser linkers at both N- and C-termini (*cybC-kan*^{r}*-GGS-1*) and a second with a degenerate nucleotide sequence (NNS) encoding a single random amino acid (X) at both N- and C-termini (*cybC-kan*^{r}*-X-1*) (Table 3.4).

As previously described (Section 3.2.6) MuDel can insert itself randomly within a target gene and therefore the random break produced upon MuDel removal from *egfp*, by *Mly*I restriction digestion, is not always in frame. Ligation of a DNA cassette into random breaks in the second or third reading frame would produce out of frame constructs with the likely introduction of a premature stop codon.

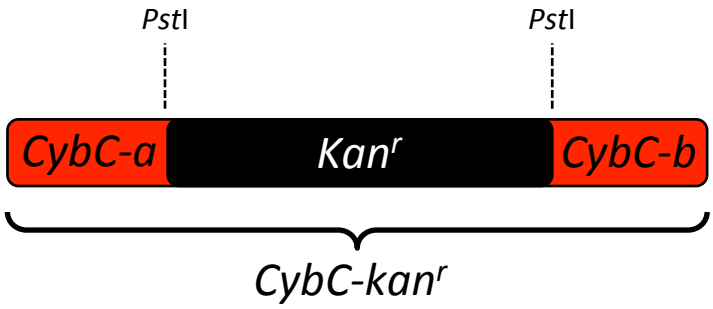


Fig 3.9 Schematic of *cybC-kan*^{r} cassette. The *cybC-kan*^{r} DNA cassette comprises the *cybC* gene (red), encoding the cyt b_{562} domain without its periplasmic location signal sequence, containing a kanamycin resistance gene (kan^r:black) within *Pst*I restriction sites.

Table 3.2. Primer combinations for the generation of cybC-GGS and cybC-X cassettes in all three open reading frames.

Primers[a]	CybC cassette name	ORF[b]	CybC Cassettes[c]
DDJdi023 DDJdi024	cybC-kan^r-GGS-1	I	GGTGGGAGCGCA CybC-kan^r AGGGGTGGGAGC
JAJA003 JAJA004	cybC-kan^r-GGS-2	II	NNGGTGGGAGCGCA CybC-kan^r AGGGGTGGGAGCN
JAJA005 JAJA006	cybC-kan^r-GGS-3	III	NGGTGGGAGCGCA CybC-kan^r AGGGGTGGGAGCNN
WREcbf005 WREcbf006	cybC-kan^r-X-1	I	NNSGCA CybC-kan^r AGGNNS
WREcbf007 WREcbf008	cybC-kan^r-X-2	II	NSNNSGCA CybC-kan^r AGGNNSN
WREcbf009 WREcbf010	cybC-kan^r-X-3	III	SNNSGCA CybC-kan^r AGGNNSNN

[a]Primer sequences can be found in Table 2.2.
[b] ORF refers to the open reading frame of each cassette.
[c]CybC-kan^r sequence contributing to the cassette is highlighted in white on a black background with linking sequences displayed as appendages. Random nucleotides maintaining the reading frame are underlined. N refers to A, C, G or T and S refers to a C or G.

With the addition of single nucleotides to the 5' and 3' ends of the DNA cassette both the second and third reading frames can be sampled (Table 3.2) after insertion into out of frame breaks in the library. The DNA cassettes cybC-kan^r-*GGS-1* and cybC-*kan^r*-*X-1* were used as PCR templates to generate cassettes in open reading frame (ORF) 2 and 3 (*cybC-kan^r*-GGS-2, cybC-*kan^r*-GGS-3, cybC-*kan^r*-X-2, *cybC-kan^r*-X-3) (Section 2.3.4) by the addition of either one or two random nucleotides to the 5' or 3' ends (Table 3.2) (Fig 3.10 b).

The six *cybC* cassettes were ligated into the linear *egfp*Δ library producing six separate DNA cassette insertion libraries (GGS1, GGS2, GGS3, X1, X2, X3) (Fig 3.10). To make sure the cassette insertion libraries sizes were large enough to potentially cover insertions throughout *egfp*, *E. coli* DH5α bacteria transformed with the libraries (5%) were grown on six separate LB agar plates, supplemented with 25 μg/ml kanamycin. The rest of the transformed cells (95%) were grown in LB Broth supplemented with 25 μg/ml kanamycin (Fig 3.10 a).

The GGS1, GGS2 and GGS3 transformed cells returned 771, 352 and 226 colonies respectively on the LB agar plates therefore each liquid culture can be said to comprise of 14,649, 6688, and 4294 variants respectively. The same procedure was applied to the X1, X2 and X3 libraries generating 63, 32 and 33 colonies, respectively, for a subset (5%) of the transformed cells extrapolating to 1197, 608 and 627 variants in the three libraries. The X1, X2 and X3 libraries have far fewer variants than the GGS libraries but still all six contain enough variants to sample insertions throughout *egfp*.

The GGS1 library had >2-two fold and >3-fold more variants than the GGS2 and GGS3 libraries respectively as did the X1 library with almost 2 fold more variants than the X2 or X3 libraries. Knowing the number of variants in each of the libraries a proportional amount of DNA isolated from the GGS1, GGS2 and GGS3 was pooled together so that one particular reading frame did not over represent itself in the pooled library (GGS_{pool}) (Fig 3.10 a). The same procedure was also carried out with the X1, X2 and X3 libraries to produce a second proportionally pooled library (X_{pool}).

A *Pst*I restriction digest of the GGS_{pool} or X_{pool} libraries removed the kanamycin resistance gene from within the *cybC-kan^r* cassettes (Fig 3.10 c). When separated by agarose gel electrophoresis two products are generated corresponding to

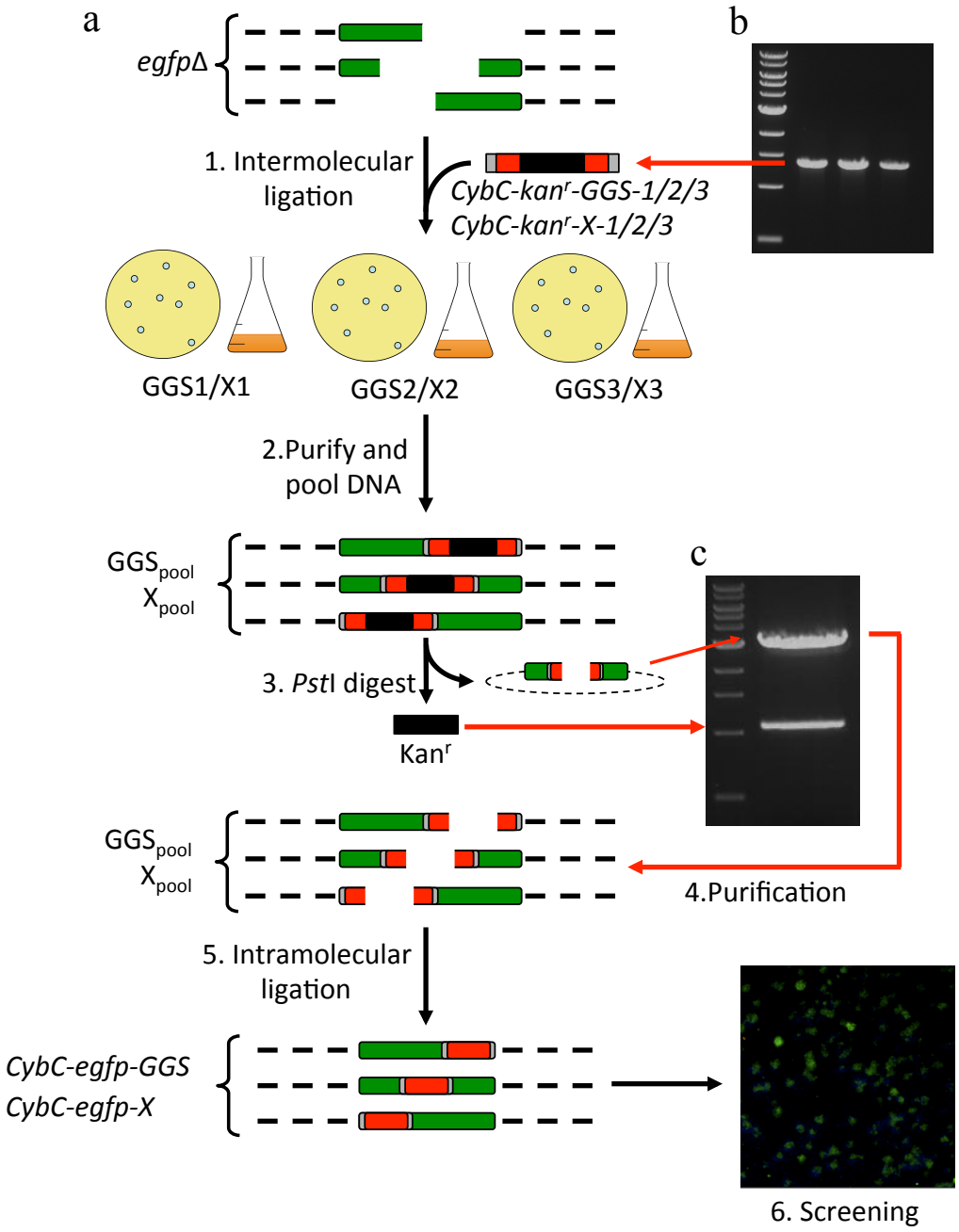


Fig 3.10 Schematic representation for the construction and screening of a domain insertion library. a, *Step 1*, Intermolecular ligations between the linear *egfp*Δ library (green) and *cybC-kan^r* cassettes (*cybC*: red, *kan^r*: black, linkers: grey) are used to transform *E.coli* DH5α cells. Transformed cells are grown on LB agar plates and in LB Broth to calculate the size of the libraries. ***Step 2*,** Plasmid libraries (GGS1, GGS2, GGS3, X1, X2 and X3) are purified from the liquid cultures and pooled so that each ORF is equally represented to produce libraries GGS$_{pool}$ and X$_{pool}$. ***Step 3*,** *Pst*I restriction digestion removes the *kan^r* gene (black) from the library. ***Step 4*,** Isolation and purification of the linear GGS$_{pool}$ and X$_{pool}$ libraries. ***Step 5*,** Intramolecular ligation produces the domain insertion libraries *cybC-eGFP-GGS* and *cybC-egfp-X*. ***Step 6*,** Bacterial colonies transformed with the domain insertion libraries are screened by their fluorescence after excitation with a UV transilluminator. **b,** *CybC-kan^r* DNA cassettes analysed by agarose gel electrophoresis. **c,** *Pst*I restriction digest of GGS$_{pool}$ separated by agarose gel electrophoresis.

the kanamycin resistance gene (1077 bp) and linear GGS_{pool} or X_{pool} library DNA at ~3200 bp (Fig 3.10 c). The library DNA was purified from the agarose gel followed by intramolecular ligation generating *cybC-egfp-GGS or cybC-egfp-X* plasmid libraries (Fig 3.10 a). The *cybC-egfp-GGS or cybC-egfp-X* libraries were used to transform *E. coli* TUNER™ (DE3) cells (Section 2.1.3, Table 2.1) and plated onto M9 minimal media agar plates supplemented with ampicillin (100 μg/ml) and IPTG (150 μM).

3.2.8 Screening *cybC-egfp-GGS* and *cybC-egfp-X* libraries.

E. coli TUNER™ (DE3) cells transformed with *cybC-egfp-GGS or cybC-egfp-X* libraries were spread on M9 minimal media agar plates, supplemented with ampicillin (100 μg/ml) and IPTG (150 μM), and incubated at 37 °C for 24 hrs. The colonies were analysed by eye for a green colour phenotype on excitation with a UV transilluminator (Section 2.3.6) and selected along with colonies presenting no colour, upon excitation, for a PCR screen to estimate the *cybC* insertion positions within *egfp*. The plates were then stored at 4 °C with colonies selected for PCR colony screens after 24 hrs and 72 hrs. After the initial 24 hr incubation at 37 °C 5% of the *cybC-egfp-GGS or cybC-egfp-X* colonies exhibited a green colour phenotype. After storage at 4 °C for up to 72 hrs an additional 1.5% of the colonies exhibited a green phenotype.

Using a primer specific for pNOM-XP3 in combination with a primer specific for *cybC* a PCR product between 456 and 1167 bp in length, depending on the position *cybC* inserted within *egfp*, is generated (Fig 3.11 a). A PCR product of 456 bp indicated *cybC* was inserted towards the 5' end of *egfp* whilst a product of 1167 bp indicated *cybC* was inserted towards the 3' end of *egfp* (Fig 3.11 b).

The PCR products produced from screening colonies transformed with *cybC-egfp-GGS or cybC-egfp-X* libraries display a trend between the size of the product and the time at which the variant gained fluorescence. In both libraries colonies with a green phenotype after the first 24 hrs at 37 °C produce PCR products of ~450 bp or ~1150 bp indicating insertions have taken place towards the 5' and 3' ends of *egfp* only (Fig 3.11 bi). The *cybC-egfp-X* variants only appeared to have insertions in the 5' end of *egfp* where as the *cybC-egfp-GGS* variants had insertions within the 5' and 3' ends of *egfp*.

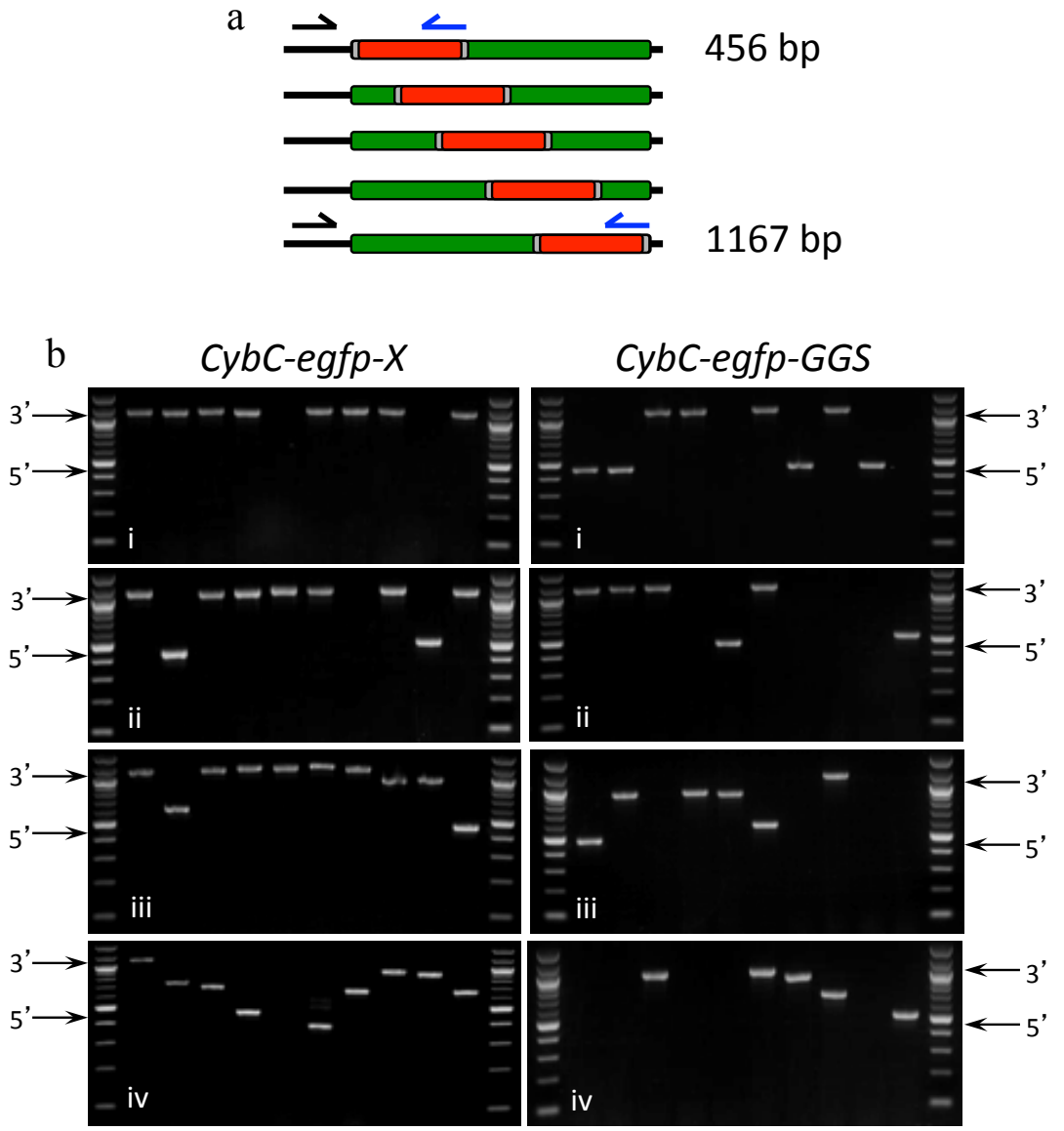


Fig 3.11. Analysis of cybC insertion within egfp. a, Screening of colonies containing *cybC-egfp-GGS* and *cybC-egfp-X* by GoTaq PCR (Section 2.2.3) using primers pEXP-F (black arrow) specific for pNOM-XP3 and WREcbr012 (blue arrow) specific for CybC (Table 2.2) generates products between 456 bp and 1167 bp depending on the insertion position within *e*gfp. **b,** Colony screen analysis of *cybC-efgp-X* (left panel) or *cybC-egfp-GGS* (right panel) libraries by 1.0% (w/v) agarose gel electrophoresis to estimate the site of CybC insertion within egfp. Products ~456 bp indicate an insertion towards the 5' end of egfp and products ~1167 bp indicate an insertion towards the 3' end of egfp as shown in the figure. **i,** Fluorescent colonies screened after 24 hrs at 37°C. **ii,** Screened colonies that gained fluorescence after 24 hrs at 4°C. **iii,** Screened colonies that gained fluorescence after 72 hrs at 4°C. **iv,** Screened non-fluorescent colonies.

Colonies that gained a green colour phenotype after 24 hrs at 4 °C produced PCR products that indicate some of the variants have *cybC* insertions away from the 5' and 3' ends and more integral to *egfp* (Fig 3.9 bii). After 72 hrs at 4°C colonies containing *cybC-egfp-GGS* variants that gained a green colour phenotype produced PCR products that indicated the majority of the variants have *cybC* insertions integral to *egfp*.

Although the *cybC-egfp-X* variants had some insertions towards the middle of *egfp* after 72 hrs at 4 °C the majority of variants screened still had insertions towards the 5' end (Fig 3.9 b iii). Screening of non-fluorescent *cybC-egfp-GGS* and *cybC-egfp-X* variants showed that insertions are taking place throughout *egfp* and the high abundance of 5' end insertions in the *cybC-egfp-X* variants is not due to bias for a particular insertion site in the libraries (Fig 3.9 b iv). Due to the fact that the *cybC-egfp-X* library produced lots of *cybC* insertions towards the 5' end of *egfp* with very few variants having insertions integral to *egfp* all further work was carried out on *cybC-egfp-GGS* variants.

Some of the colony PCR screens produce no PCR product due to the nature of the blunt end ligation between the *cybC* cassettes and the linear library; *cybC* can end up in a correct or reverse orientation. Therefore there is a 50% chance of every colony screened having *cybC* in the reverse orientation to *egfp*. This is seen in the non-fluorescent *cybC-egfp-GGS* variants with 208 colonies being screened, 86 (41%) of which produced no PCR product inferring a reverse *cybC* orientation.

Of the 55 fluorescent *cybC-egfp-GGS* variants screened ten (18%) produced no PCR product from the screen inferring a reverse *cybC* orientation. DNA sequence analysis of the screened fluorescent variants identified four more variants with *cybC* in the incorrect orientation. *CybC* in a reverse orientation usually introduces a premature stop codon resulting in a non-fluorescent protein, however 14 of the colonies screened with reverse *CybC* insertion in *egfp* presented a green colour phenotype.

From the DNA sequence analysis and colony PCR screen the 25% reverse insertions identified in the fluorescent variants took place within 36 bp of the 5' end of *egfp*. The last 36 bp of *egfp* code for an 11 amino acid C-terminal dynamic loop not required for fluorescence to mature, therefore a premature stop codon introduced in this region will still result in colonies presenting a green colour phenotype. These

variants however did not produce a mature cytochrome domain and therefore were discounted from the library and further characterisation.

The colony PCR screen of 26 non-fluorescent *cybC-egfp-X* variants all produced PCR products except in two (7 %) instances. Of the 56 fluorescent *cybC-egfp-X* variants screened by colony PCR 5 (9 %) produced no PCR product implying reverse *cybC* orientation. This is a much lower percentage than the 50% *cybC* reverse orientation expected from a blunt end ligation and that produced by the fluorescent *cybC-egfp-GGS* variants.

This could be because preferential insertion of the *cybC-kan^r-X* cassettes takes place in a particular orientation into the single breaks within the *egfp* library. In this instance there could be read through to the *kan^r* gene within the *cybC-kan^r-X* cassette, from other promoters in the same orientation in the pNOM-XP3 plasmid, giving cells greater resistance to kanamycin and a growth advantage over others. This phenomenon was not seen with the *cybC-kan^r-GGS* cassettes though, which only differ in sequence by ~1% to the *cybC-kan^r-X* cassettes.

From the 55 fluorescent and 208 non-fluorescent *cybC-egfp-GGS* variants screened 37 and 116 were selected, respectively, for sequence analysis.

3.2.9 Sequence analysis of *cybC-egfp-GGS* variants

As previously stated (Section 3.2.6) the nature of MuDel insertion and removal can result in out of frame breaks in the *egfp* coding sequence. This means that there is a 33% chance of a *cybC* cassette being in frame once ligated into the random breaks. An out of frame *cybC* cassette also results in premature stop codons being introduced into the gene usually resulting in a non-fluorescent variant.

Out of the non-fluorescent variants it is expected that 50% could have *cybC* in the reverse orientation and only 33% having *cybC* in the correct reading frame. This means of the 208 non-fluorescent colonies screened, by PCR, only 16% of the screened variants will be in the correct orientation and in the correct reading frame. From the 208 variants screened by PCR 86 (41%) variants had the *cybC* cassette in the incorrect orientation, evident from a lack of PCR products for these variants (data not shown). From the 116 non-fluorescent variants selected for sequence analysis 33 (28%) were determined to be in frame (Fig 3.12). Therefore from the total 208 variants screened 33 (16%) were deemed to be in frame and have *cybC* in the correct orientation within *egfp*, which was expected.

DNA sequence analysis of the 37 fluorescent variants identified 6 (16%) that contained the *cybC* cassette out of frame (Fig 3.12) and four variants were identified during sequence analysis with *cybC* in a reverse orientation. All out of frame or reverse orientation *cybC* insertions were located within the last 36 bp of *egfp*. As mentioned previously (Section 3.8) this region of *egfp* encodes a C-terminal dynamic loop in EGFP not required for fluorescence. Therefore incorporation of a premature stop codon, due to out of frame or reverse orientation *cybC* insertions, at the 3' end of *egfp* can still result in fluorescent proteins that lack the mature cytochrome domain.

After sequence analysis of both fluorescent and non-fluorescent variants 83 unique sites within *egfp* were sampled for domain insertion. A more detailed analysis and characterisation of these variants has been carried out in Chapter 5 and 6.

3.2.10 MuDel target site specificity

MuDel has been shown in many instances to be able to insert itself throughout a target gene producing a diverse number of insertion sites. Until now there hasn't been enough sequence data to identify if a target site-specific consensus sequence exists with which the MuDel/MuA system has a higher preference for. Screening and DNA sequencing data from the TND and domain insertion libraries yielded 152 unique insertion sites throughout *egfp*. Sequence analysis of the 152 unique 5 bp stretches of the target gene duplicated upon MuDel transposition was performed to try to identify a target site preference.

Using an online sequence analysis tool, called Weblogo, the propensity with which a particular nucleotide appears at one of the five possible positions is presented by a graphical representation of a frequency plot (Fig 3.13). From the graphical representation of the sequences there isn't a clear consensus sequence although there are a higher proportion of G/C rich pentamer sequences implying MuDel may have a higher target site preference for G/C rich stretches of DNA.

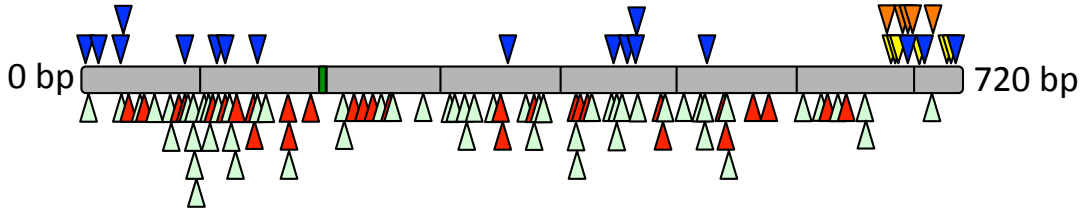


Fig 3.12. CybC insertion positions within *egfp*. Sequence analysis of fluorescent (blue triangles) and non-fluorescent variants (red triangles) selected during the screening process identified the position of the triplet nucleotide deleted from *egfp* (grey bar). Reverse *cybC* insertions or out of frame *cybC* insertions resulting in fluorescent variants are highlighted by orange and yellow triangles respectively. Non-Fluorescent variants due to an out of frame cyt b insertion and subsequent introduction of a premature stop codon are highlighted by green triangles.

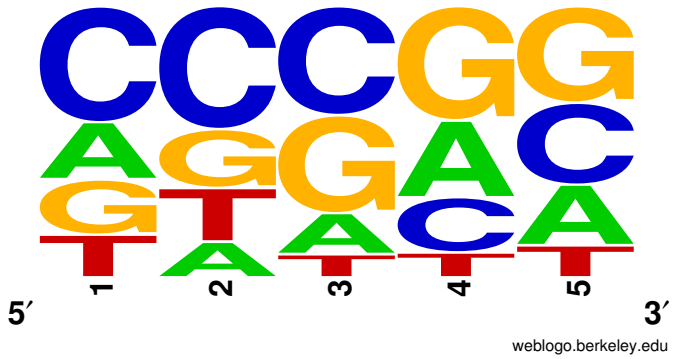


Fig 3.13 MuDel target site sequence analysis. A graphical representation of the frequency at which nucleotides appear at one of the 5 positions of the target site duplication, introduced during MuDel transposition (Section 3.1, Fig 3.1). The graphical representation was produced using the WebLogo application (http://weblogo.threeplusone.com/create.cgi) from 152 unique sequences sampled by MuDel from the TND and domain insertion libraries.

3.3 Discussion

Directed evolution approaches towards protein engineering have many benefits over rational design. Structural knowledge of the target protein is not required, many different variants can be produced at the same time and residues sampled that may not usually be targeted using rational design. However, directed evolution is normally restricted to sampling simple mutation events such as base substitutions, which restricts the sequence space hence conformational space a protein will sample [100]. Here it is demonstrated that using a transposon-based approach, mutations that affect the protein backbone as well as side chain can be sampled by directed evolution.

3.3.1 Comparison of MuDel, ME-4A-Tn5 and OE-Tn5 transposon systems

Here we have described and compared three transposon-based systems; MuDel, ME-4A-Tn5 and OE-Tn5. Due to the mechanism by which the transposons insert themselves into the target gene and the subsequent removal by *Mly*I restriction digestion two libraries can be made with either a triplet nucleotide deletion or a triplet nucleotide duplication of a target gene (Section 3.1, Fig 3.1). This alone can be used for studying the effects of single amino acid deletions (MuDel system) or single amino acid duplications (Tn5 system) within target proteins.

Two Tn5 transposon systems were used here to test the efficiency of the two different transposons OE-Tn5 and ME-4A-Tn5. The two transposons only differed in sequence by two nucleotides within their respective transposase recognition elements (TREs) (Fig 3.1 c). The ME-4A-Tn5 transposon was developed from a hyperactive ME-Tn5 transposon, which shows up to 8-fold increased transposition efficiency with respect to the OE-Tn5 transposon [98]. However the single nucleotide mutation in the ME-4A-Tn5 transposons TREs used in this study, to introduce the required *Mly*I restriction site, reduced its transposition efficiency by 5-fold with respect to that of the OE-Tn5 transposon (Section 3.2.2).

The decrease in activity of the ME-4A transposition reaction could be attributed to the change in the base pair at position 4 of the transposase recognition sequence from a thymine to an adenine [98]. It has been shown that the C5 methyl group on thymine 4 of the ME recognition sequence is required for efficient transpososome complex formation (a pre transposition DNA/protein complex, Section

1.3.3.2) and that alteration of the thiamine base to the corresponding OE sequence (adenine) results in reduced transposition efficiency [98].

Despite the OE-Tn5 transposon system having similar transposition efficiencies to the MuDel system (Section 3.2.2) further library production steps indicted that the OE-Tn5 libraries contained bias for particular transposon insertion positions within the target gene, *egfp* (Fig 3.6 d). For this reason the OE-Tn5 library was not used to sample potential triplet nucleotide duplications within *egfp*. Previous studies have identified significant target site preference for Tn5 transposons at GC rich target DNA sequences [101, 102]. Engineering of the DNA binding domains in the Tn5 transposase could potentially alleviate its target site specificity and make it a more useful system for creating directly evolved libraries.

The MuDel system was taken forward and used to produce two further libraries, a triplet nucleotide deletion library (Section 3.2.6) and a domain insertion library (Section 3.2.7). In contrast to the Tn5 system both individual libraries sampled a diverse range of insertion positions (Fig 3.8 and 3.12) within *egfp* with no obvious bias for any specific position (Fig 3.6 b and c). Sequence analysis of the 5 bp target site duplications, introduced due to the transposition mechanism (Fig 3.1), from the 152 target sites sampled across the TND and domain insertion libraries failed to identify a specific consensus sequence for which MuDel has a higher preference (Fig 3.13). This further reinforces the benefits of the MuDel system over the Tn5 systems investigated here.

3.3.2 Directed evolution approaches for sampling deletion mutations and domain insertion

Until recently the most common way of introducing single amino acid deletions was by rational design, requiring structural knowledge of the target protein [39], and the requirement of separate oligonucleotides for every mutation required. Advancement in directed evolution strategies has established a few alternative methods for sampling deletion mutations for example random insertion and deletion mutagenesis (RID) [48]. The RID mutagenesis method to date has only been used for insertion and substitution mutagenesis but could potentially be used for sampling deletion mutations [48]. The limitation of this method is the use of Ce(IV)-EDTA to chemically cleave the target DNA as multiple nicks can be introduced into the target gene and the procedure is complicated and time exhaustive. On top of this the

resulting libraries produced by RID mutagenesis exhibit extensive bias for sites sampled within a target gene [48].

What is integral to all of these procedures (MuDel or Tn5 transposition and RID) is the generation of single random breaks throughout the target gene. There are already several techniques available to achieve random breaks within a target gene, such as the non-specific nuclease activity of DNaseI, RID mutagenesis, and other transposon based techniques. However it is notoriously difficult to get single breaks within a target gene using DNAseI, as with Ce(IV)-EDTA chemical cleavage, regularly producing tandem repeats and nested deletions of uncontrollable sizes. Other transposon-based methods can be used to introduce random breaks throughout target genes. However, for the majority of these methods removal of the bulk of the transposon by restriction digestion results in the transposase recognition elements (TREs) and the target site duplication left behind in the target gene (Chapter 1, Section 1.3.1.3). This reduces the ability to design and control the size and composition of the DNA inserted.

The strength of the MuDel method described here is the ability to sample a diverse number of insertion positions and introduce a controlled triplet nucleotide deletion from the target gene leaving a library of blunt ended breaks (Fig 3.7). Unlike with other transposon based methods the MuDel TREs and target site duplication are removed from the library upon *Mly*I restriction digestion. As mentioned previously this gives the freedom and control for downstream mutagenesis techniques to sample single amino acid deletions, domain insertions or for the incorporation of non-natural amino acids to increase the chemical diversity of proteins.

The MuDel insertion library created within *egfp* described here has been used for the incorporation of non-natural amino acids into EGFP [50]. The single break generated, after removal of MuDel from the library, had a DNA cassette (Subseq) comprising a kanamycin resistance gene flanked by *Mly*I restriction sites with an amber stop codon (TAG) at its 5' end ligated into the library. The *Mly*I restriction sites were critically placed so as to remove the kanamycin resistance genes but leaving the TAG codons within the *egfp* library. Using an amber suppression technique [50] a non-natural amino acid can be incorporated at the position of an in frame TAG codon during translation. This technology allows for new chemistry to be sampled in proteins not accessible given the 20 naturally occurring amino acids.

Owing to the nature of MuDel transposition and subsequent removal by *Mly*I restriction digestion the single random breaks produced within a target gene are not always in frame. However, when creating domain insertion libraries the linking sequences separating the two domains can be encoded within the DNA cassette inserts (Section 3.2.7 and Table 3.2). This opens the possibility to sample out of frame insertion sites within the libraries by adding one or two random nucleotides to the 5' or 3' ends of the DNA cassettes, increasing the potential library sample size 3-fold.

Design of the sequences that link two domains in a discontinuous domain architecture is often given little thought even though they have been shown to play an essential role in the functional coupling of resulting integral domain fusion scaffolds [31, 37]. Here we investigated the effects different length linkers play on the tolerance of EGFP to cyt b_{562} domain insertion. Single random amino acid linkers (X) and longer more flexible tripeptide linkers (GlyGlySer) were encoded at the 5' and 3' ends of the *cybC* cassettes (Table 3.2), used for insertion into the blunt ended breaks within the linear *egfp*Δ library (Section 3.2.7). It was evident from PCR screens of colonies containing the *cybC-egfp-GGS* and *cybC-egfp-X* libraries that EGFP was more tolerant to the integral insertion of cyt b_{562} when the two domains were separated by the longer GlyGlySer linkers (Fig 3.11). This is probably due to the inherent difference between the flexibility the different linkers would impart on the integral domain constructs; the longer more flexible linkers will provide the flexibility for the parent domain to fold into its native structure, whilst shorter more rigid linkers would reduce the conformational flexibility required for the parent domain to fold.

It was also observed that EGFP was more tolerable to cyt b_{562} domain insertions towards its C-terminus evident by the high proportion of *cybC* insertions towards the 3' end of *egfp* (Fig 3.11). This is not due to bias in the library for insertions positions at the 3' end of the *egfp* gene but due to the nature of the C-terminal region of EGFP. The last 36 bp of *egfp* encode an 11 amino acid dynamic loop that is not required for fluorescence to mature [62] and is therefore highly tolerant to domain insertions.

Between the two libraries (TND and domain insertion) described in this Chapter, 139 unique sites were sampled for the potential mutagenesis of *egfp*. With further screening of the TND and domain insertion libraries it is highly probable that more unique sites would be identified within the *egfp* library for sampling mutational

events. With the development of improved high throughput screening methods and approaches the number of novel variants that can be identified from diverse libraries will increase, allowing the full potential of directly evolved protein libraries to be exploited. Further analysis and characterisation of identified fluorescent variants from the TND library and the domain insertion library will be discussed in Chapters 4, 5 and 6.

Chapter 4: Characterization and analysis of single amino acid deletion mutants of EGFP

4.1 Introduction

Substitution mutations have been utilized extensively for over 20 years to engineer proteins with new or improved properties and to study mechanisms, folding and stability [5]. Substitution mutations only alter the side chain and therefore leave the protein backbone unaltered. Sampling InDel mutations so as to target the protein backbone would allow new sequence space and hence conformational arrangements to be sampled [19, 39, 100]. Insertion and deletion (InDel) mutations are sampled during protein evolution [15, 16, 103], as is evident by length variations between homologous proteins (Fig 4.1). Due to selection pressures InDel mutations usually occur in segments of 3 base pairs, or multiples thereof, so as to maintain the original open reading frame of the gene [16]. Given this requirement, InDel mutations at the protein level are sampled less often in nature than substitution mutations that only require DNA point mutations.

The problem with mutations that change the backbone is that their effects can be difficult to predict due to the local and global structural rearrangements required to accept removal of an amino acid from the polypeptide [100]. Dogma suggests that loops are likely to be more tolerant to InDel mutations due to the inherent conformational flexibility of these regions, whilst InDel mutations in secondary structures can result in registry shifts (Chapter 1, Fig 1.1) [15]. However, there are very few studies that survey the effects of Indel mutations on protein structure and function. Single amino acid deletion mutagenesis was performed on the antibiotic resistance protein, TEM-1 β-lactamase. A wide variety of residues were identified that tolerated deletion and some variants even displayed altered properties such as improved activity towards normally poor substrates [19, 21].

Given that InDel mutagenesis is not a commonly adopted strategy in protein engineering, the impact of such an important class of mutations on the protein structure-function-folding relationship has not been fully explored. Thus, the established dogma of the generally limited structural sampling and perceived detrimental effect on protein stability persists. This is exacerbated by the lack of directed evolution approaches to sample InDel mutations that would allow sampling

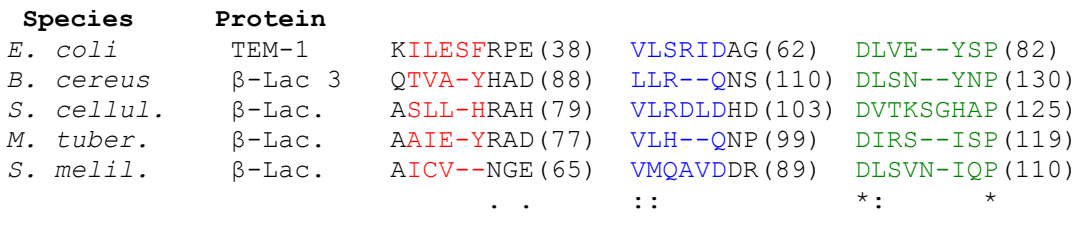


Fig 4.1 Sequence allignments for homologous β-lactamases. Length variations for homologous proteins highlight the importance of InDel mutations in the protein evolution process. Selected stretches of sequence β-lactamases were aligned using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/). * indicates identical residues, : indicates conserved residues, . indicates semi-conserved residues. Red sequence corresponds to residues in a β-strand, blue sequence corresponds to residues in an α-helix and green sequence corresponds to a region in TEM-1 β-lactamase important for binding the β-lactamase inhibitor protein BLIP. β-lac 3 stands for class A β-lactamase III, β-lac stands for β-lactamase with no defined name.

of single amino acid deletions across the polypeptide backbone and variants screened for tolerance and improved activity. A directed evolution approach would allow a broader sampling of InDel mutations allowing variants with altered properties to be observed and analysed thus allowing the effect of the mutation rationalized.

Therefore the aim of this Chapter is to improve our understanding of the impact of the most common Indel event, the single amino acid deletion [16] on EGFP function, structure and folding. A survey of tolerated and non-tolerated single amino acid deletion events was undertaken using a directed evolution approach and to identify variants with altered and beneficial properties. The transposon-based directed evolution approach has already been described in Chapter 3. Selected EGFP variants were then analysed in more detail, including determining their structural and folding properties. On commencement of this project, no structure was available for EGFP despite its wide-ranging use as a tool in cell biology. Therefore, to allow a true interpretation of our observed deletions in terms of EGFP structure, the 3D atomic resolution structure of EGFP was solved by X-ray crystallography.

4.2 Results

4.2.1 EGFP crystallization and structure determination

Prior to this study the atomic level tertiary structure for EGFP had not been determined, which is surprising given its widespread use as a tool in cell biology. To address this and to provide a suitable structural model for analysis of the effect of deletion mutations, the crystal structure of EGFP was solved. EGFP protein samples (10 mg/ml in 50 mM Tris HCl, pH 8.0, 150 mM NaCl) were screened for crystal formation using the sitting drop vapour diffusion method at 4 °C. The crystals for EGFP were obtained from 0.1 M MES/NaOH, pH 6.5, 0.2 M calcium acetate and 18% PEG 8000. The crystallographic statistics are shown in Table 4.1. Crystals grew in the space group $P2_12_12_1$ and contained a single molecule in the asymmetric unit. The EGFP crystals gave good diffraction with an optimum resolution of 1.35 Å. Structure determination was carried out by molecular replacement with a previously solved GFP structure (PDB:2HQZ) using the CCP4 program MOLREP. Structure refinement was performed with the program REFMAC (Section 2.7.2). An R-Free value of 17.8% for the refined EGFP structure implied the model was a good representation of the experimental data.

Table 4.1. Data collection and refinement statistics.

Data collection statistics	
Space group	*P*2(1)2(1)2(1)
a (Å)	51.1
b (Å)	62.2
c (Å)	69.6
Resolution range (Å)	40 - 1.35
Total reflections	225405
Unique reflections	49477
Completeness (%) (last shell)	99.9 (99.8)
I/σ (last shell)	5.0 (2.0)
R(sym) (%) (last shell)	9.3 (38.6)
B(iso) from Wilson ($Å^2$)	9.1
Refinement statistics	
Protein atoms excluding H	1808
Solvent molecules	368
R-factor (%)	15.3
R-free (%)	17.8
Rmsd bond lengths (Å)	0.006
Rmsd angles (°)	1.2
Ramachandran core region (%)	92.7
Ramachandran allowed region (%)	7.3
Ramachandran additionally allowed region (%)	0.0
Ramachandran disallowed region (%)	0.0

Superposition of the structure obtained for EGFP on wt GFP (pdb: 1GFL) shows that the structures are very similar (Fig 4.2) with an RMSD of 0.33 Å across all backbone atoms (0.95 Å across all atoms), indicating that the F64L and S65T mutations of EGFP do not effect the overall protein structure but have more subtle effects. Secondary structure prediction taking into account hydrogen bond energies and main chain dihedral angles [104] showed the secondary structure boundaries between wt GFP and EGFP are very similar except for a difference in the predicted secondary structure adjacent to the chromophore (Table 4.2). In wt GFP the central helix running through the core of the β-barrel is predicted to be a mixture of a 3_{10} helix and an α-helix, whilst in EGFP the entire helix is predicted to be a 3_{10} helix. This would imply the central helix in EGFP is packed better into the core of the β-barrel, which is potentially a consequence of the F64L mutation.

4.2.1.1 Structural effect of the F64L mutation in EGFP

Given the high resolution of the structure, the exact placement of side chains can be defined with high confidence enhancing the molecular description of EGFP. The F64L mutation confers increased folding efficiency to GFP at 37 °C, however, the structural consequences of this mutation had not been investigated prior to this study. The most obvious effect of the F64L mutation comprises the exchange of the bulky phenylalanine side chain for a smaller leucine side chain in the central chromophore containing α-helix. The substitution causes the tighter packing of β-strand 2 into the core of the β-barrel structure, with the largest deviation being between the C_α atom of residue V29 (~0.59Å) from wt GFP and EGFP (Fig 4.2). The V29 residue shifts in position closer to the chromophore given the additional space made available by loss of the bulky F64 side chain. There is also a slight shift of residue W57 away from the surface of the protein towards L64 with an RMSD over the side chain atoms of 0.42 Å (Fig 4.2). The movement of W57 towards the core of the protein is confirmed by a decrease in its solvent accessible surface area in EGFP (12.82 $Å^2$) with respect to wt GFP (15.16 $Å^2$) (solvent accessible surface area calculated as per [105]). Another noticeable effect was a rotation of the L18 isobutyl side chain away from the edge of the β-barrel towards the core of the protein with a 1.82 Å RMSD between the wt GFP

Table 4.2 Secondary structure prediction

Residue[a]	Secondary structure prediction[b]	
	EGFP	wt GFP[c]
K52	Coil	Coil
L53	Coil	Coil
P54	Coil	Coil
V55	Coil	Coil
P56	Coil	Coil
W57	3_{10} helix	3_{10} helix
P58	3_{10} helix	3_{10} helix
T59	3_{10} helix	3_{10} helix
L60	3_{10} helix	3_{10} helix
V61	3_{10} helix	α-helix
T62	3_{10} helix	α-helix
T63	3_{10} helix	α-helix
L64/F64	Coil	α-helix
Cro	Chromophore	Chromophore
V68	Coil	Turn
Q69	3_{10} helix	3_{10} helix
C70	3_{10} helix	3_{10} helix
F71	3_{10} helix	3_{10} helix
S72	Coil	Coil

[a] Residue numbering as for GFP [106]
[b] Secondary structure prediction determined as per [104]
[c] wt GFP PDB:1GFL
Grey shading highlights the residues that comprise the central helix

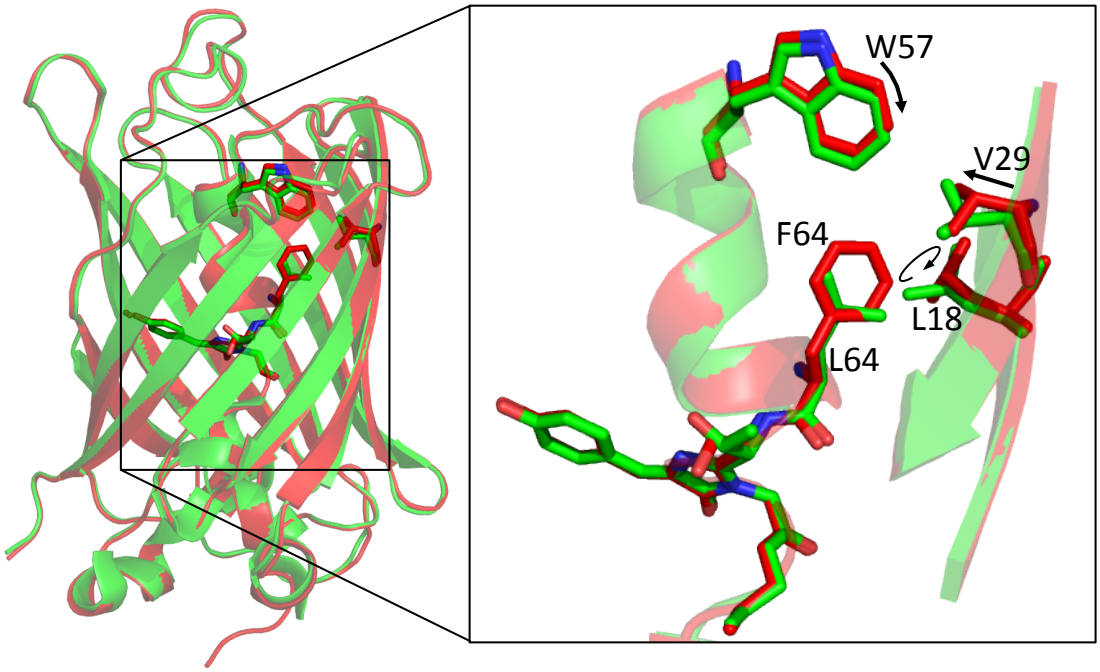


Fig 4.2 Structural effect of the F64L mutation in EGFP. Cartoon representation of EGFP (green) superposed on wt GFP (red, pdb:1GFL) shows that the overall structures are very similar. **Inset:** The subtle structural effects of the F64L mutation are a shift in the position of W57 towards L64 the closer packing of V29 towards L64, and a rotation of the L18 side chain towards the core of the protein.

and EGFP L18 side chain atoms. The repositioning of these residues could potentially be influencing the folding of EGFP at 37 °C by instigating better packing of the hydrophobic residues surrounding the central helix and of the central helix itself.

4.2.1.2 Structural effect of the S65T mutation in EGFP.

The S65T mutation has proved to be a more general mutation that can be transplanted to other green fluorescent protein variants to alter their spectral properties (λ_{ex} shift from 375 nm to ~480 nm) [69, 72], increase the rate of oxidation during chromophore maturation and increase the brightness of the fluorescent proteins [72]. Analysis of the local environment around the chromophore can explain the molecular basis for the observed spectral properties (Fig 4.3 a). The 2Fo-Fc electron density map produced after molecular replacement and structural refinement shows a tridental density for E222, which was successfully modeled as two conformations of the carboxylate side chain (Fig 4.3 b). Due to the S65T mutation, the hydroxyl group of T65 in EGFP occupies a different position to the corresponding hydroxyl group of S65 in wt GFP (Chapter 1, Fig 1.12), probably for steric reasons given the additional methyl group of T65. Consequently the hydroxyl group of T65 donates a hydrogen bond to the backbone carbonyl of V61. This results in the carboxylate of E222 being able to occupy two conformations (Fig 3 a, i and ii).

In both conformations the E222 $O_{\delta 1}$ is hydrogen bonded to two conserved water molecules, whilst in one conformation the E222 $O_{\delta 2}$ donates a hydrogen bond to the hydroxyl group of S205 (Fig 4.3 a i) and in its second conformation it donates a hydrogen bond to the hydroxyl group of T65 (Fig 4.3 a ii). In order for E222 to donate hydrogen bonds its carboxylate group must be protonated and therefore neutral. This allows charge stabilization on the deprotonated tyrosyl group of Y66 by hydrogen-bonding interactions from H148, T203 and a conserved water molecule coordinated between the backbone carbonyl group of N146 and the side chain hydroxyl group of S205. The neutral charge on E222 also removes any potential electrostatic clashes between the negative charges on E222 and the chromophore, thus allowing the chromophore to be deprotonated in the ground state. This explains why EGFP has a single excitation peak corresponding to the deprotonated state.

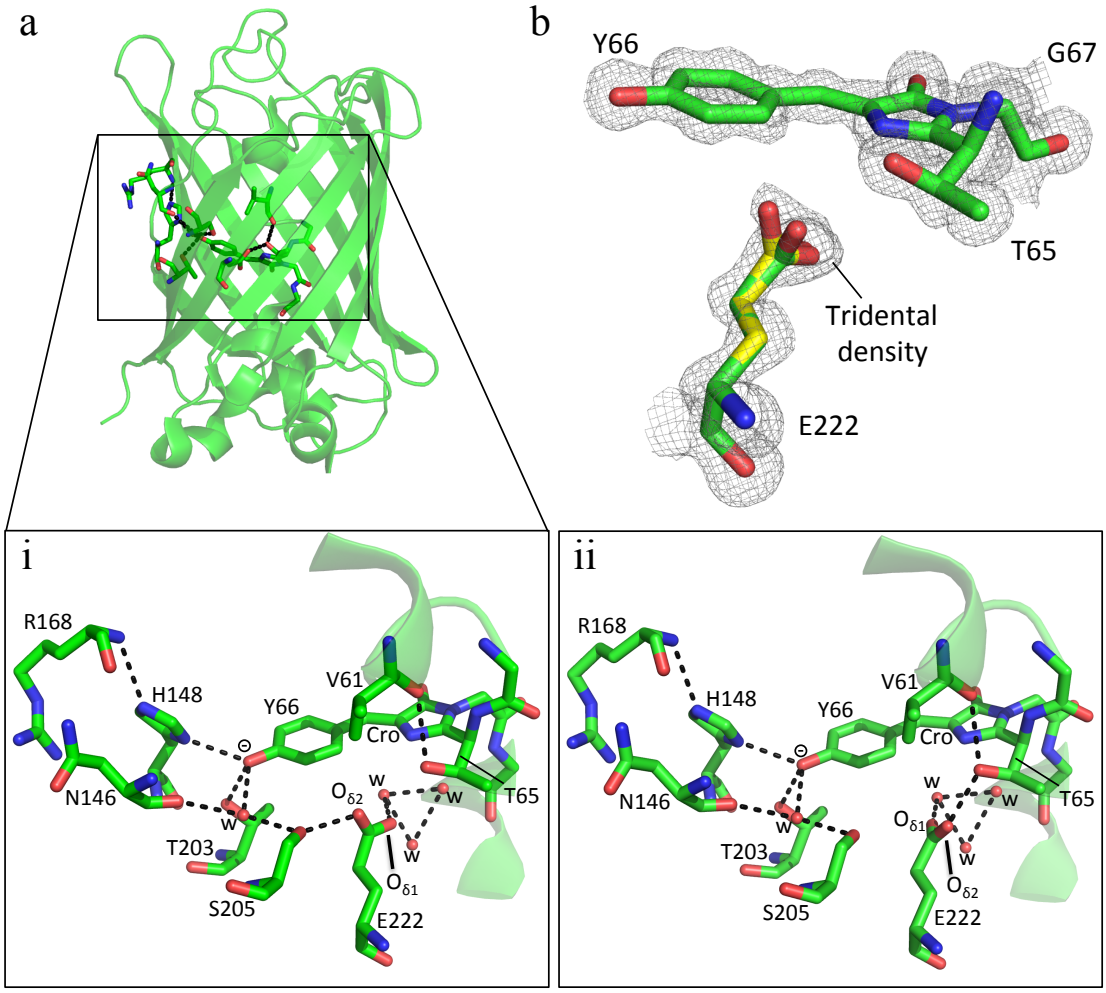


Fig 4.3 Effect of the S65T mutation in EGFP. a, Cartoon representation of EGFP (green) with the chromophore and surrounding residues shown as sticks. The S65T mutation in EGFP results in E222 populating two conformations donating a hydrogen bond (black dotted lines) to **i,** S205 or **ii,** T65. All residues are shown in stick representation with the central α-helix in cartoon representation, Cro signifies cromophore, W signifies water molecules. **b,** F_0-F_c electron density map for the chromophore and residue E222 (sticks) contoured to 1σ. The tridental density for E222 was successfully modeled by two side chain conformations (yellow or green).

4.2.2 Analysis of tolerated and non-tolerated single amino acid deletion positions in EGFP

The construction of a triplet nucleotide deletion library within *egfp*, for sampling single amino acid deletion mutations in EGFP, has been described in detail in Chapter 3. *E.coli* BL21 (DE3) Gold cells transformed with plasmids coding for EGFP single amino acid deletion variants (EGFPΔ), were grown on M9 minimal media agar plates (Section 2.1.4) supplemented with 100 μg/ml ampicillin and 150 μM IPTG, at 37 °C for 24 hr. After overnight incubation colonies presenting a green colour phenotype were noted and incubated for a further 24 hrs at 25 °C, followed by incubation at 4 °C for up to 2 weeks.

After incubation for 2 weeks at 4 °C colonies presenting a green colour phenotype when illuminated on a UV-transilluminator were selected for DNA sequence analysis of the EGFPΔ genes. From-DNA sequence analysis of the EGFPΔ gene from 89 colonies that displayed green fluorescence, 42 unique variants were observed (Table 4.3). Of the 66 colonies that did not fluoresce, DNA sequence analysis revealed 45 unique variants (Table 4.4). Analysis of the single amino acid deletion positions within the secondary (Fig 4.4) and tertiary structure (Fig 4.5) revealed a clear difference in the clustering of mutations giving rise to fluorescent and non-fluorescent variants.

EGFP is comprised of 43% loops, 11% α-helices and 46% β-strands, therefore if all ordered secondary structures and loops comprising EGFP were equally tolerant to single amino acid deletion the proportion of variants identified in these structures should be the same. However, the majority of the mutations giving rise to fluorescent variants were found in the connecting loops (60%), as expected, with the rest equally distributed across α-helices (19%) and the termini of β-strands (21%) (Fig 4.4 a). This indicates that the loops in EGFP are more tolerant to single amino acid deletion whilst β-strands are less tolerant. There are also 2-fold more tolerated deletions from α-helices than would be expected. In particular there are a high proportion of amino acid deletions tolerated C-terminally of β-strands 7 and 11, and in the loops connecting β-strands 7, 8, 9 and 10, all of which are adjacent to one another (Fig 4.4 a) in the tertiary structure.

Of the 45 unique mutations giving rise to non-fluorescent variants 71% were clustered to the middle of β-strands, 18% were located in loops and 11% in α-helices

Table 4.3 Tolerated TNDs in *egfp* and subsequent amino acid mutations

Nucleotide deletion[a]	Amino acid Mutation[b]	Frequency	Secondary structure[c]
$_{3}$G**TG A**GC$_{10}$	V1Δ S2G	2	N-terminus
$_{9}$AA**G GG**C$_{16}$	K3N G4Δ	4	H1
$_{12}$G**GC G**AG$_{19}$	G4Δ	8	H1
$_{12}$GG**C GA**G$_{19}$	E5Δ	2	H1
$_{18}$**GAG**$_{22}$	E6Δ	1	H1
$_{27}$A**CC G**GG$_{34}$	T9Δ G10R	6	H1
$_{27}$AC**C GG**G$_{34}$	G10Δ	2	Loop H1-S1
$_{36}$**GTG**$_{40}$	V12Δ	1	S1
$_{75}$**CAC**$_{79}$	H25Δ	2	S2
$_{114}$**ACC**$_{118}$	T38Δ	3	Loop S2-S3
$_{144}$**TGC**$_{148}$	C48Δ	1	S3
$_{147}$**ACC**$_{151}$	T50Δ	1	Loop S3-H2
$_{150}$A**CC G**GC$_{157}$	T50Δ G51S	2	Loop S3-H2
$_{159}$C**TG C**CC$_{166}$	L53Δ	1	Loop S3-H2
$_{225}$C**CC G**AC$_{232}$	P75Δ D76H	2	H3
$_{225}$**GAC**$_{229}$	D76Δ	2	H3
$_{237}$**AAG**$_{241}$	K79Δ	1	H3
$_{396}$GA**G GA**C$_{403}$	E132D D133Δ	1	Loop S6-S7
$_{411}$**GGG**$_{415}$	G138Δ	2	Loop S6-S7
$_{459}$A**TG G**CC$_{466}$	M153Δ A154T	2	S7
$_{462}$G**CC G**AC$_{469}$	A154Δ	5	S7
$_{465}$**GAC**$_{469}$	D155Δ	4	S7
$_{474}$AA**G AA**C$_{481}$	K158Δ	1	Loop S7-S8
$_{480}$**GGC**$_{484}$	G160Δ	1	S8
$_{513}$AT**C GA**G$_{520}$	I171M E172Δ	3	Loop S8-S9
$_{522}$**GGC**$_{526}$	G174Δ	2	Loop S8-S9
$_{525}$**AGC**$_{529}$	S175Δ	1	Loop S8-S9
$_{567}$G**GC G**AC$_{574}$	G189Δ	1	Loop S9-S10
$_{570}$G**AC G**GC$_{577}$	D190Δ	1	Loop S9-S10
$_{576}$C**CC G**TG$_{583}$	P192Δ V193L	3	Loop S9-S10
$_{588}$**CCC**$_{592}$	P196Δ	1	Loop S9-S10
$_{591}$**GAC**$_{595}$	D197Δ	1	Loop S9-S10
$_{594}$**AAC**$_{598}$	N198 Δ	1	Loop S9-S10
$_{633}$C**CC A**AC$_{640}$	P211Δ N212H	3	Loop S10-S11
$_{678}$G**CC GCC G**GG$_{687}$	A226Δ A227Δ	1	S11
$_{681}$G**CC G**GG$_{688}$	A227Δ	5	S11
$_{681}$GC**C GG**G$_{688}$	G228Δ	2	C-terminus
$_{690}$AC**T CT**C$_{697}$	L231Δ	1	C-terminus
$_{699}$A**TG G**AC$_{706}$	M233Δ D234N	2	C-terminus
$_{702}$GA**C GA**G$_{709}$	D234E E235Δ	2	C-terminus
$_{705}$**GAG**$_{709}$	E235Δ	1	C-terminus
$_{711}$**TAC**$_{715}$	Y237Δ	1	C-terminus

[a] Numbers refer to gene sequence numbering for *egfp* developed by [107]

[b] Δ after a residue number signifies that residue has been deleted, protein numbering as per GFP [106]

[c] Secondary structure elements as defined in Fig 4.4, helices (H), strands (S).

Table 4.4 Non-tolerated TNDs in *egfp* and subsequent amino acid mutations

Nucleotide deletion	Amino acid Mutation	Frequency	Secondary structure
$_{9}$A**AG G**GC$_{16}$	K3Δ G4S	1	H1
$_{60}$G**GC G**AC$_{67}$	G20Δ	3	S1
$_{81}$T**TC A**GC$_{88}$	F27Δ S28C	1	S2
$_{90}$T**CC G**GC$_{97}$	S30Δ G31C	3	S2
$_{99}$GG**C GA**G$_{106}$	E34Δ	2	S2
$_{105}$GG**C GA**T$_{112}$	D36Δ	1	S2
$_{135}$A**AG T**TC$_{142}$	K45Δ F46I	1	S3
$_{168}$CC**C TG**G$_{175}$	W57Δ	1	H2
$_{171}$**TGG**$_{174}$	W57Δ	3	H2
$_{189}$AC**C CT**G$_{196}$	L64Δ	1	Loop H2-H3
$_{192}$C**TG AC**C$_{199}$	L64Δ T65P	2	Loop H2-H3/Cro
$_{198}$T**AC G**GC$_{205}$	Y66Δ G67C	1	Cro
$_{216}$**AGC**$_{220}$	S72Δ	1	H3
$_{219}$**CGC**$_{223}$	R73Δ	1	Loop H3-H4
$_{261}$**GCC**$_{265}$	A87Δ	2	H5
$_{279}$G**TC CA**G$_{286}$	V93Δ Q94E	1	S4
$_{282}$**CAG**$_{286}$	Q94Δ	1	S4
$_{300}$T**TC A**AG$_{307}$	F100Δ K101STOP	1	S4
$_{309}$G**AC G**GC$_{316}$	D103Δ	1	Loop S4-S5
$_{321}$A**AG AC**C$_{328}$	K107Δ	1	S5
$_{330}$G**CC GA**G$_{337}$	A110Δ	3	S5
$_{330}$GC**C GA**G$_{337}$	E111Δ	1	S5
$_{360}$**GTG**$_{364}$	V120Δ	3	S6
$_{360}$G**TG A**AC$_{367}$	V120Δ N121D	1	S6
$_{381}$G**GC A**TC$_{388}$	G127Δ I128V	1	S6
$_{390}$T**TC A**AG$_{397}$	F130Δ K131STOP	1	Loop S6-S7
$_{411}$**CTG**$_{415}$	L137Δ	1	Loop S6-S7
$_{435}$**TAC**$_{439}$	Y145Δ	1	Loop S6-S7
$_{444}$**CAC**$_{448}$	H148Δ	3	S7
$_{450}$G**TC TA**T$_{457}$	V150Δ Y151D	3	S7
$_{450}$GT**C TA**T$_{457}$	Y151Δ	2	S7
$_{486}$**AAG**$_{490}$	K162Δ	1	S8
$_{507}$**CAC**$_{511}$	H169Δ	3	S8
$_{510}$A**AC A**TC$_{516}$	N170Δ	1	S8
$_{540}$**GAC**$_{544}$	D180Δ	1	S9
$_{546}$TA**C CA**G$_{553}$	Y182STOP Q183Δ	2	S9
$_{561}$**CCC**$_{565}$	P187Δ	1	S9
$_{600}$TA**C CT**G$_{607}$	Y200STOP L201Δ	1	S10
$_{609}$AC**C CA**G$_{616}$	Q204Δ	1	S10
$_{615}$TC**C GC**C$_{622}$	A206Δ	1	S10
$_{618}$GC**C CT**G$_{625}$	L207Δ	1	S10
$_{621}$C**TG AG**C$_{628}$	L207Δ S208R	1	S10
$_{654}$AT**G GT**C$_{661}$	M218I V219Δ	1	S11
$_{660}$**CTG**$_{664}$	L220Δ	1	S11
$_{663}$**CTG**$_{667}$	L221Δ	1	S11

[a] Numbers refer to gene sequence numbering for *egfp* developed by [107]
[b] Δ after a residue number signifies that residue has been deleted, protein numbering as per GFP [106]
[c] Secondary structure elements as defined in Fig 4.4, helices (H), strands (S).

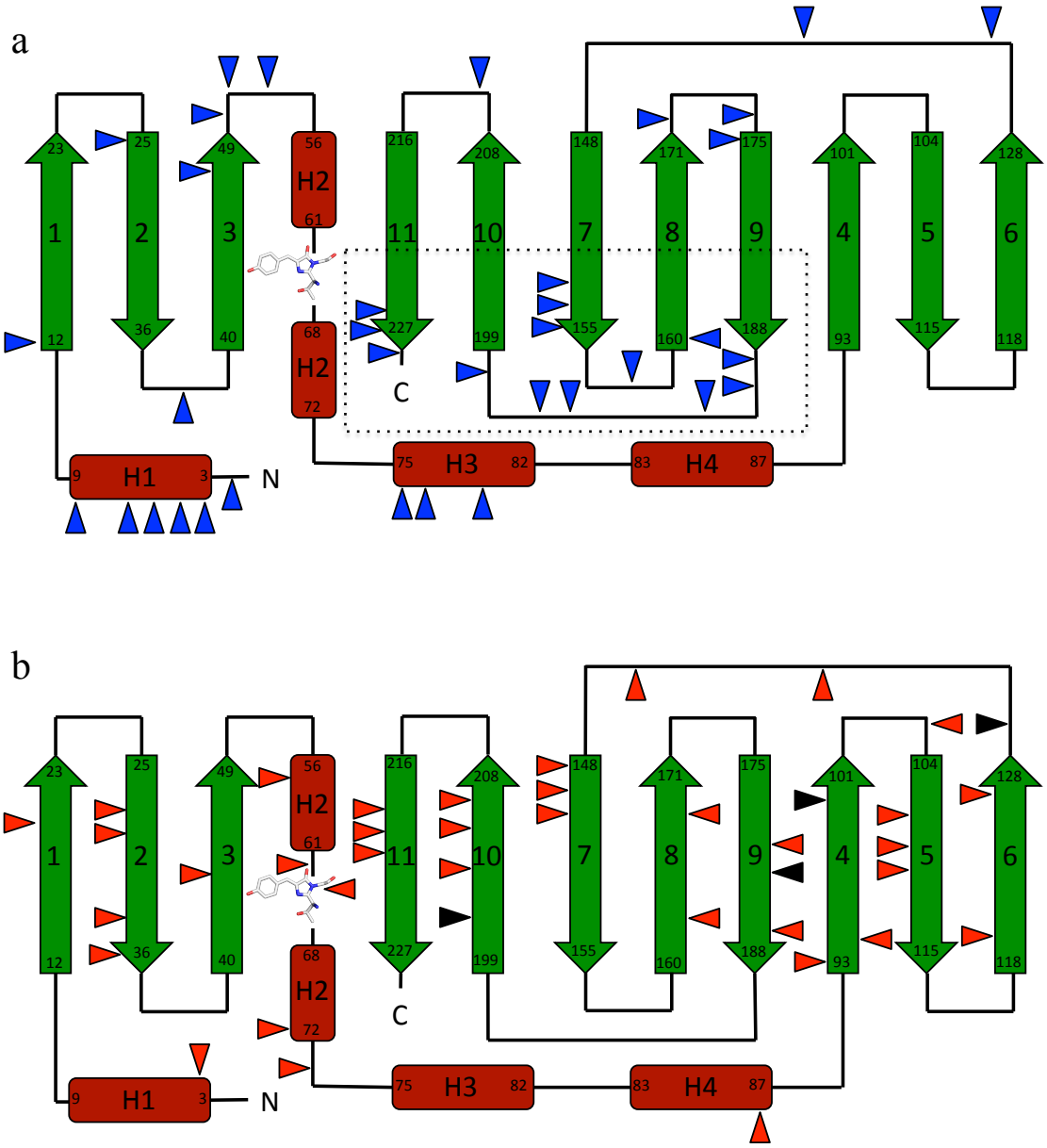


Fig 4.4 Mapping deletion mutations with respect to EGFP secondary structure. The secondary structure arrangement and overall topology of EGFP shows the arrangement of β-strands (green), α-helices (red) and loops (black) **a.** Tolerated single amino acid deletions are indicated by blue triangles with an area particularly tolerant to deletion mutations surrounded by a dotted line. **b.** Non-Tolerated amino acid deletions are indicated by red triangles. Black tringles show the position of premature stop codons introduced due to out of frame triplet nucleotide deletions.

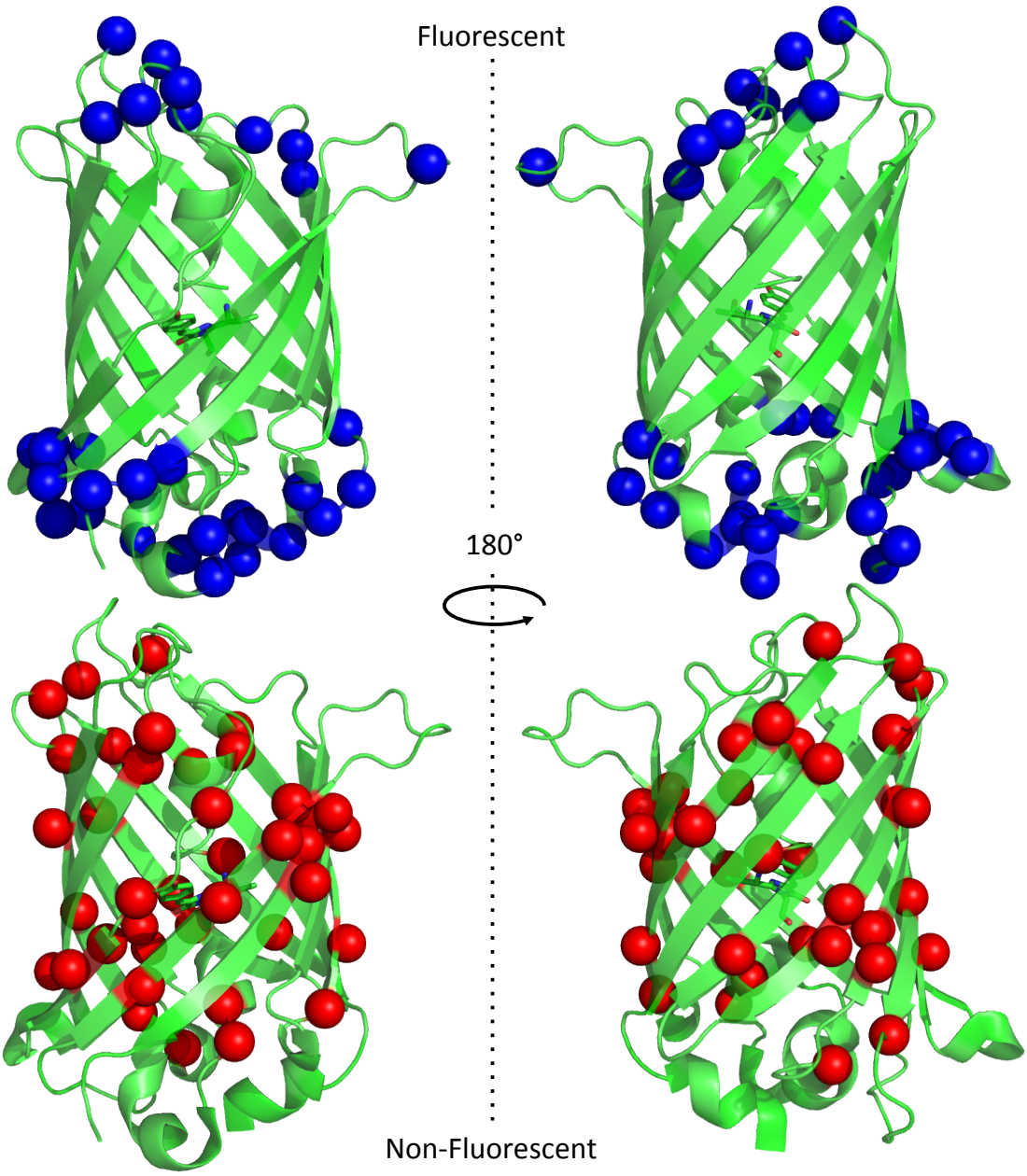


Fig 4.5. Map of single amino acid deletions onto the tertiary structure of EGFP. Cartoon representation of EGFP (green) with tolerated deletions indicated by blue spheres (top) and non-tolerated deletions (bottom) indicated by red spheres.

(Fig 4.4 b). Deletion mutations were observed in four variants in the central α-helix containing the residues that form the chromophore (T65, Y66 and G67), resulting in non-fluorescent variants.

Due to the nature with which MuDel inserts itself randomly throughout a target gene and subsequent removal, by *Mly*I restriction digestion, the triplet nucleotide deletion (TND) is not always in frame. This means that a TND can give rise to the deletion of an amino acid and cause a point mutation (Table 4.3 and 4.3). Out of frame TND can therefore give rise to a stop codon being introduced prematurely into *egfp* depending on the sequence surrounding the TND site (Table 4.4). From the 45 non-tolerated deletions analysed four were due to the introduction of premature stop codons and production of truncated protein (Table 4.4).

Loop regions connecting secondary structures are usually the most tolerant structures to mutagenesis as amino acid deletions are normally accommodated by loop shortening (Chapter 1, Fig 1.1). However, deletions in loops were also found to have a negative impact on EGFP (Table 4.4). L137 and Y145 both reside in the longest loop in EGFP yet their deletion removed the ability of EGFP to confer a fluorescence phenotype on *E. coli*. However, removal of D133 or G138 in the same loop did not result in an inactive protein (Table 4.3). Deletion of D103, also situated in a loop connecting β-strands 4 and 5, was detrimental to fluorescence.

One variant identified had two amino acids deleted (A226 and A227) from the C-terminal end of the last β-strand of EGFP, whilst maintaining fluorescence. This two amino acid deletion could have been produced from a double transposon insertion event or by *Mly*I restriction endonuclease non-specifically removing additional nucleotides upon MuDel removal. The latter is more probable given that extended incubation with *Mly*I or if the enzyme is not fresh can result in star activity and non-specific DNA cleavage.

Observing the tolerated and non-tolerated deletion positions in the tertiary structure of EGFP shows the clustering of tolerated deletions to the two ends of the β-barrel, primarily in loops connecting secondary structures (Fig 4.5)

4.2.3 Fluorescence properties of active EGFP deletion variants

The deletion variants that conferred cellular fluorescence were analysed further to assess if mutations influence the fluorescence properties of EGFP. Variants were expressed in *E.coli* BL21-Gold (DE3) cells (Table 2.1) overnight at 25 °C in M9

minimal media supplemented with 100 µg/ml ampicillin and 150 µM IPTG. Excitation and emission spectra were measured by whole cell fluorescence analysis (Section 2.6.1.1) of culture samples.

All cultures were diluted to an $O.D._{600}$ = 0.1 in TNG (50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 10% (v/v) glycerol) buffer so that any differences in fluorescence intensity between the variants would be a reflection of expression levels or increased brightness (product of extinction coefficient and quantum yield) and not cell density. All of the variants had similar excitation and emission spectra to EGFP, ~488 nm and ~510 nm respectively, but with differing intensities (Fig 4.6 and Table 4.5). In particular two of the single amino acid deletion variants, $EGFP^{D190\Delta}$ and $EGFP^{A227\Delta}$ exhibited increased fluorescence (~1.5 and ~1.8 fold respectively) with respect to EGFP (Fig 4.6 and Table 4.5). $EGFP^{A227\Delta}$ was also identified five times during library screening as colonies expressing this variant observed on a UV transilluminator, had a brighter green colour phenotype with respect to other variants.

Another variant, $EGFP^{G4\Delta}$, was identified 8 times during the library screen due to colonies expressing this variant having a significantly brighter green colour phenotype, when illuminated with a UV-transilluminator, than other variants (Section 4.2.4, Fig 4.9 b). The deleted residue, G4, is distant from the chromophore and only a couple of residues into the first α-helix of EGFP (Fig 4.7). Further analysis and characterization of this variant has been performed in Sections 4.2.4 onwards.

A fourth variant, $EGFP^{K158\Delta}$, was selected for further study because the deletion site is located in a distorted type-I β-turn between two antiparellel β-strands (S7 and S8) (Fig 4.7). Despite the deletion from this secondary structure $EGFP^{K158\Delta}$ maintained similar levels of fluorescence with respect to EGFP (Fig 4.6 and Table 4.4). The three variants identified as having improved fluorescence properties ($EGFP^{G4\Delta}$, $EGFP^{D190\Delta}$ and $EGFP^{A227\Delta}$) and variant $EGFP^{K158\Delta}$ were selected for more detailed expression studies and further characterization.

4.2.4 Protein expression studies

Protein expression studies were performed on EGFP and the selected single amino acid deletion variants to determine if the mutations had affected the levels of expressed protein. The four variants selected were $EGFP^{G4\Delta}$, $EGFP^{D190\Delta}$, $EGFP^{K158\Delta}$ and $EGFP^{A227\Delta}$, with EGFP included as a control. These four variants all have single amino acid deletions situated towards the bottom of the β-barrel (Fig 4.7).

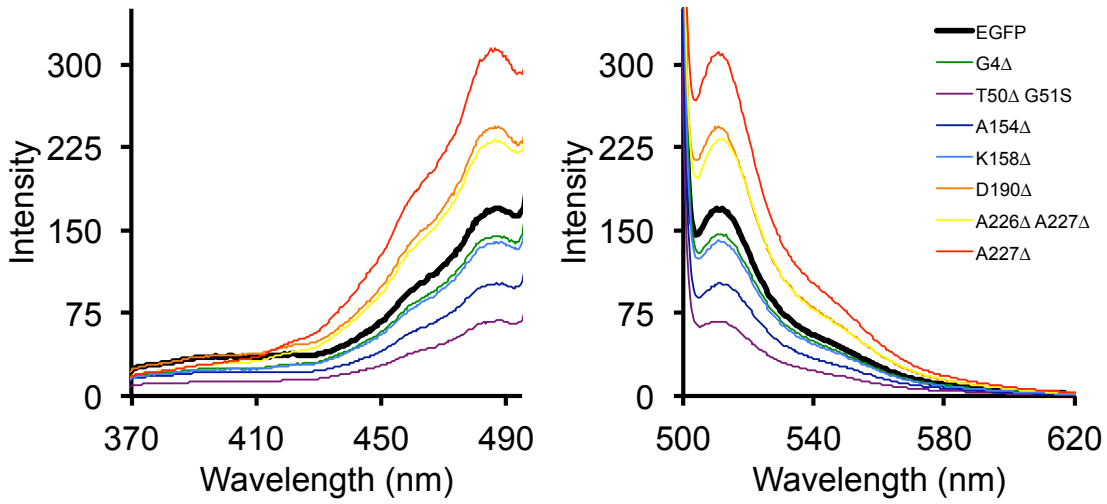


Fig 4.6 Excitation and emission spectra of EGFP and selected EGFPΔ variants. Excitation (left panel) and emission (right panel) spectra were measured by whole cell fluorescence analysis of cultures expressing EGFP and EGFPΔ variants. Cultures were normalized by cell density so that any differences in fluorescence intensity would be attributed to differences in expression levels or altered spectral properties.

Table 4.5 Fluorescent properties of EGFP and EGFPΔ variants

Mutation	λ_{ex} (nm)	λ_{em} (nm)	Fold change in fluorescence[a]
EGFP	488	512	1.0
K3N G4Δ	486	511	0.4
G4Δ	487	511	0.9
E6Δ	487	512	1.1
T9Δ G10R	486	511	0.7
T38Δ	488	511	0.1
C48Δ	488	511	0.2
T50Δ G51S	488	511	0.4
L53Δ	489	509	0.3
P75Δ D76H	487	512	0.2
D76Δ	487	512	1.0
E132Δ	488	510	0.2
A154Δ	487	511	0.6
K158Δ	486	511	0.8
G160Δ	487	511	0.8
G174Δ	487	511	0.2
S175Δ	486	511	0.4
G189Δ	488	510	0.3
D190Δ	487	511	1.4
P211Δ N212H	488	511	0.2
A226Δ A227Δ	487	511	1.4
A227Δ	487	511	1.8
G228Δ	487	510	0.2
M233Δ D234N	486	511	0.5

[a] Fold difference in fluorescence determined with respect to EGFP fluorescence

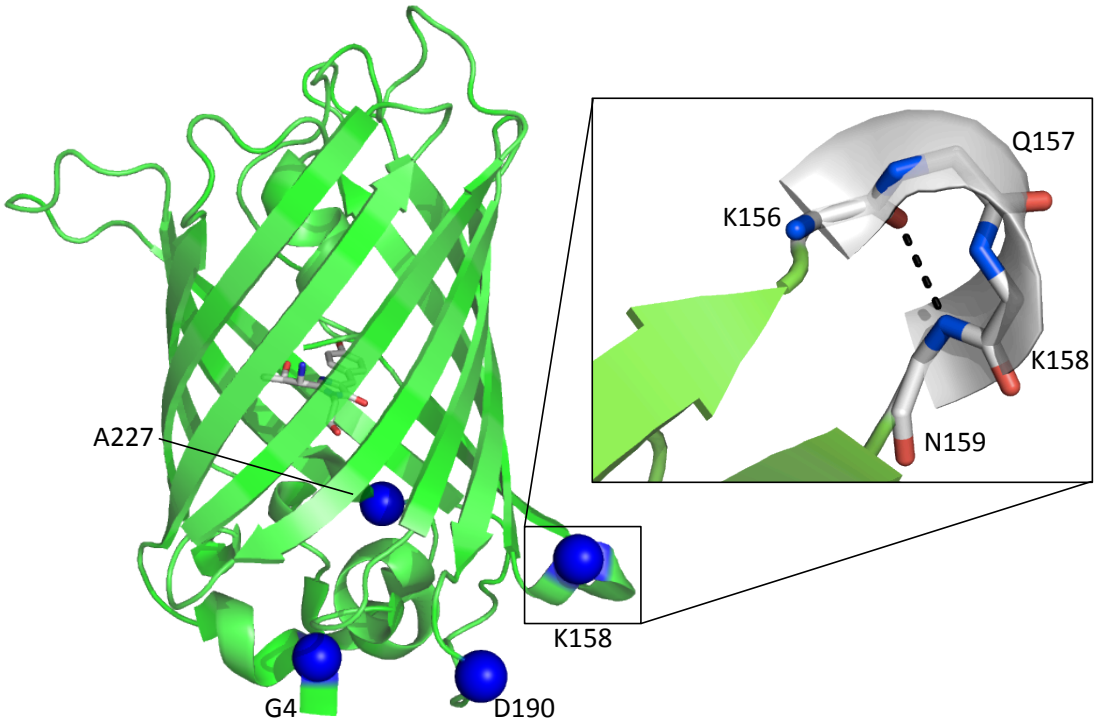


Fig 4.7. Selected amino acid deletion positions in the tertiary structure of EGFP. Cartoon representation of EGFP (green) with amino acid deletion positions of selected variants indicated by blue spheres. **Inset,** Residue K158 is located in a tight β-turn (white cartoon) comprised of just 4 residues (stick representation of backbone atoms only).

Expression studies were performed using *E. coli* BL21-Gold (DE3) cells in LB broth supplemented with 100 μg/ml ampicillin, induced with 150 μM IPTG, at two different temperatures (25 °C or 37 °C). As temperature can affect folding efficiency, sampling two different temperatures would allow production levels of functional protein to be assessed with respect to EGFP, which was constructed to generate more fluorescent protein at 37 °C.

Whole cell fluorescence analysis (Section 2.6.1.1) was performed on cultures grown in triplicate for 18 hr at 37 °C or 28 hr at 25 °C. Mean excitation and emission spectra for each variant, at both temperatures, indicated that all the deletion variants exhibited increased fluorescence with respect to EGFP (Fig 4.8). With respect to EGFP, at 37 °C $EGFP^{K158\Delta}$ fluorescence was only marginally higher (~1.1 fold), $EGFP^{D190\Delta}$ and $EGFP^{G4\Delta}$ fluorescence was ~1.4 or ~1.7 fold higher respectively. $EGFP^{A227\Delta}$ was the best performer with ~ 2.6 fold greater fluorescence intensity than EGFP (Fig 4.8 a).

At 25 °C cultures producing $EGFP^{K158\Delta}$ and $EGFP^{D190\Delta}$ had similar fluorescence intensities to cultures grown at 37 °C with increased fluorescence intensity of ~1.2-fold with respect to EGFP (Fig 4.8 b). In contrast, the increase in fluorescence intensity for $EGFP^{A227\Delta}$ seen at 37°C was reduced at 25 °C to ~1.3-fold with respect to EGFP (Fig 4.8 b). This 2-fold decrease in fluorescence intensity relative to EGFP at the different temperatures could be a consequence of reduced protein expression levels at lower temperatures or the increase in EGFP folding efficiency. $EGFP^{G4\Delta}$ however exhibited almost a 4-fold increase in fluorescence intensity compared to EGFP when grown at 25 °C (Fig 4.8 b).

To investigate the temperature relationship between protein production and fluorescence for $EGFP^{G4\Delta}$ SDS-PAGE and western blot analysis was performed on whole cell, soluble and insoluble fractions from culture grown at 37°C (Fig 4.9 a). Cultures expressing the four different variants were harvested by centrifugation (Section 2.5.2) and the cell pellets resuspended to an apparent $O.D._{600}$ = 10 to normalize the number of cells across all the samples. Any observed increase in protein band density would therefore be a reflection of protein quantity and not the number of cells used to prepare the sample.

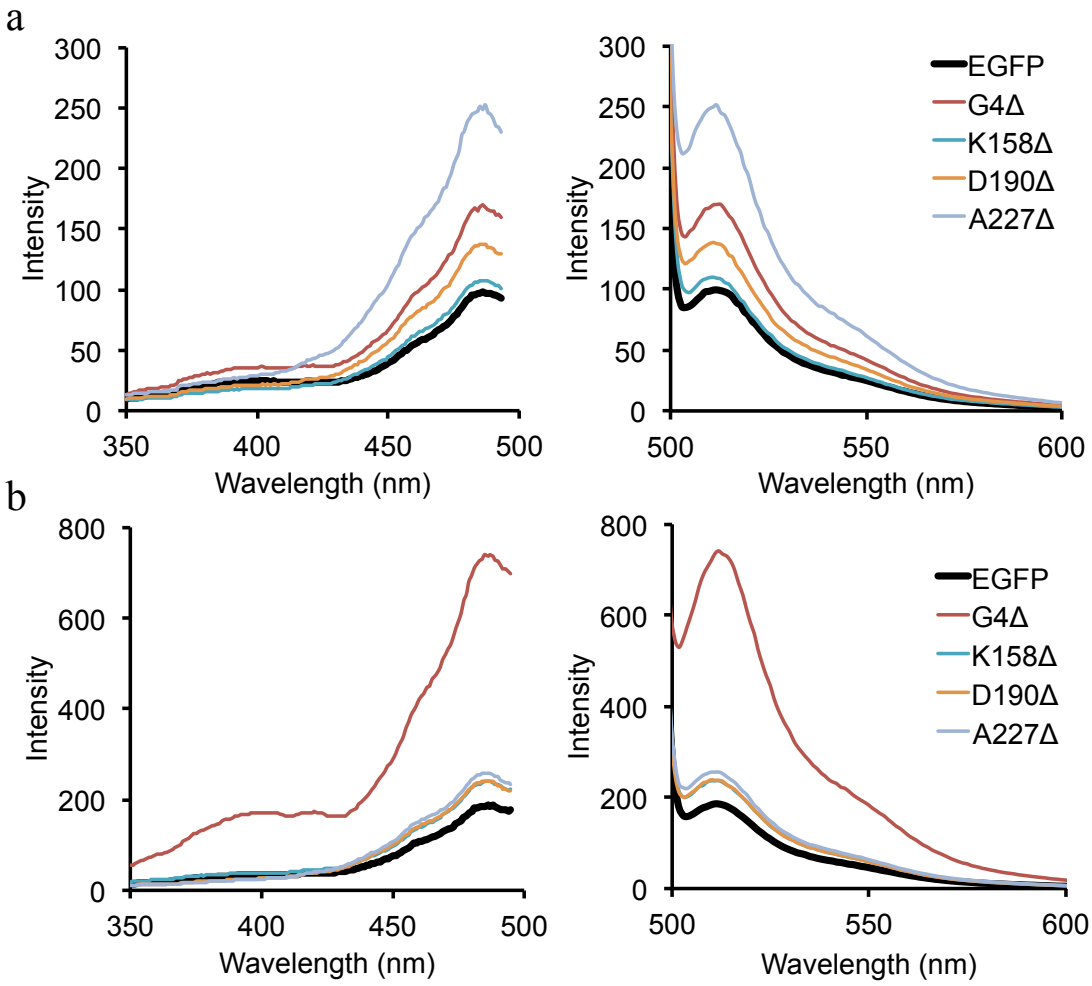


Fig 4.8. Fluorescence intensity of whole cell samples. Excitation (left panel) and emission (right panel) spectra measured by whole cell fluorescence analysis using cultures expressing EGFP and selected EGFPΔ variants (as annotated on spectra) at **a,** 37 °C or **b,** 25 °C

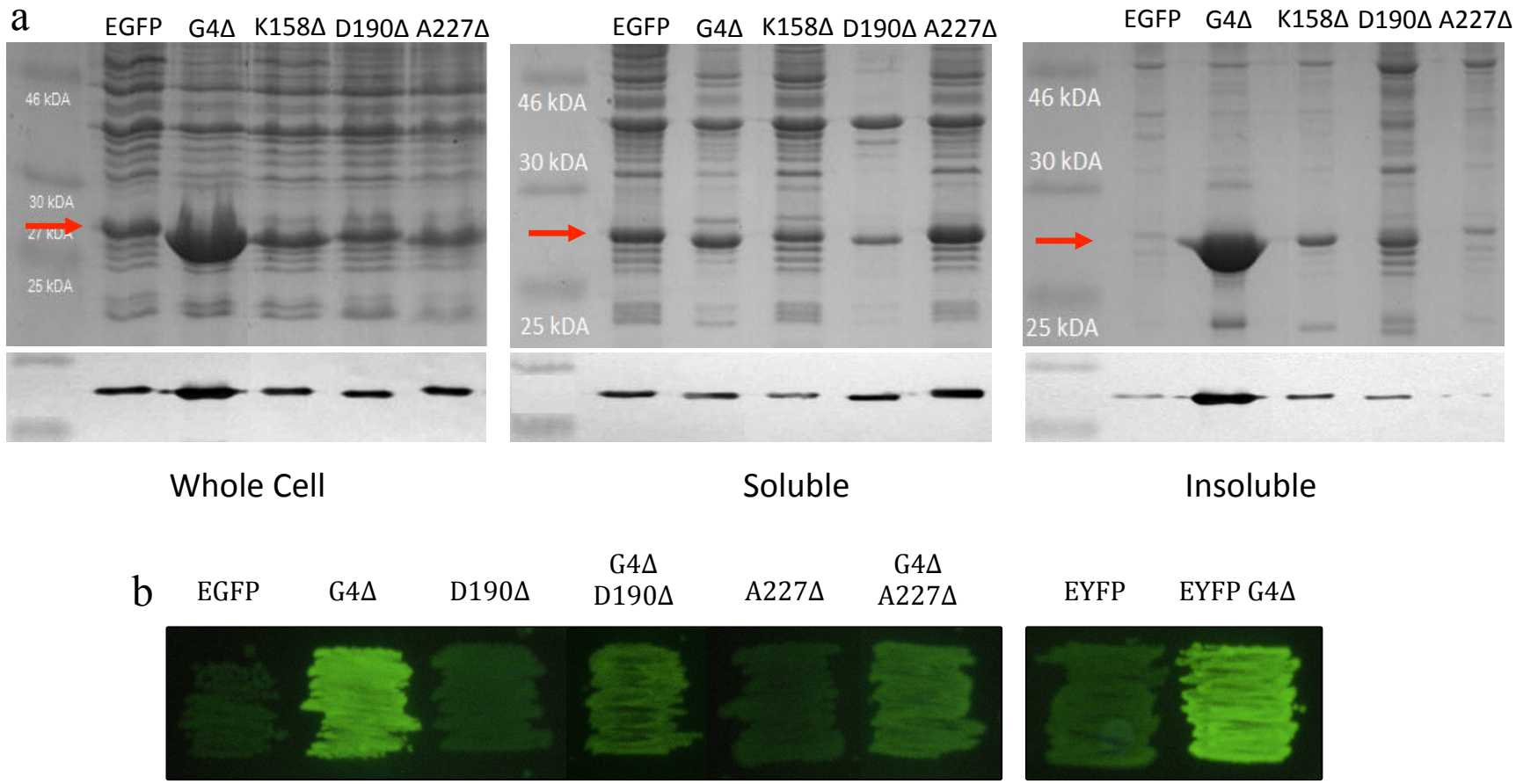


Fig 4.9 Analysis of cellular production of EGFP deletion variants. a. SDS-PAGE (top panels) and western blot (bottom panels) analysis of whole cell, soluble and insoluble fractions from a 37 °C expression of EGFP and EGFPΔ variants (red arrow) as indicated in the figure. Antibodies specific for EGFP were used for western blot analysis. **b.** Cellular fluorescence of colonies expressing EGFP, EGFP single and double deletion variants, EYFP and EYFP$^{G4\Delta}$ streaked out on LB agar plates supplemented 150 μM IPTG illuminated with a UV transilluminator.

The SDS-PAGE analysis showed that EGFP$^{G4\Delta}$ was produced to a much higher level than EGFP and the other variants, indicated by greater protein band density on the gel for the whole cell sample (Fig 4.9 a). The majority of the protein appears to be insoluble at 37 °C as can be seen in the gel for the insoluble fraction (Fig 4.9 a) and explains the decreased fluorescence observed in cultures expressing G4Δ at 37 °C.

The SDS-PAGE analysis also showed that EGFP and EGFP$^{A227\Delta}$ were produced to similar levels, indicated by similar protein band density on the gel for the whole cell sample (Fig 4.9 a). However, the SDS-PAGE and Western blot analysis of soluble and insoluble fractions for EGFP and EGFP$^{A227\Delta}$ showed that there was more EGFP$^{A227\Delta}$ in the soluble fraction and less in the insoluble fraction with respect to EGFP (Fig 4.9 a). Although the EGFP$^{K158\Delta}$ deletion is located in a tight β-turn in the tertiary structure of EGFP (Fig 4.7) it didn't appear, from the expression studies, to benefit protein production (Fig 4.9 a) or fluorescence intensity (Fig 4.8), therefore no further analysis was carried out on this variant.

Given that EGFP$^{G4\Delta}$ gets produced to such high levels, based on increased band intensity (Fig 4.9 a) on the SDS-PAGE gel, temperatures may be influencing the amount of correctly folded soluble G4Δ and would explain the four fold increase in fluorescence with respect to EGFP seen in cultures grown at 25 °C. However, SDS-PAGE and Western blot analysis of culture samples expressing EGFP$^{G4\Delta}$ at 25 °C would be required to confirm this.

As mentioned previously (see 4.2.4) colonies grown on LB agar expressing EGFP$^{G4\Delta}$ had a much brighter green colour phenotype than any other variant (Fig 4.9 b). This contradicts the previous findings that expression of EGFP$^{G4\Delta}$ in liquid cultures at 37 °C produced ~1.7-fold increase in fluorescence with respect to EGFP.

To investigate if the property of increased cellular fluorescence observed for EGFP$^{G4\Delta}$ is a transferrable trait, double amino acid deletion mutants EGFP$^{G4\Delta D190\Delta}$ and EGFP$^{G4\Delta A227\Delta}$ were constructed (Section 2.4.2). The G4Δ mutation was also engineered into enhanced yellow fluorescent protein (EYFP) to assess if the mutation was EGFP specific or transferrable to other colour variant fluorescent proteins (Section 2.4.1).

E.coli BL21-Gold (DE3) cells were transformed with relevant plasmids containing the genes for different deletion variants and were grown on LB agar plates overnight at 37 °C. Single colonies were selected and resuspended in 200 µl of LB broth and incubated in a shaking incubator for 2 hrs. The cultures were then streaked

onto LB agar plates supplemented 150 μM IPTG and incubated overnight at 37 °C. Coloured phenotypes were noted after illumination of the streaked out cultures with a UV-transilluminator (Fig 4.9 b).

All colonies producing variants carrying the G4Δ mutation resulted in a brighter colour phenotype than variants without the mutation. Although the double amino acid deletion variants are not as fluorescent as $EGFP^{G4\Delta}$ variant they are more fluorescent than their single amino acid deletion counterparts (Fig 4.9 b). The G4Δ mutation is also transferrable to EYFP, increasing the intensity of the cellular colour phenotype (Fig 4.9 b). The transferability of the G4Δ mutation to other fluorescent proteins highlights this mutation as being beneficial. In order to determine how the mutations are affecting EGFP fluorescence, EGFP, $EGFP^{G4\Delta}$, $EGFP^{D190\Delta}$ and $EGFP^{A227\Delta}$ were expressed (Section 2.5.3) and purified (Section 2.5.8) for further characterization.

4.2.5 Fluorescent characterisation of EGFP and EGFPΔ variants.

To determine if the increased fluorescence of $EGFP^{G4\Delta}$, $EGFP^{D190\Delta}$ and $EGFP^{A227\Delta}$ with respect to EGFP was due to a difference in the production level of soluble protein or if the proteins fluorescent characteristics had been altered, a detailed characterization of their fluorescence properties was performed. All of the deletion variants have similar spectral characteristics to EGFP with λ_{em} and λ_{ex} almost identical (Table 4.6).

Minor changes in extinction coefficient or quantum yield were observed, thus, the brightness of the three variants are comparable to that of EGFP (Table 4.6). The fluorescence lifetimes for the deletion variants are also comparable to that of EGFP indicating the mutations have not substantially effected the surrounding environment of the chromophore (Table 4.6, Fig 4.10). This confirmed that the effect of altered fluorescence intensity observed in the expression studies is due to increased protein production levels and solubility due to efficient folding and not due to altered spectral characteristics.

Table 4.6. Spectral characteristics of EGFP and EGFPΔ variants

Mutation	λ_{ex} (nm)[a]	λ_{em} (nm)[a]	ε ($M^{-1}cm^{-1}$)[b]	ϕ[c]	Brightness[d] ($M^{-1}cm^{-1}$)	τ (ns)[e]
EGFP	488	511	55000	0.60	33000	2.54±0.04
G4	487	512	53070	0.59	31300	2.64±0.05
D190	486	510	53430	0.58	30990	2.56±0.05
A227	487	511	51850	0.61	31630	2.44±0.04

[a] λ_{ex} and λ_{em} determined from mean fluorescence spectra (Fig 4.8)
[b] Extinction coefficient determined from single absorbance measurement
[c] Quantum yield determined from integrated fluorescence emission against a fluorescein standard
[d] Brightness = extinction coefficient x quantum yield
[e] Fluorescence lifetimes are mean values with errors calculated from the standard deviation of 3 measurements

Table 4.7 Size exclusion chromatography analysis

Mutation	Concentration (μM)	Elution volume (ml)[a]	Calculated Mw (Da)[b]	Average Mw (Da)[c]	Estimated Mw (Da)[d]
EGFP	10	11.50	24600	25600	26941
	25	11.46	25100		
	50	11.42	25800		
	100	11.35	26800		
G4	10	11.50	24600	25700	26884
	25	11.45	25300		
	50	11.42	25800		
	100	11.34	27000		
D190	10	11.59	23300	24200	26826
	25	11.56	23800		
	50	11.53	24300		
	100	11.45	25200		
A227	10	11.60	23300	23400	26870
	25	11.63	22900		
	50	11.59	23500		
	100	11.57	23700		

[a]Elution volumes determined from peak absorbance at 488 nm (Fig 4.11)
[b]Molecular weights calculated using standard curve (Appendix A)
[c]Average calculated molecular weights over the different concentrations
[d]Estimated molecular weight from primary sequence

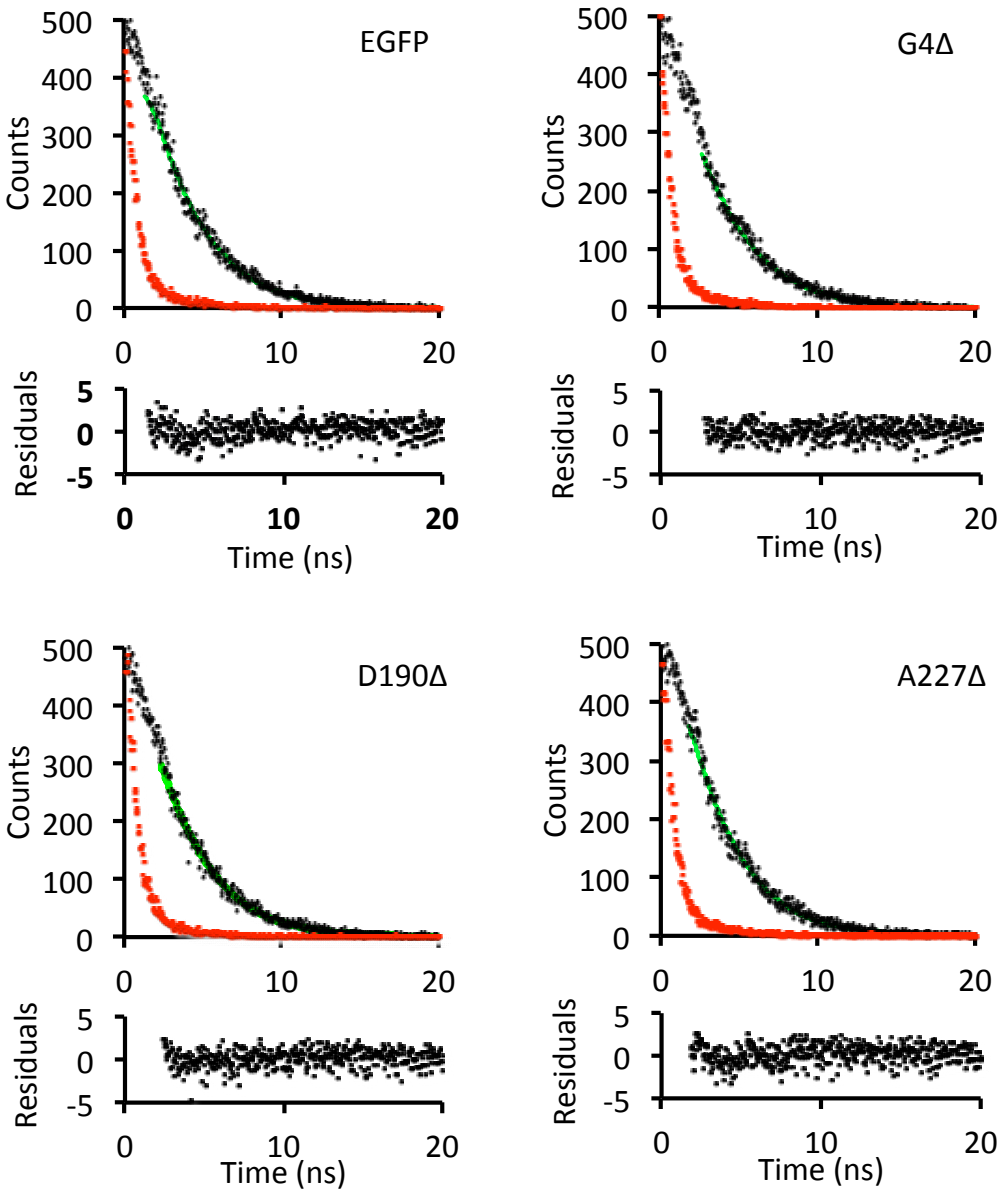


Fig 4.10. Fluorescence lifetime analysis of EGFP deletion variants. Photon counts (black dots) measured over 20 ns are fit to single exponential decay functions (green line) with the instrument response functions (IRF) (red dots). Residuals for the single exponential decay fits to the data are plotted below each of the decay curves for EGFP and EGFPΔ variants as indicated in the figure.

4.2.6 Biophysical characterization of EGFP and EGFPΔ variants

4.2.6.1 Analytical size exclusion chromatography

The propensity for fluorescent proteins (FPs) to dimerize is an unwanted attribute if they are to be used *in vivo* as fusion partners, as dimerization can alter spectral properties [62, 108], could promote protein aggregation [109], affect results from protein-protein interaction experiments [63] and hinder their use as Forster resonance energy transfer (FRET) partners [110]. With this in mind it was important to investigate the oligomeric states of EGFP and the EGFPΔ variants. This was performed by analytical size exclusion chromatography (SEC) using a Superdex^{TM} 75 gel filtration column (Section 2.6.2.1). A standard curve for the relationship between molecular weight and elution volume from the column was determined using BioRad gel filtration standards (Appendix A), so that the elution volume for EGFP and EGFPΔ variants, at different concentrations, could be related to the molecular weight of that species under native conditions.

EGFP and EGFPΔ samples were applied to the column at 10, 25, 50 or 100 μM concentrations. The elution of proteins from the column was monitored by absorbance at 488 nm and the elution volumes (Table 4.7) determined from the absorbance peak (Fig 4.11). EGFP, EGFP$^{G4\Delta}$ and EGFP$^{D190\Delta}$ all exhibited a decrease in elution volume (~0.14 - 0.16 ml), with increasing protein concentration (from 10 – 100 μM) (Table 4.7), relating to an increase in calculated molecular weight by ~10% (~24.5 - ~27 kDa).

The change in elution volume with increasing protein concentration suggests there may be a weak tendency for EGFP, EGFP$^{G4\Delta}$ and EGFP$^{D190\Delta}$ to dimerize. If there was a strong tendency for EGFP or the EGFPΔ variants to dimerize an elution volume of ~10.2 ml would be expected. The shape of the elution profiles for EGFP, EGFP$^{G4\Delta}$ and EGFP$^{D190\Delta}$ exhibit slight asymmetry indicating that any dimer interactions possibly taking place have rapid association/dissociation rate constants, as slow rates would present as two defined elution peaks in the elution profile.

EGFP$^{A227\Delta}$, in contrast to EGFP and the other deletion variants, showed almost no change in elution volume (Table 4.7) with increasing protein concentration. The elution volume for EGFP$^{A227\Delta}$ (~11.6 ml) inferred a calculated molecular weight of ~23.4 kDa (Table 4.7), almost 3.4 kDa smaller than the theoretical molecular weight of 26.8 kDa calculated from its amino acid sequence. The elution volumes for EGFP$^{D190\Delta}$

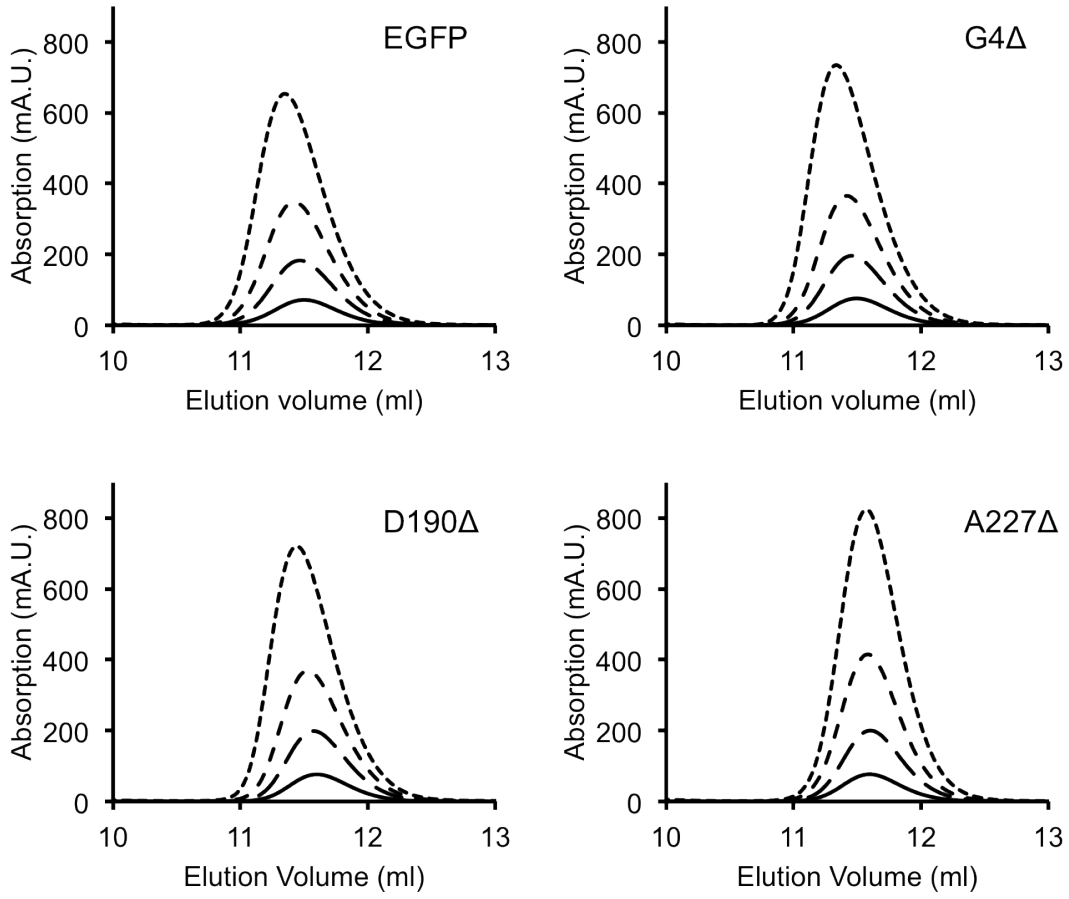


Fig 4.11 Size exclusion chromatography of EGFP or EGFPΔ variants. Samples of variants (as indicated in figure) were applied to a Superdex^{TM} 75 gel filtration column and the elution of the protein samples monitored by absorbance at 488 nm. Protein concentrations of 10 μM (solid black line), 25 μM (long dashed line), 50 μM (medium dashed line) or 100 μM (short dashed line) were applied to the column.

(~11.60 - ~11.45 ml) also equate to smaller calculated molecular weights (~23.3 - ~25.2 kDa) than the theoretical molecular weight calculated from their primary sequence (~26.8 kDa). Although estimated molecular weights can be calculated from the elution volume, analytical SEC is actually a measure of hydrodynamic volume. The difference between the theoretical and calculated molecular weights could therefore be a reflection of $EGFP^{D190\Delta}$ and $EGFP^{A227\Delta}$ being more compact than EGFP or $EGFP^{G4\Delta}$.

4.2.6.2 Guanidinium chloride induced equilibrium unfolding

To assess the deletion mutations effect on EGFP stability, guanidinium chloride (GdmCl) induced equilibrium unfolding experiments were performed (Fig 4.12). The chromophore of EGFP is very sensitive to its environment with protein unfolding and exposure to aqueous solvent significantly reducing fluorescence intensity. Thus, the chromophore is a sensitive probe to monitor protein unfolding at different concentrations of GdmCl over a wide range of equilibration times (2.5 – 250 hr).

Unlike usual GdmCl denaturation studies of small monomeric proteins, which reach equilibrium quickly, EGFP and EGFPΔ variants take up to two weeks to approach equilibrium (Fig 4.12 a-e). Although unusual, this phenomenon has been seen in studies of many fluorescent proteins [71, 111-113]. The very slow kinetics to equilibrium unfolding in GdmCl is thought to be attributed to high folding/unfolding energy barriers

It therefore must be emphasized that any analysis of the GdmCl unfolding data does not assume that the system has reached full equilibrium but is approaching equilibrium. Due to a limited number of data points above concentrations of 3.5 M GdmCl it was not possible to fit the measured unfolding data up to 22 hrs of incubation to either a two-state or three-state model. Therefore the unfolding data for EGFP and EGFPΔ acquired after 2.5, 5, 22, 48, 120 and 250 hr incubations in GdmCl were fit to an asymmetric five-parameter Equation (GraphPad Prism) to estimate apparent [GdmCl] at which 50% of the protein sample is in the native state and 50% is in the denatured state ($[GdmCl]_{50\%}^{app}$]) (Fig 4.12 a-d).

The $[GdmCl]_{50\%}^{app}$ values were plot against incubation time to estimate equilibrium kinetics (Fig 4.12 e). After fitting the data to a single exponential decay (Section 2.6.2.4, Equation 17) it was apparent that the GdmCl unfolding of the EGFP and EGFPΔ variants had approached equilibrium (Fig 4.12 e) by 250 hrs.

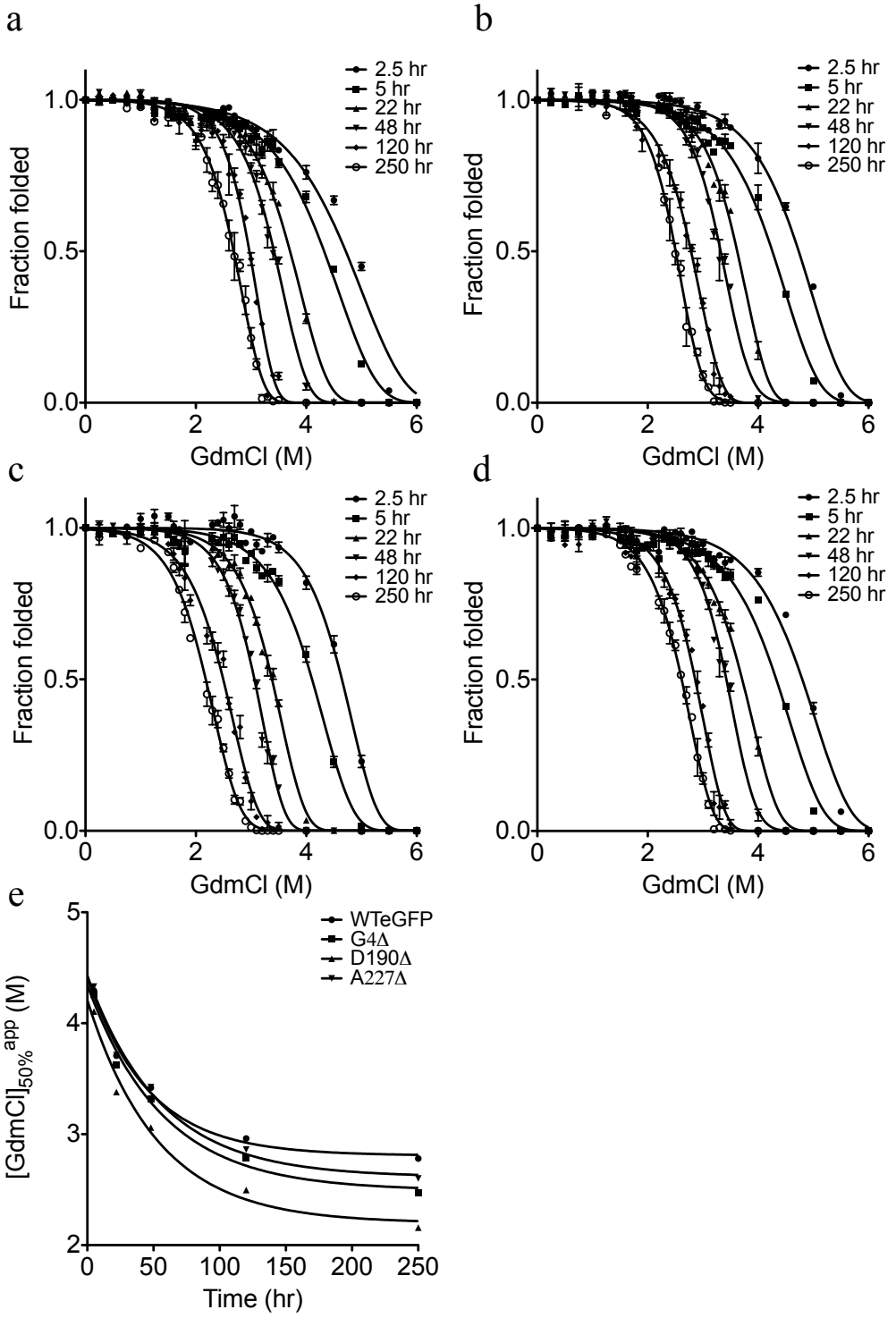


Fig 4.12 Guanidinium chloride induced equilibrium unfolding and equilibrium kinetics. Fluorescence emission at 520 nm after excitation at 480 nm was monitored for **a,** EGFP, **b,** G4Δ, **c,** D190Δ and **d,** A227Δ over 250 hrs (as indicated in the figures) and data were fit to an asymmetric 5 parameter Equation (GraphPad Prism). **e,** Apparent $[GdmCl]_{50\%}$ values (the [GdmCl] at which 50% of the samples are in the native and 50% in the denatured states) were plot against time and fit to single exponential decay curves to assure equilibrium had been reached before calculating free energies of denaturation ($\Delta G_{N-D}^{H_2O}$).

The unfolding data for EGFP and EGFPΔ variants after 250 hrs incubation in GdmCl were further analysed by a two-state unfolding model (Fig 4.13 a) (Section 2.6.2.2). Previous equilibrium unfolding studies of fluorescent proteins has indicated that an unfolding intermediate exists [113, 114] and the equilibrium unfolding data for GFPs is best fit to a three-state model. Despite this, a three-state model did not converge on the data obtained here possibly due to an intermediate state being indistinguishable from the denatured state by monitoring fluorescence under the conditions used (pH 8.0 and 37°C) for denaturation. Therefore given that the data collected here did fit well to a two state model (R^2>0.98) (Fig 4.13 a), thermodynamic parameters determined from this fit could be used for a direct comparison between samples.

Given that there is a linear relationship between ΔG^0_{N-D} and [GdmCl] (Fig 4.13 b) the free energy of denaturation in water $\left(\Delta G^{H_2O}_{N-D}\right)$ and the dependence of ΔG^0_{N-D} on [GdmCl] (m-value) could be inferred from Equation 16 (Section 2.6.2.2) (Table 4.5). The free energy of denaturation calculated for EGFP (5.28 kcal mol^{-1}) agrees closely to previous equilibrium unfolding experiments performed (5.16 kcal mol^{-1}) [111]. The G4Δ mutation conferred increased stability to EGFP with a $\Delta G^{H_2O}_{N-D}$ of 6.46 kcal mol^{-1} even though the $[GdmCl]_{50\%}^{app}$ is marginally lower (~ 0.18 M) (Table 4.8).

Although the $[GdmCl]_{50\%}^{app}$ for $EGFP^{G4\Delta}$ is lower than EGFP the ΔG^0_{N-D} dependency on [GdmCl] (m-value) is greater. The m-value for EGFP was calculated to be 1.95 kcal mol^{-1} M^{-1} whilst for $EGFP^{G4\Delta}$ the m-value is 2.61 kcal mol^{-1} M^{-1}. This results in an increased $\Delta G^{H_2O}_{N-D}$ for $EGFP^{G4\Delta}$. There are several possible causes for this: 1, the denatured protein is less compact and has less structure than that of denatured EGFP or 2, an increase in m-value could reflect a change in the degree to which an intermediate state is populated during the unfolding. The A227Δ mutation also increased overall stability of EGFP from 5.28 kcal mol^{-1} to 5.81 kcal mol^{-1}.

Although the D190Δ mutation appears to decrease the $[GdmCl]_{50\%}^{app}$ quite substantially (~ 0.52 M) the increased cooperativity of unfolding (m-value of 2.23 kcal mol^{-1} M^{-1}) with respect to EGFP results in this variant having the same overall stability as EGFP (within error) (Table 4.8).

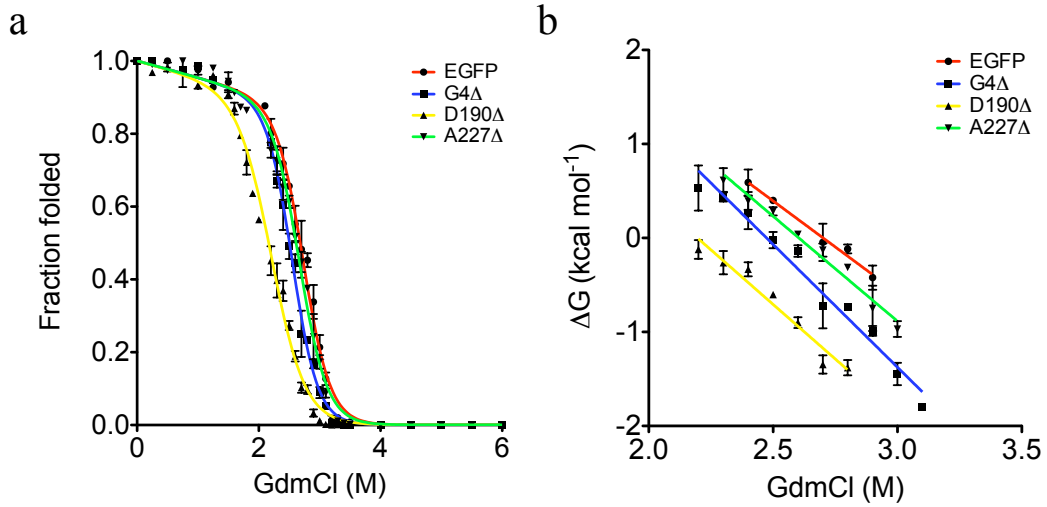


Fig 4.13. Equilibrium unfolding and linear dependence of ΔG on [GdmCl]. a, Equilibrium unfolding data, after 250 hr incubation, were fit to a 2 state model taking into account sloping baselines inherent to spectroscopic measurements of protein samples in GdmCl. **b,** The linear dependence of ΔG on [GdmCl] (m-value) was used to calculate free energies of denaturation ($\Delta G_{N-D}^{H_2O}$).

Table 4.8. Equilibrium unfolding, unfolding and refolding kinetic parameters and melting temperature.

Variant	$[GdmCl]_{50\%}$ (M)[a]	$\Delta G_{N-D}^{H_2O}$ (kcal mol^{-1})[b]	m-value (kcal mol^{-1} M^{-1})[c]	k_U (min^{-1})[d]	k_{fast} (10^{-2} s^{-1})[e]	k_{slow} (10^{-2} s^{-1})[e]	Tm (°C)
EGFP	2.74	5.28±0.36	1.95±0.13	2.09±0.00	4.32±0.08	1.00±0.02	84
G4Δ	2.56	6.46±0.41	2.61±0.15	2.15±0.01	3.78±0.25	1.15±0.05	83
D190Δ	2.22	5.10±0.57	2.32±0.23	1.81±0.02	2.96±0.06	0.92±0.17	82
A227Δ	2.68	5.81±0.37	2.23±0.13	4.59±0.01	3.93±0.15	0.89±0.04	84

[a]Concentration of GdmCl at which 50% of the protein sample is in the native and denatured state, determined from a 2 state model (Section 2.6.2.2).
[b]Free energy of denaturation ($\Delta G_{N-D}^{0} = \Delta G_{N-D}^{H_2O} - m[GdmCl]$)
[c]Measure of dependence of ΔG on denaturant concentration determined from the slope of the plots in Fig 4.13 b.
[d] Rate constant from single exponential fit of unfolding progress curves (Fig 4.14 a)
[e] Rate constants from two exponential fits of refolding progress curves (Fig 4.14 b)

However, great care must be taken when interpreting these results. The presence of an unfolding intermediate and the degree to which it is populated can result in a decreased m-value from data analysis with a 2-state model, underestimating the calculated $\Delta G_{N-D}^{H_2O}$ value. Comparison of the $[GdmCl]^{50\%}$ values may be a more reliable measure of stability in this case.

4.2.6.3 Unfolding and refolding kinetics.

None of the single amino acid deletion mutations have reduced overall stability of EGFP (Table 4.8). In the case of G4Δ and A227Δ mutation, stability has been improved. To complement the equilibrium studies, the rate at which deletion mutation variants unfold and refold was investigated.

Fluorescence was monitored for EGFP and the three deletion variants after rapid dilution of native protein into 6.8 M GdmCl (unfolding) or fully denatured protein into fresh buffer (refolding) (Fig 4.14). The data was fit to a single or double exponential function for unfolding and refolding experiments respectively (Section 2.6.2.4, Equations 17 and 18 respectively) (Fig 4.14).

Unfolding kinetics for fluorescent proteins are very slow (min – hr) in comparison to other small globular proteins, therefore it is possible to accurately measure and follow the unfolding of fluorescent proteins after rapid dilution into GdmCl. The unfolding rates of EGFP and $EGFP^{G4\Delta}$ were similar (~2 min^{-1}) (Table 4.8) and agreed with previous findings that EGFP becomes fully unfolded after ~2 mins in 6.2 - 6.5M GdmCl [111]. The unfolding rate for $EGFP^{D190\Delta}$ was only marginally decreased with respect to EGFP, however, the unfolding rate of A227Δ was ~4.5 min^{-1}, 2 fold higher than that of EGFP.

Refolding kinetics of denatured EGFP or EGFPΔ variants was performed by rapid dilution into fresh TNG buffer to a final concentration of 100 μM and 0.68 M GdmCl. There was a marked difference in total amount of fluorescence recovered (Fig 4.14). Both EGFP and $EGFP^{D190\Delta}$ were similar, recovering 77% and 79%, respectively, of their initial fluorescence. $EGFP^{A227\Delta}$ was slightly higher, with 85% of its initial fluorescence recovered. However, $EGFP^{G4\Delta}$ recovered up to 100% of its initial fluorescence before denaturation, suggesting that the full population of protein molecules regained their native structure.

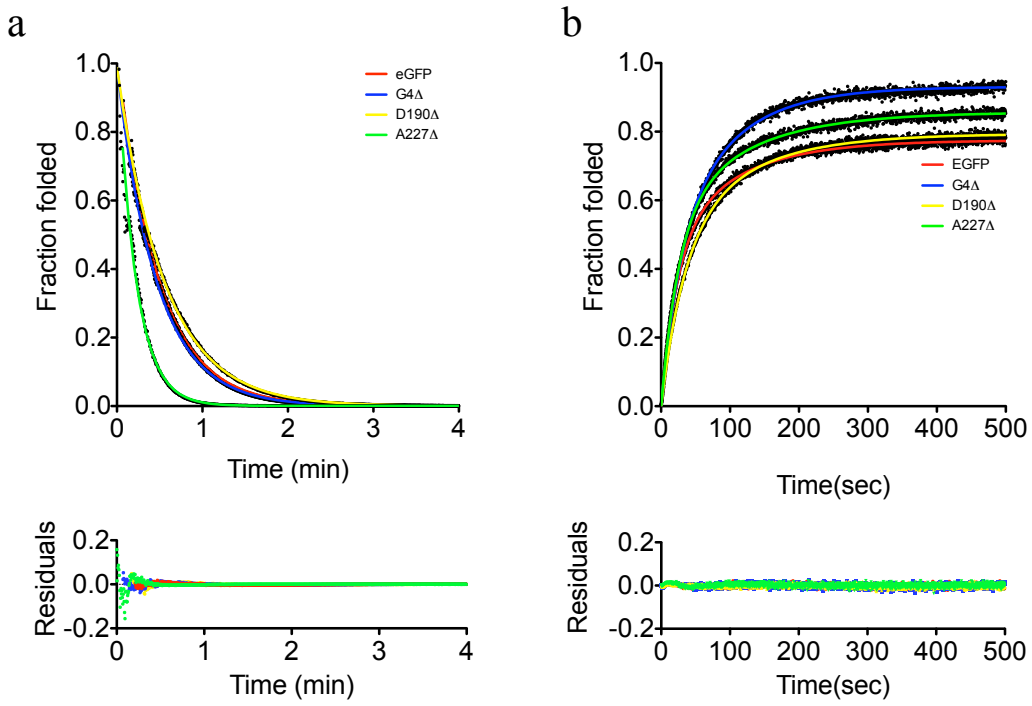


Fig 4.14. Unfolding and refolding kinetics of EGFP and the deletion variants. a, Unfolding kinetics. Rate constants were determined from a single exponential decay function to fluorescence data (excitation at 488 nm, emission at 510 nm), monitored over time, after dilution of EGFP or EGFPΔ variants (as indicated in figure) into 6.8M GdmCl. **b,** Refolding kinetics. Rate constants were determined from double exponential functions fit to fluorescence data monitored over time after dilution of fully unfolded EGFP or EGFPΔ variants into fresh TNG buffer. Residuals for curves fit to the data are shown below their respective graph

The refolding kinetics of EGFP, EGFP$^{G4\Delta}$, EGFP$^{D190\Delta}$ and EGFP$^{A227\Delta}$ show an initial fast phase with rate constants (k_{fast}) of ~4.3 x 10^{-2} s^{-1}, ~3.7 x 10^{-2} s^{-1}, ~2.9 x 10^{-2} s^{-1} and 3.9 x 10^{-2} s^{-1} respectively (Table 4.8). The fast initial phase was followed by a slow refolding phase with rate constants (k_{slow}) of ~1 x 10^{-2} s^{-1}, ~1.1 x 10^{-2} s^{-1}, ~0.9 x 10^{-2} s^{-1} and ~0.9 x 10^{-2} s^{-1} respectively (Table 4.8). *Cis/trans* isomerization has been shown to be a rate-limiting step in protein folding [115, 116] and is thought to be the reason for the slow refolding phase.

From the refolding kinetics generated here, it appears that all the deletion mutations have decreased the fast rate constants (k_{fast}). The G4Δ mutation appears to increase the *cis/trans* isomerization rate as is evident from an increase in the slow rate constant by up to 20% with respect to EGFP (Table 4.8), although this is still only a minor change.

4.2.6.4 Thermal denaturation

Temperature plays a major role during protein folding with lower temperatures promoting correctly folded proteins; given that ΔG is related to entropic considerations and temperature. Many fluorescent proteins have been shown to be thermosensitive (Section 4.2.4) [117, 118], with respect to folding, quite often resulting in misfolding and aggregation above temperatures of ~ 37°C. However, once folded the β-barrel structure of fluorescent proteins is remarkably resistant to thermal denaturation [119, 120], up to temperatures of ~85°C. To assess the effect of the deletion mutations on the thermostability of EGFP, protein samples were slowly heated (1°C/min) from room temperature (25°C) to 98°C whilst monitoring fluorescence using a qPCR thermocycler (Section 2.6.2.5) (Fig 4.15).

Apparent melting temperatures (T_m^{app}) were determined from the fluorescence data and show that the deletion variants were very similar to that of EGFP, with T_m^{app} of 82-84°C (Table 4.8) and therefore retain the high thermal stability of EGFP. However, the thermal denaturation of EGFP and the EGFPΔ variants was irreversible (data not shown) therefore implying the potential formation of aggregates making it difficult to generate accurate thermodynamic data. This can result in decreased T_m^{app} values when calculated using the method described here.

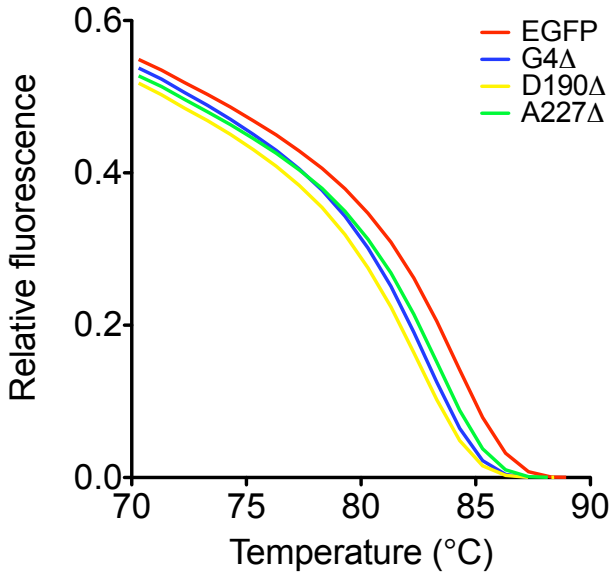


Fig 4.15. Thermal denaturation of EGFP and EGFPΔ variants. Fluorescence emission was monitored during temperature ramping from 25-98 °C at 1 °C/min using a qPCR thermal cycler. Melting temperatures (T_m^{app}) are listed in Table 4.8.

4.2.7 EGFP$^{D190\Delta}$ and EGFP$^{A227\Delta}$ crystallography

In order to determine the effects the single amino acid deletion mutations have on the EGFP structure at the molecular level the EGFP$^{D190\Delta}$ and EGFP$^{A227\Delta}$ (10 mg/ml in 50 mM Tris-HCl, pH 8.0 and 150 mM NaCl) were screened for crystal formation by the sitting drop vapour diffusion method at 4°C.

A crystal of EGFPΔ^{D190} was obtained from 0.1 M Hepes/NaOH, pH 6.6, 200 mM ammonium sulphate and 2 M (K/Na)-Phosphate and a crystal of EGFPΔ^{A227} was obtained from 0.05 M MES/NaOH, pH 5.6, 100 mM ammonium sulphate, 10 mM magnesium chloride hexahydrate and 20% (w/v) PEG 8000.

However, due to time constraints the crystal structures have not yet been refined to a high enough resolution from the X-Ray diffraction data. Preliminary structures have been generated but the R-values for EGFP$^{D190\Delta}$ and EGFP$^{A227\Delta}$ are 36.8% and 28.6%. Therefore, in their current state these structures were considered of limited value.

4.3 Discussion

As mentioned previously predicting the effects of a single amino acid deletion and where it will be tolerated within a target protein is currently very difficult and generally avoided when undertaking rational protein design. Therefore, a directed evolution approach provides us with the opportunity to survey protein tolerance to single amino acid deletion. Using the directed evolution approach described in Chapter 3 the tolerance and impact of single amino acid deletions on EGFP have been investigated. This has included the identification of variants resulting in improved cellular fluorescence and altered folding properties.

GFP has been shown to tolerate insertion of short peptides [121], long peptides [121] and whole domains [34] but was intolerant to site directed deletions from the loops connecting secondary structures [122, 123]. This implies that GFP is almost a ‘size-minimized’ protein intolerant to loop shortening, as is also evident from its compact structure with high resistance to heat [122, 123], denaturing agents [113, 114] and proteolysis [124]. However, to date amino acid deletion mutagenesis of GFP has only been studied by N- and C-terminal truncations [122, 123, 125], shortening of the

longest loop connecting β-strands 6 and 7 [122], and deletion of residues from three other loops (residues 76-81, 83-88 and 191-195) [123].

Deletion of the terminal amino acids from GFP identified the minimal domain requirement for the protein to fold to its fluorescent form [123, 125], whilst deletion of amino acids from the longest loop linking β-strands 6 and 7 from GFP has showed this region is intolerant to shortening by single amino acid deletion [122]. A survey of GFP tolerance to single amino acid deletions throughout its structure has not been performed until now.

4.3.1 The crystal structure of EGFP.

Prior to this study the crystal structure for EGFP was yet to be determined despite its very wide use as a tool in cell biology. The high resolution (1.35 Å) structure determined here (Fig 4.2) has helped to identify the structural effects of the S65T and F64L mutations in EGFP and how they confer increased folding stability at 37 °C and alter the spectral properties with respect to wt GFP. The structure of EGFP was very similar to that of wt GFP with superposed structures having an RMSD across backbone atoms of only 0.33 Å.

Mutation of the bulky side chain of F64 for the smaller leucine side chain results in the better packing of hydrophobic core residues, especially residues L18, V29, and W57 that lie close to the chromophore (Fig 4.2). L64 together with these three residues form a hydrophobic interaction network, with W57 normally partially exposed becoming more buried (12.82 $Å^2$ surface exposed) compared to wt GFP (15.16 $Å^2$ surface exposed). This improved packing is a potential reason for the increased folding stability at 37 °C.

The T65 side chain in the EGFP structure determined here is in the same orientation as previously determined S65T-GFP structures [69]. However, in contrast to other S65T-GFP structures the electron density of the E222 carboxylate side chain in EGFP suggests that it occupies two distinct conformations. Both conformations result in the promotion of an anionic form of the chromophore, This is due to the E222 carboxylate group being neutral therefore allowing charge stabilisation on the phenyl group of the chromophore by H148, T203 and a conserved water molecule coordinated between residues N146 and S205. However, given the heterogeneity in the local environment of the chromophore due to the alternate conformations of E222 it would

be expected that this would be reflected by spectral heterogeneity. This is not the case as is evident from single exponential fluorescence lifetime decays and is potentially due to the E222 carboxylate being neutral, therefore small alterations in its conformation is not likely to effect the electrostatic environment surrounding the chromophore.

The two side chain conformations observed for E222 could be a crystallographic artifact and both may not be populated in solution. Alternatively both conformations may exist but could be in a dynamic equilibrium and transiently exchanging between the two conformations. A third possibility could be that upon folding the E222 side chain is trapped in one conformation or the other. Further structural analysis by NMR could potentially identify how dynamic the E222 side chain is in solution and whether the two conformations observed in the crystal structure are true.

4.3.2 Tolerance of EGFP to single amino acid deletion.

The use of the transposon-based directed evolution approach on the target EGFP has provided to date one of the most comprehensive lists of residues tolerant or intolerant to single amino acid deletions: 42 unique single amino acid deletions were tolerated (Table 4.3) and 45 non-tolerant positions were also identified throughout EGFP (Table 4.4). This is in contrast to previous amino acid deletion studies, which suggested EGFP to be largely intolerant to amino acid deletions [122, 123].

Analysis of the tolerated and non-tolerated amino acid deletion positions by mapping to the secondary structure topology (Fig 4.4) and tertiary structure of EGFP (Fig 4.5) showed a clear divide between regions tolerable and non-tolerable to deletion mutagenesis. Our results here agree to an extent with the dogma that deletion mutations are better tolerated in loops rather than ordered secondary structure. The majority (60%) of the tolerated amino acid deletions were situated in loops while the majority (71%) of non-tolerated amino acid deletions located towards the middle of β-strands. This can be explained by removal of a single amino acid from a secondary structure causing registry shifts in the strands (Chapter 1, Fig 1.1). Given that EGFP fluorescence relies on its stable regular structure, altering the registry of the β-strands would have obvious detrimental effects.

Although the majority of tolerated amino acid deletions were located in the loop regions, 21% were situated towards the termini of β-strands, in particular towards the end of β-strands 7 and 8 (Fig 4.4). There is also a high tolerance to amino acid deletions in the loops adjacent to these sites (Fig 4.4) possibly identifying this region of EGFP as being particularly tolerant to mutagenesis. β-strands may be more tolerant to single amino acid deletions positioned towards their N- and C-termini as a registry shift may be accommodated by a small loss of secondary structure and replacement by a loop, which may not affect the overall structure of the protein.

Although the loops connecting secondary structures were most tolerant to single amino acid deletions only two mutations, D133Δ and G138Δ, were identified in the longest loop connecting β-strands 6 and 7 of EGFP, which resulted in maintained fluorescence. The longest loop in EGFP (I128-L141) forms a cap over one end of the β-barrel, protecting the chromophore from the external environment (Fig 4.16 a), and has been shown previously to be intolerant to amino acid deletion [122]. There are several residues (F130, I136, L137, L141 and Y145) with buried side chains or backbone atoms acting as anchor points for the loop to the top of the β-barrel (Fig 4.16 b). Disruption of these anchor points could destabilize the protein structure or result in misfolding. Deletion of both L137 or Y145, situated in the longest loop, was detrimental to EGFP fluorescence (Table 4.4) agreeing with previous findings [122].

However, in contrast to previous studies deletion of D133 or G138 were tolerated. This could potentially be due to the fact that these residues are surface exposed (Fig 4.16 b) and therefore their deletion may not disrupt the anchor points formed by the buried residues.

Besides buried side chains, residues in the longest loop make hydrogen-bonding interactions with residues in adjacent loops, potentially further stabilizing overall EGFP stability. Residue D103 is situated in a loop connecting β-strands 4 and 5 adjacent to the N-terminal end of the longest loop (Fig 4.16 c), deletion of which resulted in loss of fluorescence. The D103 side chain carboxylate forms a hydrogen bond to the amide nitrogen of K131 situated in the longest loop (Fig 4.16 b). Deletion of D103 would result in the loss of this stabilizing interaction and could potentially destabilize both the longest loop and the loop connecting β-strands 4 and 5. This in turn could affect protein folding and stability explaining the loss of fluorescence in the D103 deletion mutant.

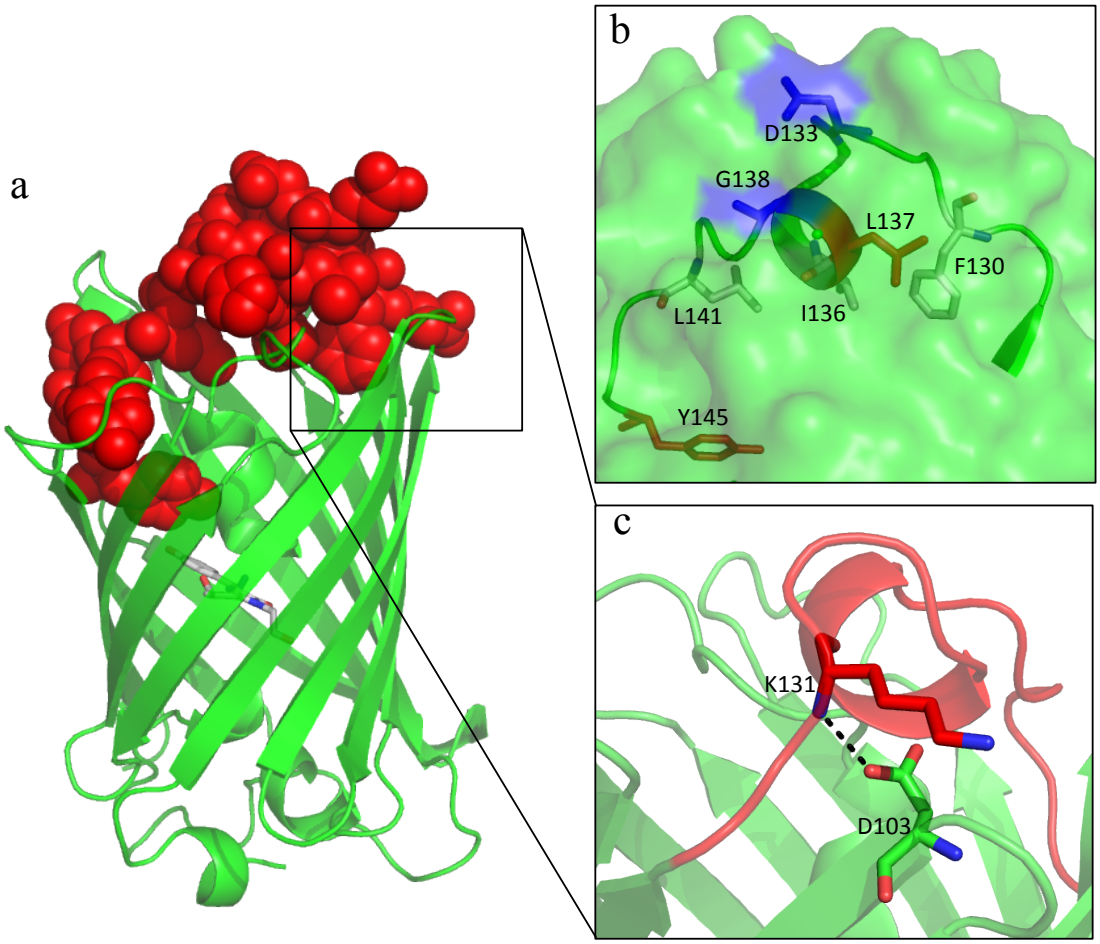


Fig 4.16 Buried residues and stabilizing interactions in the longest loop of EGFP. a, The longest loop (residues I128-L141, red spacefill) in EGFP (green cartoon) forms a cap over the end of the β-barrel protecting the chromophore (sticks) from the external environment. **b,** Buried residues act as anchor points for the longest loop. Deletion of the buried residue L137 (red) or Y145 (N-terminal end of the longest loop: red) result in loss of fluorescence. Deletion of solvent exposed residues D133 or G138 from the longest loop are tolerated. **c,** Stabilizing hydrogen bond (black dashed line) interaction between D103 and K131 (green or red sticks respectively). Deletion of D103 results in loss of fluorescence.

Amino acid deletions were also identified in the central α-helix, containing the chromophore forming residues, running through the core of the β-barrel structure (Table 4.4). All of the deletions in this region resulted in loss of fluorescence. Deletions in and around the chromophore forming residues are likely to disrupt correct folding and packing around the chromophore and therefore affect the maturation process, with the probability of producing non-fluorescent variants.

There was a high proportion of single amino acid deletions tolerated in other α-helix secondary structures (H1 and H3, Fig 4.4), in particular in the first α-helix. This is potentially due to the fact that truncation of the first five amino acid residues can be tolerated and fluorescence can still mature [125], therefore any structural disruption of this α-helix due to a single amino acid deletion is also likely to be tolerated.

The third α-helix (H3) in EGFP is part of one of the longer loops that spans one end of the β-barrel and whilst important the structure is not conserved across all fluorescent proteins. This loop spanning the end of the β-barrel in some structures of fluorescent proteins is one long distorted helix structure, in some is split into two helices and in others is a single short helical segment. Given that this helix differs greatly between different fluorescent protein structures implies a degree of flexibility and therefore is the likely reason it is more tolerant to single amino acid deletions than other secondary structures in EGFP.

The fact that tolerated and non-tolerated amino acid deletions have been identified in loops and in organized secondary structures highlights the necessity of a directed evolution approach. However from the survey of tolerated and non-tolerated single amino acid deletion mutations carried out here it is possible to devise some very preliminary rules for rationally introducing deletion mutations.

1. Loop regions connecting secondary structures are more likely to tolerate the deletion of amino acids.

2. Deletion of amino acids from β-strands are more likely to be tolerated when positioned towards the termini of that secondary structure

3. Deletion of amino acids from helix structures are also more likely to tolerate deletions from their termini although may also tolerate deletions from the middle of the structure dependent on the structural importance of the α-helix.

It is important to state that without crystal structure determination of all the amino acid deletion mutants of EGFP it is very difficult to characterize the exact effect of the deletions at the molecular level and thus it is still difficult to devise defined rules for introducing deletion mutations rationally. The rules defined here are to be used with caution given the lack of structural information but also they are dependent on the protein being studied, in particular rules 2 and 3. Deletion of amino acids from organized secondary structures although may be tolerated in EGFP towards their termini may not in other proteins. It is therefore important to have a detailed knowledge of the protein structure to be studied and to make educated decisions on whether the organized secondary structures are likely to be important for structural stability or function.

4.3.3 Identification of novel EGFP variants through single amino acid deletion mutagenesis.

Whole cell fluorescence analysis of cultures producing the tolerated amino acid deletion variants of EGFP (Fig 4.6, Fig 4.8), and the observation of brighter colour phenotypes in colonies grown on LB Agar (Fig 4.9 b) identified three single amino acid deletions, G4Δ, D190Δ and A227Δ that produced a marked increase in fluorescence with respect to EGFP. At 37 °C $EGFP^{A227\Delta}$ exhibited a 2.6-fold increase in fluorescence with respect to EGFP. However, the G4Δ mutation resulted in an increase in EGFP fluorescence when expressed in cell culture at 25 °C and 37 °C (~4-fold or ~1.7 fold respectively) (Fig 4.8). The effect of the G4Δ mutation was not just specific to EGFP as transfer of this mutation to EYFP also resulted in increased cellular fluorescence (Fig 4.9 b). Therefore, the G4Δ could be considered to be having a generic effect on GFP protein and should be incorporated into other variants to improve cellular fluoresence.

Although the identified deletion mutations increased fluorescence when expressed in cell culture the effects of the deletion mutations were not mediated through changes to the spectral properties as was evident from calculated brightness

values comparable to that of EGFP (Table 4.6). The effect is likely to be manifested through changes in protein stability and folding. Guanidine hydrochloride induced equilibrium unfolding experiments (Fig 4.12) showed that, with respect to EGFP, the $EGFP^{G4\Delta}$ variant has increased stability by ~1.18 $kcal.mol^{-1}M^{-1}$ and that the $EGFP^{A227\Delta}$ variant has marginally increased stability by ~0.53 $kcal.mol^{-1}M^{-1}$, despite having similar $[GdmCl]_{50\%}^{app}$ values (~2.5 – ~2.7 M) (Table 4.8). The increase in stability is due to an increase in cooperativity with which $EGFP^{G4\Delta}$ and $EGFP^{A227\Delta}$ unfold (~1.3-fold) with respect to EGFP, evident from the increased m-values (Fig 4.12 and Table 4.8). The increased stability for the $EGFP^{G4\Delta}$ variant may be due to the formation of additional stabilizing interactions. A lysine residue adjacent to G4 in EGFP (K3) may move to a more favorable position within the first α-helix and form stabilizing interactions between its ε-amino group and other residues in its immediate vicinity. Structural determination of $EGFP^{G4\Delta}$ would greatly aid in confirming this notion and would help towards identifying the molecular basis for its increased stability.

The data measured here were best fit to a 2 state unfolding model. However, if the data were to be fit to a 3-state model two m-values would be determined for a transition from a native to an intermediate state and from the intermediate state to the denatured state. This would allow a more accurate determination of the nature of EGFP unfolding with respect to the stability and degree of remaining structure of the intermediate state. Further equilibrium unfolding experiments would be required to assess a three state unfolding model for EGFP and the EGFPΔ variants. Performing the experiments at lower pH or higher temperature may help toward identifying an unfolding intermediate as seen in other equilibrium unfolding studies [113, 126].

Given that other GFP proteins have been shown to unfold by a three state model [113, 114] it is expected that unfolding kinetics would reflect this with two rate constants, one for the native to intermediate state and one for the intermediate to denatured state. However this would only be true for a system in which the formation of the intermediate from the native state was described by the fast rate constant with the formation of the denatured state from the intermediate being described by the slow rate constant. If the formation of the intermediate from the native state is described by a low rate constant it is unlikely that the fast rate constant would be distinguishable. A single unfolding kinetic rate constant was determined here for the unfolding of EGFP and the EGFPΔ variants (Fig 4.14 and Table 4.8). This is probably due to the protein

unfolding being performed in 6.8 M GdmCl. At this high concentration of GdmCl it is unlikely that any unfolding intermediates will be occupied, or they will unfold so rapidly to the denatured state that a rate constant for their formation, from the native state, will be unidentifiable by fitting to a double exponential decay function. Unfolding kinetic studies of a GFP variant (GFP mut2) at lower concentrations of GdmCl (< 5 M) have previously been shown to fit well to double exponential decay processes, but above 5 M GdmCl only a single exponential decay can be resolved from the data [127, 128].

The EGFPΔ variants exhibited similar unfolding kinetic rate constants (~2 min^{-1}) to that of EGFP (Table 4.8 and Fig 4.14) except for $EGFP^{A227\Delta}$, which unfolded with a more than two-fold faster rate constant (4.59 min^{-1}). This was an unusual result given that $EGFP^{A227\Delta}$ was found to be more stable than EGFP from equilibrium unfolding studies (Fig 4.12 and Table 4.8), which suggests that the two-state analysis for equilibrium unfolding is insufficient for determining EGFP stability. A more detailed study on unfolding kinetics for this variant at varying concentrations of GdmCl would be required to confirm the result seen here and to ascertain a dependence between unfolding rates and [GdmCl].

In agreement with previous work on protein refolding kinetics [71, 115, 128] EGFP and the EGFPΔ variants refolding data was best fit to a double exponential function. In contrast to refolding kinetics determined for urea denatured EGFP in previous studies [115], the fast and slow rate constants determined here are three-fold and >six-fold faster respectively (Table 4.8). The amount of fluorescence intensity recovered by EGFP (77%) (Fig 4.14) did agree with the previous findings for refolding studies [115], in contrast to $EGFP^{G4\Delta}$ which was capable of recovering up to 100% of its initial fluorescence (Fig 4.14).

The $EGFP^{G4\Delta}$, $EGFP^{D190\Delta}$ and $EGFP^{A227\Delta}$ variants all showed a decrease in the fast rate constant by up to 30% with respect to EGFP. The slow rate constants for $EGFP^{D190\Delta}$ and $EGFP^{A227\Delta}$ were also decreased by up to 11% whilst the G4Δ variant resulted in an increase of k_{slow} by up to 20% (Table 4.8). The two rate constants are thought to arise from *cis/trans* isomerization of the peptide bond between Met88 and Pro89 (M88-P89) [115]. In the native protein this peptide bond is in the *cis* configuration however in the denatured state the bond can exist in both configurations, *cis* or *trans*. Usually *trans* peptide bonds are favoured over *cis* in unfolded proteins

(*trans:cis*, 1000:1) however in an X-Pro peptide bond this ratio can drop to 3:1, due to the restricted torsion angles of proline residues making both isomers almost energetically favorable [115, 116].

The initial fast refolding phase is thought to be due to unfolded protein with the M88-P89 peptide bond already in the *cis* conformation whilst the slow rate is thought to be due to this peptide bond being in the *trans* configuration and therefore requiring isomerization for correct folding [115]. Residue G4 is in close proximity to the M88-P89 peptide bond (Fig 4.17). Deletion of G4 may cause favorable interactions between other residues surrounding the M88-P89 peptide bond promoting *cis/trans* isomerization during folding, therefore increasing the slow rate constant (k_{slow}) with respect to EGFP.

However, the kinetics of refolding calculated here can only be attributed to the folding of the β-barrel around the chromophore and protection from the external environment for fluorescence to be regained. The fact that the deletion variants increased the levels of fluorescence during the expression studies (Section 4.2.4) may partly be due to the altered rates of folding of the β-barrel (Table 4.8), however the effect of the deletion mutations on the rate at which the chromophore forms have not been determined here. Further work would be required to see if the deletion mutations have an effect on chromophore maturation, which could also play a major role in the rate of fluorescence development.

Determining the structures for the single amino acid deletion variants of EGFP would help a lot towards understanding the mechanisms by which they are affecting folding and stability. Determination of the $EGFP^{D190\Delta}$ and $EGFP^{A227\Delta}$ variant atomic resolution structures is still ongoing. Attempts to crystalize $EGFP^{G4\Delta}$ have been made but to date crystallization conditions are still to be identified that provide crystals for structure determination.

4.3.4 Effect of single amino acid deletion mutations on the oligomeric state of EGFP

GFPs can form weakly associated dimers at increased concentrations with a dissociation constant (K_d) of ~100 μM [108]. Under certain crystalisation conditions wild-type GFP crystalized as a dimer identifying extensive dimerization contacts

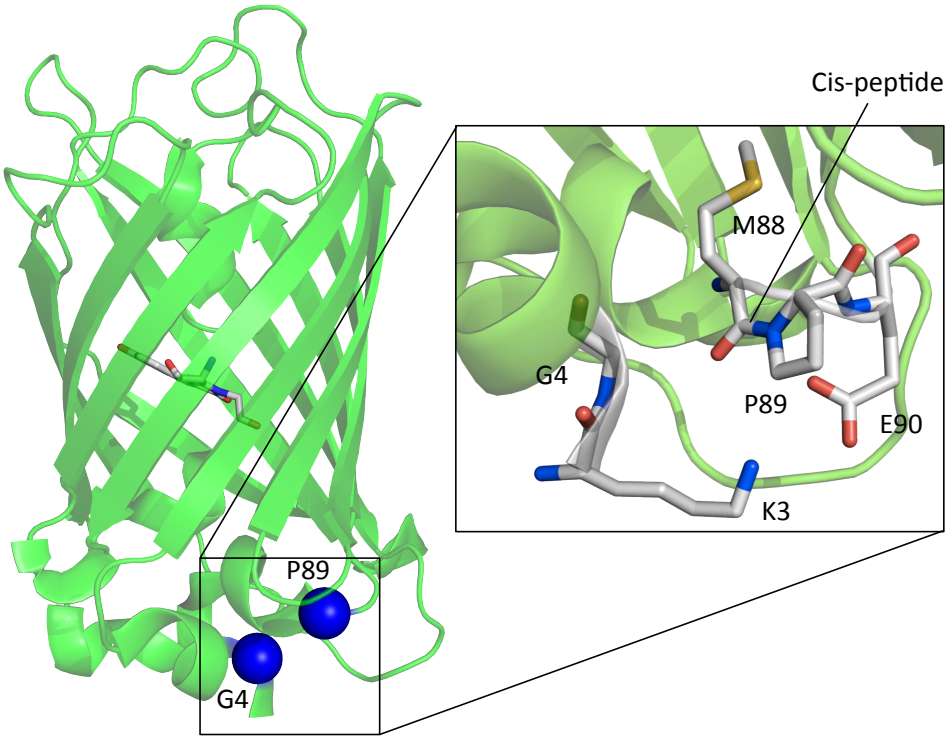


Fig 4.17. Location of G4 in EGFP with respect to the native M88-P89 cis-peptide bond. G4 is adjacent to P89 (blue spheres) in the tertiary structure of EGFP. Deletion of G4 would result in an altered environment around the native cis-peptide bond between M88 and P89 (inset) possibly influencing the rate of cis/trans isomerization during folding.

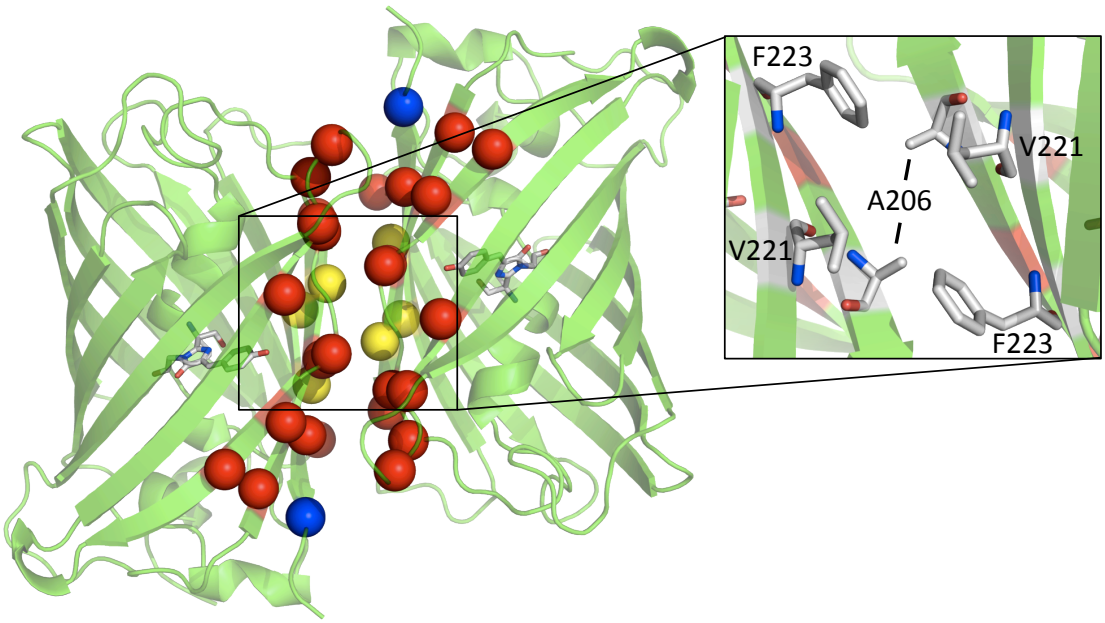


Fig 4.18. GFP dimerization interface. The dimer interface of wild-type GFP (pdb:1GFL) is comprised of 3 hydrophobic residues from each monomer (yellow spheres and inset) with the majority of the interactions being contributed from hydrophilic residues (red spheres). A227 (blue sphere) is positioned at the periphery of the dimerization interface.

between three hydrophobic residues (A206, L221 and F223) and many hydrophilic contacts (Fig 4.17). The single substitution mutation A206K has been shown to suppress the tendency for GFPs to dimerize. Analytical size exclusion chromatography showed that the EGFP$^{G4\Delta}$ and EGFP$^{D190\Delta}$ had similar oligomeric properties to that of EGFP (Fig 4.11) with a small shift in elution volume (~0.15 ml) with increasing protein concentration.

In contrast EGFP$^{A227\Delta}$ showed no change in elution volume with increasing protein concentration (Fig 4.11) implying a decrease in the tendency of this variant to form weakly associated dimers. Although A227 is not considered to be a residue that contributes to the dimerization of GFPs it is in close proximity to the residues that do form the dimerization interface (Fig 4.18), possibly decreasing the propensity for EGFP to form weak dimers when deleted.

4.3.5 Conclusion

Contrary to current dogma, it has been shown here that deletion of a single amino acid can be tolerated at a wide variety of positions, including secondary structure. Furthermore, it has been shown that variants with beneficial properties can be generated validating the use of deletion mutations as an approach to engineer useful properties into a protein. Many of the identified mutations may be considered to be non-intuitive so the use of a directed evolution approach has allowed a non-biased assessment of the impact of deletion mutations. Contrary to the thought that amino acid deletions destabilize a protein, the basis for variants with improved cellular fluorescence was at the level of protein stability and folding.

Chapter 5: Characterisation and analysis of cytb_{562}-EGFP integral fusion scaffolds

5.1 Introduction

Construction of tailored protein scaffolds not currently present in nature is integral to synthetic biology and bionanotechnology [129-132]. The generation of protein scaffolds linking the disparate functions of two unrelated proteins [132-134] is particularly attractive as subsequent chimeras have the potential to act as biomolecular switches, encompassing: sensors, modulators, transducers and energy transfer components for use in natural and artificial contexts.

Single polypeptide chains that fold into multi-domain architectures are common in nature, with the relative arrangement of individual domains generally important for functional linkage [23]. Most multi-domain proteins are linked in a 'head-to-tail' manner with the C-terminus of one domain linked directly to the N-terminus of the next [25]. A significant minority (9%) exhibit integral fusion architecture with one protein domain inserted within another [24, 25].

Integral fusion architecture decreases the degrees of freedom the 'insert' domain has with respect to the accepting or 'parent' domain, thereby intimately linking the two structures. This intimate linkage has the potential to communicate stimulus-induced changes in one domain through to the other therefore coupling the structure and function of the two normally disparate proteins. A classic example in nature is the concept of allostery where binding to a regulatory site distant from the active site can modulate the function (in a positive or negative manner) through propagated structural changes [135].

Our limited understanding of the functional linkage, between the two domains, of engineered domain insert protein scaffolds is exacerbated by a lack of detailed structural information. This ultimately makes identifying sites within a parent protein that will not only tolerate the insertion of a domain but also retain and link the functions of the two domains very difficult. There has been limited success using rational design approaches [32, 34, 35], with this technique relying on a detailed knowledge of both protein structures, often resulting in only modest switching magnitudes.

Directed evolution provides an alternative approach and has been applied in a few cases to generate libraries of integral domain inserts from which variants with

switching characteristics have been identified [31, 37, 136, 137]. The limitation of previously used directed evolution methods are that random breaks introduced into the target DNA, into which genes encoding whole domains were inserted, were generated using DNAseI. It is notoriously difficult to generate single random breaks into target DNA with DNAseI and digestion with this nuclease resulting in tandem duplications and nested deletions of unpredictable sizes within the parent gene [32]. DNAseI also shows considerable target site specificity resulting in DNA libraries with bias [48].

Here we have used a transposon based directed evolution approach to introduce defined random breaks throughout a target gene, *egfp*, coding for enhanced green fluorescent protein (EGFP), into which DNA cassettes coding for cytochrome b_{562} (cyt b_{562}) were inserted (Chapter 3, Section 3.2.7). The subsequent library of cyt b_{562}-EGFP integral fusion proteins was screened for the functional coupling of their normally disparate functions for the use as an efficient energy transfer component and the potential to act as a sensor for redox state.

EGFP is a perfect model protein for studying integral fusion scaffold design as it has intrinsic auto fluorescence upon folding making domain insertion tolerance and coupling easy to determine. Furthermore, EGFP is one of the most common tools used in cell biology for studying gene expression, protein localisation, protein trafficking and protein-protein interactions to name a few. The chromophore is very sensitive to changes in its local environment and therefore functional coupling can easily be detected by changes in fluorescence characteristics upon the binding of an effector or other input.

Cyt b_{562} is a four helix bundle protein that binds haem non-covalently with a methionine (Met 7) and histidine (His 102) co-ordinating the iron moiety [73-75]. Cyt b_{562} undergoes a major change in conformation on binding haem most notably at the C-terminal helix, changing from a dynamic unordered structure to a structurally ordered α-helix [73]. This conformational change, upon haem binding, has the potential to act as the structural switch with which to modulate a parent domains function, in this case EGFP fluorescence.

In naturally evolved integral fusion scaffold proteins the N- and C-termini of the insert domain are generally juxtaposed with an average distance of ~8 Å from one another [24]. The N-and C-termini of apo-cyt b_{562} are ~12 Å from one another making it a suitable candidate for an insert domain in an artificial integral fusion scaffold.

Haem is a biologically important small molecule that acts as a cofactor to many different proteins encompassing a wide range of roles such as; oxygen transport [78], catalysis [79], electron transfer [80] and sensing. Haem has also been shown to be an efficient fluorescence quencher [81], including that of EGFP [82, 138], with quenching occurring *via* energy transfer to the haem moiety bound to cyt b_{562}, the efficiency of which is dependent on the proximity and orientation of the two chromophores to one another. Haem affinity for cyt b_{562} is also dependent on the iron oxidation state [76] with bound haem being sensitive to oxidative modification [77] making it a potential sensor for changes in redox conditions.

5.2 Results

5.2.1 Library construction

The first step is to identify sites within EGFP that tolerate domain insertion. Several factors can influence tolerance: (1) frame of trinucleotide deletion, (2) frame of domain cassette insert, (3) domain insert position within the structure of EGFP and (4) nature of linker sequence separating the two domains. The first two are dependent on the MuDel transposon approach to library construction (Chapter 3) that can be overcome through experimental modification with relation to the domain cassette insert and variant screening. This will allow a wider number of variants to be constructed and thus sampled increasing the likelihood of identifying useful variants. The last two are more difficult to predict so can only be addressed through variant screening. The benefit of the MuDel transposon based approach is that the nature of break is always defined (in comparison to DNaseI) so making the construction of the crucial linker segments easier to incorporate into the library design process.

The linkers joining the two domains in an integral domain scaffold can be crucial to achieving functional coupling and the switching magnitude that can be achieved. Two different types of linkers were used in this study: short single random amino acid linkers (X) or longer more flexible Gly-Gly-Ser linkers as outlined in Chapter 3. Screening of the cyt b_{562}-EGFP libraries with either X or Gly-Gly-Ser linkers identified a greater tolerance of EGFP to cyt b_{562} domain insertion when the two domains were separated by the longer more flexible linkers (Chapter 3, Section 3.2.8). Therefore, only variants from the *cybC-egfp-GGS* library were selected for further characterisation.

In order to sample all possible open reading frames (ORF) in the library, *cybC* cassettes were designed with additional random nucleotides at the 5' and 3' ends to maintain the reading frame of *cybC* in ORFs 2 or 3 (Fig 5.1). Due to the random nucleotides at the 5' and 3' end of the ORF 2 and 3 *cybC* DNA cassettes, substitution mutations can also be introduced into EGFP when sampling a random break in *egfp* in these reading frames (Fig 5.1).

Consequently genes encoding cyt b_{562}-EGFP integral domain fusions with Gly-Gly-Ser linkers were isolated from *E. coli* cells and had their DNA sequence determined to identify the nature of the *cybC* DNA cassette insertion into *egfp*.

5.2.2 Analysis of tolerated and non-tolerated cytochrome b_{562} domain insertions in EGFP.

The library of integral domain insertion variants was screened for EGFP tolerance to domain insertion. The library was used to transform *E. coli* TUNER™ (DE3) cells (Table 2.1) subsequently grown overnight at 37 °C on M9 minimal media agar plates supplemented with 100 μg/ml ampicillin and 150 μM IPTG followed by storage at 4 °C for up to 72 hrs. Colonies that displayed a green fluorescent phenotype upon excitation with a UV-transilluminator (n = 37) (Section 2.3.6) were selected and the gene encoding the integral fusion protein sequenced (Section 2.2.16). A variant *egfp* gene from 116 non-fluorescent colonies was also sequenced to determine what gave rise to inactive EGFP.

The DNA sequence of 36 different variants from the *cybC-egfp* domain insertion library that conferred cellular fluorescence (Table 5.1) and 116 different variants that conferred no observable colour phenotype (Table 5.2) were analysed. Of the 36 variants that conferred cellular fluorescence, 26 were identified to contain an in-frame *cybC* insertion sampling 15 different insertion positions within *egfp*. From the 15 different insertion positions sampled 23 unique cyt b_{562}-EGFP integral domain scaffolds were generated (Table 5.1). The disparity between the number of different insertion sites sampled and the number of unique variants identified arises due to different linker sequence being sampled at the same insertion site (Table 5.1).

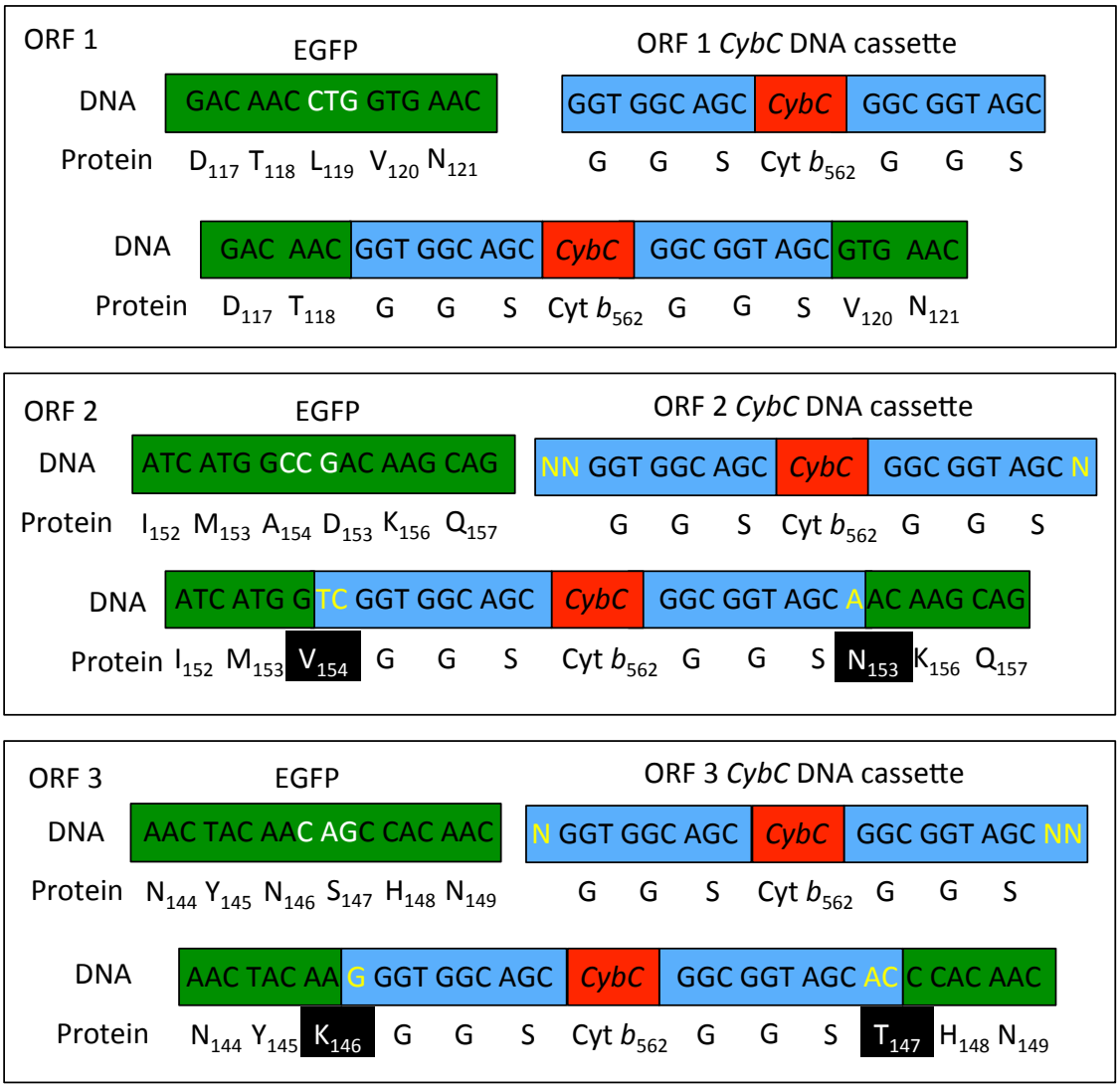


Fig 5.1 Sampling *cybC* insertion sites in all ORFs and subsequent mutations sampled. Due to the nature of MuDel insertion and subsequent removal, a triplet nucleotide (white sequence) is deleted from *egfp* (green) and results in a random break in the target gene that is not always in frame. Insertion of a *cybC* DNA cassette (red) encoding Gly-Gly-Ser linkers (blue) into the random breaks generates *cybC-egfp-GGS* DNA constructs that encode cyt b_{562}-EGFP integral fusion proteins. In order to sample the random breaks in ORF 2 and 3 *cybC* DNA cassettes with additional random nucleotides (yellow sequence) are used to maintain the reading frame. The resulting *cybC-egfp-GGS* DNA constructs therefore encode for cyt b_{562}-EGFP integral fusion proteins that can also have substitution mutations (black background with white sequence) flanking the cyt b_{562}. Representative DNA and encoded protein sequences of ORF 1 (top panel), 2 (middle panel) and 3 (bottom panel) *cybC* insertions are of variants NFCG28 (Table 5.2), CG14 and CG8 (Table 5.1) respectively.

Table 5.1 DNA Sequence and amino acid mutations of fluorescent integral cyt b_{562}-EGFP fusion proteins.

Variant	Nucleotides deleted[a]	Amino acid mutations[b]	Secondary Structure[c]	ORF[d]	Fusion protein[e]
CG1	$_{}$ATG GTG$_{7}$	V1*M	N-LP	2	GFP$_{1}$-GGS-AD-cytb-YR-GGSM-$_{2}$GFP
CG2	$_{12}$GGC GAG$_{19}$	E5Q	H1	3	GFP$_{4}$-GGS-AD-cytb-YR-GGSQ-$_{6}$GFP
CG3	$_{27}$ACC GGG$_{34}$	T9A	H1	2	GFP$_{8}$-AGGS-AD-cytb-YR-GGS-$_{10}$GFP
CG4	$_{27}$ACC GGG$_{34}$	G10R	LP H1-S1	3	GFP$_{9}$-GGS-AD-cytb-YR-GGSR-$_{11}$GFP
CG25	$_{27}$ACC GGG$_{34}$	G10K	LP H1-S1	3	GFP$_{9}$-GGS-AD-cytb-YR-GGSK-$_{11}$GFP
CG5	$_{90}$TCC GGC$_{97}$	-	S2	2	GFP$_{30}$-GGS-AD-cytb-YR-GGS-$_{31}$GFP
CG6	$_{117}$TAC GGC$_{124}$	Y39F	LP S2-S3	2	GFP$_{38}$-FGGS-AD-cytb-YR-G-$_{40}$GFP
CG7	$_{150}$ACC GGC$_{157}$	T50N, G51S	LP S3-H3	2	GFP$_{49}$-NGGS-AD-cytb-YR-GGSS-$_{52}$GFP
CG8	$_{438}$AAC AGC$_{445}$	N146K, S147T	LP S6-S7	3	GFP$_{145}$-KGGS-AD-cytb-YR-GGST-$_{148}$GFP
CG15	$_{438}$AAC AGC$_{445}$	S147C	LP S6-S7	3	GFP$_{146}$-GGS-AD-cytb-YR-GGSC-$_{148}$GFP
CG26	$_{456}$ATC ATG$_{463}$	M153L	S7	2	GFP$_{152}$-GGS-AD-cytb-YR-GGSL-$_{154}$GFP
CG17	$_{456}$ATC ATG$_{463}$	M153L	S7	3	GFP$_{152}$-GGS-AD-cytb-YR-GGSL-$_{154}$GFP
CG9	$_{462}$GCC GAC$_{469}$	A154G	S7	2	GFP$_{153}$-GGGS-AD-cytb-YR-GGS-$_{155}$GFP
CG14	$_{462}$GCC GAC$_{469}$	A154V, D155N	S7	2	GFP$_{153}$-VGGS-AD-cytb-YR-GGSN-$_{156}$GFP
CG19	$_{462}$GCC GAC$_{469}$	-	S7	2	GFP$_{154}$-GGS-AD-cytb-YR-GGS-$_{155}$GFP
CG23	$_{462}$GCC GAC$_{469}$	A154G, D155H	S7	2	GFP$_{153}$-GGGS-AD-cytb-YR-GGSH-$_{156}$GFP
CG13	$_{462}$GCC GAC$_{469}$	D155N	S7	2	GFP$_{154}$-GGS-AD-cytb-YR-GGSN-$_{156}$GFP
CG10	$_{462}$GCC GAC$_{469}$	D155N	S7	3	GFP$_{154}$-GGS-AD-cytb-YR-GGSN-$_{156}$GFP
CG18	$_{522}$GGC AGC$_{529}$	G174V, S175C	LP S8-S9	2	GFP$_{173}$-VGGS-AD-cytb-YR-GGSC-$_{178}$GFP
CG11	$_{693}$CTC GGC$_{700}$	-	C-LP	2	GFP$_{231}$-GGS-AD-cytb-YR-GGS-$_{232}$GFP
CG22	$_{693}$CTC GGC$_{700}$	L231H, G232S	C-LP	2	GFP$_{230}$-HGGS-AD-cytb-YR-GGSS-$_{233}$GFP
CG21	$_{693}$CTC GGC$_{700}$	G232S	C-LP	2	GFP$_{231}$-GGS-AD-cytb-YR-GGSS-$_{233}$GFP
CG20	$_{708}$CTG TAC$_{715}$	L236P, Y237N	C-LP	2	GFP$_{235}$-PGGS-AD-cytb-YR-GGSN-$_{238}$GFP
CG12	$_{711}$TAC AAG$_{718}$	Y237W, K238Q	C-LP	2	GFP$_{236}$-WGGS-AD-cytb-YR-GGSQ

[a]DNA sequence numbering as per the *egfp* gene [107]
[b]Amino acid numbering as per GFP [106], V2 of EGFP is numbered V1*.
[c]LP signifies loop, H signifies α-helix, S signifies β-strand, N- or C- signify N- or C-terminal respectively.
[d] ORF is the open reading frame into which the *cybC* DNA cassette is inserted
[e]Position of insertion within GFP denoted by subscript. Mutated residues in black, non-underlined. Linker sequences black underlined. The sequence segment representing cyt b is in red with just the terminal residues shown, the rest of the sequence is denoted by 'cytb'

Table 5.2 DNA Sequence and amino acid mutations of non-fluorescent integral cyt b_{562}-EGFP fusion proteins.

Variant	Nucleotides deleted[a]	Amino caid mutations[b]	Secondary structure[c]	ORF[d]	Fusion protein[e]
NFCG38	$_{45}$C**TG G**TC$_{52}$	L15P	S1	2	GFP$_{14}$-PGGS-AD-cytb-YR-GGS-$_{16}$GFP
NFCG26	$_{82}$T**TC A**GC$_{89}$	F27S, S28G	S2	2	GFP$_{26}$-SGGS-AD-cytb-YR-GGSG-$_{29}$GFP
NFCG3	$_{82}$T**TC A**GC$_{89}$	F27S	S2	2	GFP$_{26}$-SGGS-AD-cytb-YR-GGS-$_{28}$GFP
NFCG17	$_{91}$T**CC G**GC$_{98}$	S30Y	S2	2	GFP$_{29}$-YGGS-AD-cytb-YR-GGS-$_{31}$GFP
NFCG5	$_{112}$G**CC A**CC$_{119}$	A37V	LP S2-S3	2	GFP$_{36}$-VGGS-AD-cytb-YR-GGS-$_{2}$GFP
NFCG13	$_{127}$C**TG A**CC$_{134}$	L42P, T43P	S3	2	GFP$_{41}$-PGGS-AD-cytb-YR-GGSP-$_{43}$GFP
NFCG1	$_{142}$A**TC T**GC$_{149}$	C48S	S3	2	GFP$_{47}$-GGS-AD-cytb-YR-GGSS-$_{49}$GFP
NFCG23	$_{142}$AT**C TG**C$_{149}$	C48D	S3	3	GFP$_{47}$-GGS-AD-cytb-YR-GGSD-$_{49}$GFP
NFCG24	$_{170}$CC**C TG**G$_{177}$	W57S	H2	3	GFP$_{56}$-GGS-AD-cytb-YR-GGSS-$_{58}$GFP
NFCG7	$_{171}$**TGG**$_{175}$	W57Δ	H2	1	GFP$_{56}$-GGS-AD-cytb-YR-GGS-$_{58}$GFP
NFCG6	$_{189}$**ACC**$_{193}$	T63Δ	H2	1	GFP$_{62}$-GGS-AD-cytb-YR-GGS-$_{64}$GFP
NFCG16	$_{227}$CC**C GA**C$_{234}$	-	H3	3	GFP$_{75}$-GGS-AD-cytb-YR-GGS-$_{76}$GFP
NFCG14	$_{236}$AT**G AA**G$_{243}$	M78I, K79E	H3	3	GFP$_{77}$-IGGS-AD-cytb-YR-GGSE-$_{80}$GFP
NFCG25	$_{239}$AA**G CA**G$_{246}$	K79N, Q80K	H3	3	GFP$_{78}$-NGGS-AD-cytb-YR-GGSK-$_{81}$GFP
NFCG11	$_{259}$T**CC G**CC$_{266}$	S86C, A87P	H4	2	GFP$_{85}$-CGGS-AD-cytb-YR-GGSP-$_{88}$GFP
NFCG8	$_{259}$T**CC G**CC$_{266}$	S86W	H4	2	GFP$_{85}$-WGGS-AD-cytb-YR-GGS-$_{87}$GFP
NFCG10	$_{355}$A**CC C**TG$_{362}$	T118N	S6	2	GFP$_{117}$-NGGS-AD-cytb-YR-GGS-$_{119}$GFP
NFCG28	$_{357}$**CTG**$_{361}$	L119Δ	S6	1	GFP$_{118}$-GGS-AD-cytb-YR-GGS-$_{120}$GFP
NFCG34	$_{375}$CT**G AA**G$_{382}$	K126Q	S6	3	GFP$_{125}$-GGS-AD-cytb-YR-GGSQ-$_{127}$GFP
NFCG36	$_{411}$C**TG G**GG$_{418}$	L137Q, G138W	LP S6-S7	2	GFP$_{136}$-QGGS-AD-cytb-YR-GGSW-$_{139}$GFP
NFCG12	$_{411}$C**TG G**GG$_{418}$	L137P	LP S6-S7	2	GFP$_{136}$-PGGS-AD-cytb-YR-GGS-$_{138}$GFP
NFCG9	$_{414}$**GGG**$_{418}$	G138Δ	LP S6-S7	1	GFP$_{137}$-GGS-AD-cytb-YR-GGS-$_{139}$GFP
NFCG37	$_{423}$C**TG GA**G$_{430}$	L141P	LP S6-S7	2	GFP$_{140}$-PGGS-AD-cytb-YR-GGS-$_{142}$GFP
NFCG33	$_{489}$G**TG A**AC$_{496}$	V163A, N164D	S8	2	GFP$_{162}$-AGGS-AD-cytb-YR-GGSD-$_{165}$GFP
NFCG19	$_{489}$G**TG A**AC$_{496}$	V163G	S8	2	GFP$_{162}$-GGGS-AD-cytb-YR-GGS-$_{164}$GFP
NFCG2	$_{535}$C**TC G**CC$_{542}$	L178P, A179T	S9	2	GFP$_{177}$-PGGS-AD-cytb-YR-GGST-$_{180}$GFP
NFCG29	$_{537}$GC**C GA**C$_{544}$	D180S	S9	3	GFP$_{179}$-GGS-AD-cytb-YR-GGSS-$_{181}$GFP
NFCG15	$_{562}$C**CC A**TC$_{569}$	P187R, I188L	S9	2	GFP$_{186}$-RGGS-AD-cytb-YR-GGSL-$_{189}$GFP
NFCG21	$_{577}$C**CC G**TG$_{584}$	P192Q, V193M	LP S9-S10	2	GFP$_{191}$-QGGS-AD-cytb-YR-GGSM-$_{194}$GFP
NFCG4	$_{577}$C**CC G**TG$_{584}$	P192H, V193L	LP S9-S10	2	GFP$_{191}$-HGGS-AD-cytb-YR-GGSL-$_{194}$GFP
NFCG32	$_{621}$C**TG A**GC$_{628}$	L207P, S208G	S10	2	GFP$_{206}$-PGGS-AD-cytb-YR-GGSG-$_{209}$GFP
NFCG30	$_{645}$CG**C GA**T$_{652}$	D216Y	S11	3	GFP$_{215}$-GGS-AD-cytb-YR-GGSY-$_{217}$GFP

[a]DNA sequence numbering as per the *egfp* gene [107]

[b]Amino acid numbering as per GFP [106], V2 of EGFP is numbered V1*.

[c]LP signifies loop, H signifies α-helix, S signifies β-strand, N- or C- signify N- or C-terminal respectively.

[d] ORF is the open reading frame into which the *cybC* DNA cassette is inserted

[e]Position of insertion within GFP denoted by subscript. Mutated residues in black, non-underlined. Linker sequences black underlined. The sequence segment representing cyt b is in red with just the terminal residues shown, the rest of the sequence is denoted by 'cytb'

All 23 unique variants identified contained the *cybC* cassette insertion in ORF 2 and 3 with no fluorescent variants being sampled in ORF 1 (Table 5.1). This could potentially be due to the fact that the additional random nucleotides in the *cybC* DNA cassettes for ORF 2 and 3 encode for an additional amino acid making the linkers slightly longer (Fig 5.1). Very few variants, that retained fluorescence, were identified during library screening of cyt b_{562}-EGFP chimeras with a random single amino acid linker separating the two domains (Section 3.2.8). This is thought to be due to the reduced conformational flexibility, inherent to shorter linkers, impeding the correct folding of the EGFP domain allowing fluorescence to mature. The additional amino acid coded for by the random nucleotides in ORF 2 and 3 *cybC* cassettes (Fig 5.1) may provide the additional flexibility in the linking peptides between cyt b_{562} and EGFP required for the EGFP domain to fold correctly.

From the DNA sequence analysis of the 36 fluorescent variants, 10 were identified to have a *cybC* DNA cassette insertion either out-of-frame or in the reverse orientation; six variants had the *cybC* cassette insertion out of frame and four had a *cybC* cassette in the reverse orientation. Out of frame insertions can occur due to the mechanism of MuDel removal by *Mly*I as outlined in Section 5.2.1. Due to the nature of blunt end ligations the *cybC* DNA cassette can also be inserted in a reverse orientation with respect to *egfp*.

Despite having an out of frame or reverse *cybC* cassette insertion all 10 chimeras retained fluorescence. DNA sequence analysis of these 10 variants identified the *cybC* insertion was within the last 36 bp of the *egfp* gene, the last 33 bp of which encodes an 11 amino acid dynamic C-terminal loop not required for fluorescence to mature (Fig 5.2) [123, 125]. Given that this region of EGFP is not required for fluorescence to mature the insertion of a *cybC* DNA cassette out of frame or in a reverse orientation to *egfp* can still result in the EGFP portion of the construct folding and gaining fluorescence. However, all of these variants lack a mature cyt b_{562} domain and were therefore excluded from the library with no further analysis or characterisation being performed.

Of the 15 unique insertion sites identified resulting in fluorescent variants, 60% were located in loops, 27% were located in β-strands and 13% were located in

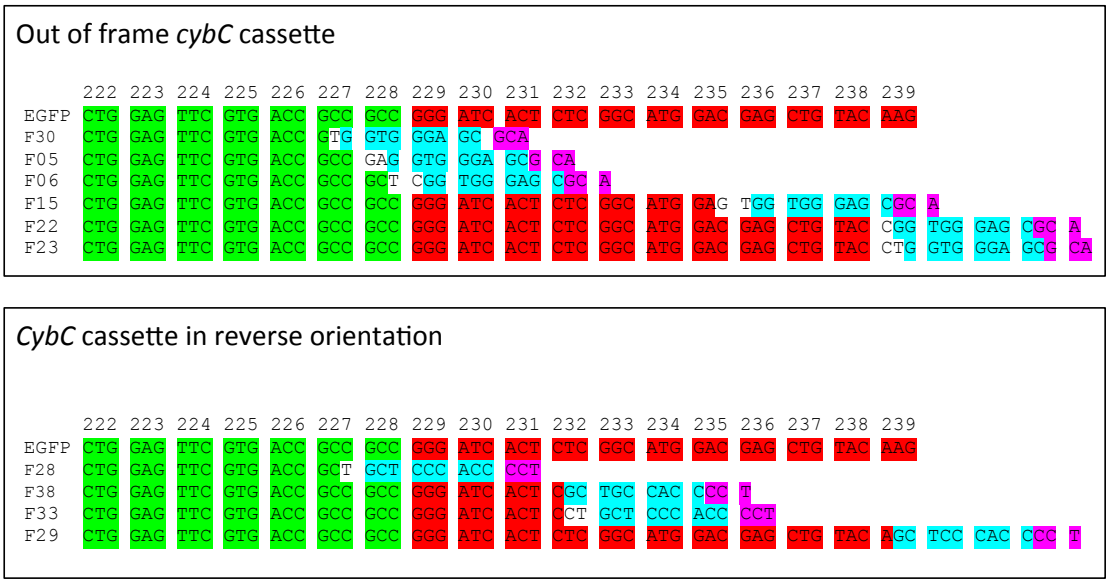


Fig 5.2 Sequence alignment of out of frame and reverse orientation *cybC* cassette insertions in *egfp*. Sequence alignments for fluorescent variants identified from the *cybC-egfp-GGS* library with either a *cybC* cassette insertion out-of-frame (top panel) with *egfp* (green) or with a *cybC* cassette in a reverse orientation (bottom panel) to *egfp*. The last 33 bp of *egfp* (red) encode for an 11 amino acid C-terminal dynamic loop not required for fluorescence to mature. Numbers above codons refer to the amino acid they code for in EGFP. Blue sequence corresponds to DNA encoding the Gly-Gly-Ser linkers, pink sequence corresponds to the first codon of the *cybC* gene (for out of frame) or last codon of *cybC* gene (reverse orientation). Non-highlighted sequence corresponds to random nucleotides encoded by the ORF 2 and 3 *cybC* DNA cassettes (Fig 5.1)

α-helices (Table 5.1, Fig 5.3 and Fig 5.4). One of insertion sites (A154) was sampled five times with each cyt b_{562} insertion resulting in different combinations of secondary point mutations at the N-and C-terminal ends of the linker sequence (CG9, CG13, CG14, CG19, CG23) (Table 5.1).

Insertion sites with a single amino acid either side of residue A154 (CG17, CG10) were also observed suggesting this region of EGFP may be highly tolerant to domain insertion. This region located at the C-terminal end of β-strand 7 (Fig 5.3, Fig 5.4) corresponds to the same β-strand observed in Chapter 4 to be tolerant to single amino acid deletion (Fig 4.4). Variant CG5 has cyt b_{562} inserted at residue S30 of EGFP, which is situated towards the middle of β-strand 2 (Fig 5.5). Library screening also identified two variants, CG1 and CG12, which are effectively classical N- and C-terminal fusions between cyt b_{562} and EGFP (Table 5.1).

Most of the tolerated insertion positions within the primary sequence of EGFP are distant from the chromophore forming residues. However, in the tertiary structure of EGFP some of the insertion sites are in close proximity to the mature chromophore (CG5, CG6, CG8, CG15), in particular variants CG8 and CG15, which have cyt b_{562} inserted adjacent to the tyrosyl group of the chromophore (Section 5.2.4, Fig 5.5). Mutation around this site has previously been shown to modify the protonation state of the chromophore, therefore altering the spectral characteristics [34, 139, 140].

DNA sequence analysis of 116 non-fluorescent variants identified 33 (28%), with *cybC* inserted in frame with *egfp*, sampling 28 unique insertion sites. The rest of the variants (83) contained a *cybC* DNA cassette insertion out-of-frame to *egfp* resulting in premature stop codons being introduced. For this reason in depth sequence analysis of these variants was not performed but the *cybC* insertion positions have been included in Appendix B. In-frame *cybC* DNA cassette insertion into *egfp* resulting in cyt b_{562} domain insertion into β-strands made up the majority of sites sampled (54%), with 21% identified in loops and 25% located in α-helices (Table 5.2, Fig 5.3, Fig 5.4).

As with the TND library (Chapter 4) there is a higher percentage of non-tolerated insertions into ordered secondary structures with less insertion sites sampled in the loop regions. There were three non-tolerated cyt b_{562} domain insertions identified in the chromophore containing central α-helix. Domain insertions into this α-helix would always result in non-fluorescent variants for the obvious reason that the β-barrel

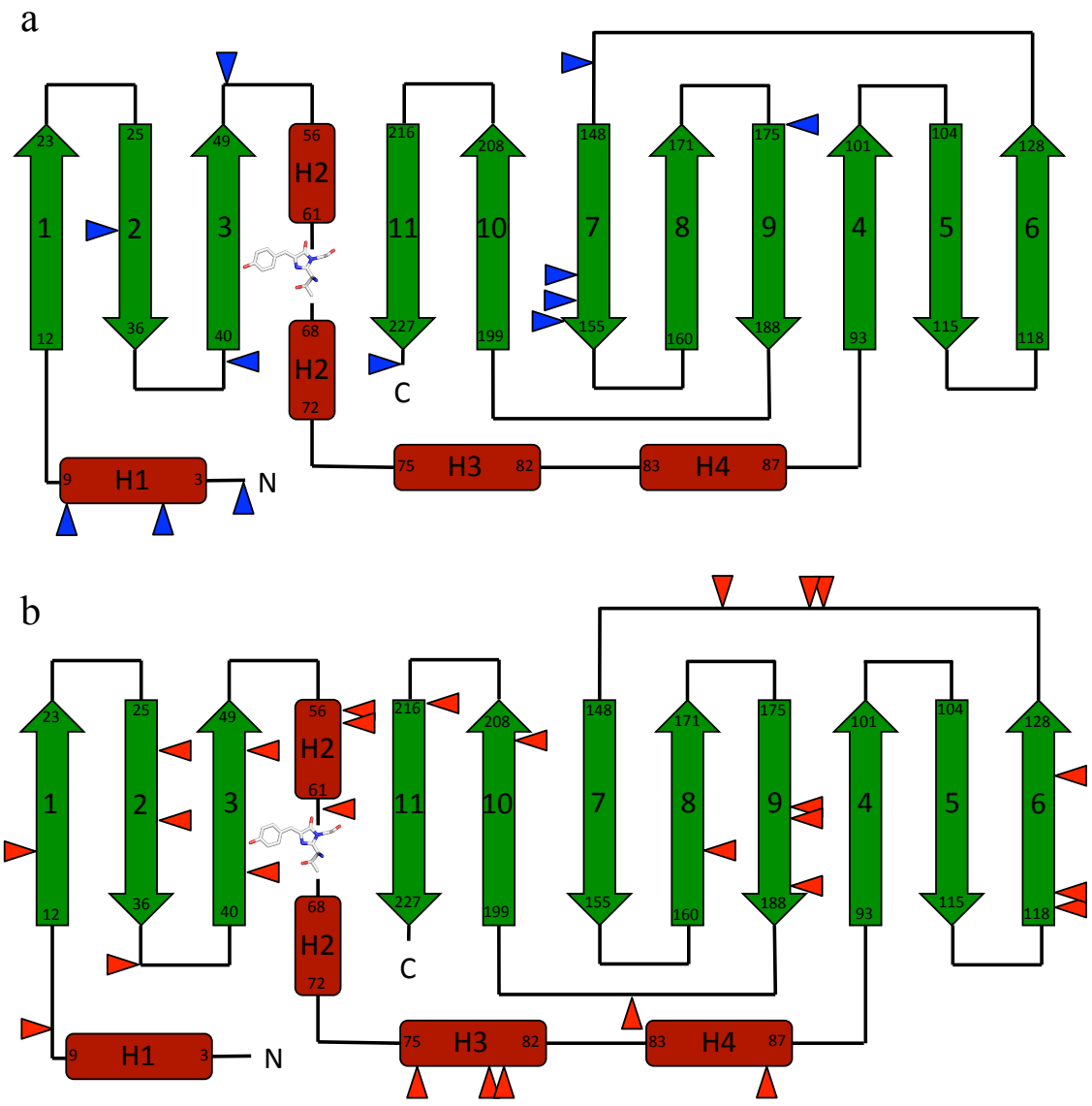


Fig 5.3 Secondary structure topology for EGFP. The secondary structure topology of EGFP shows the arrangement of β -strands (green), α-helices (red) and loops (black) with respect to one another in the folded protein. **a,** Tolerated cyt b_{562} domain insertions resulting in fluorescent variants are indicated by blue triangles. **b,** Non-tolerated cyt b_{562} domain insertions resulting in non-fluorescent variants are indicated by red triangles.

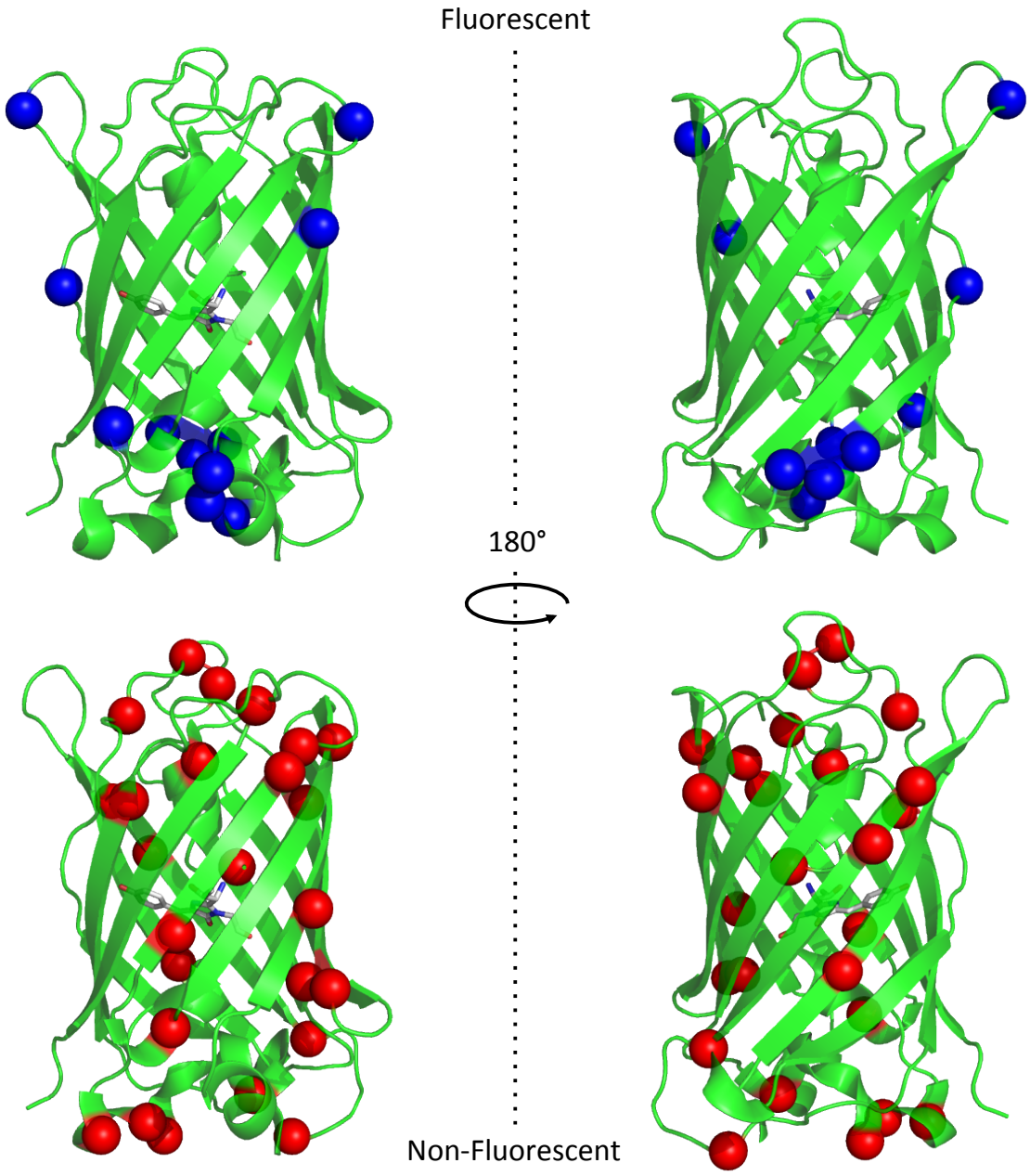


Fig 5.4 Tolerated and non-tolerated cyt b_{562} insertion positions in the tertiary structure of EGFP. Cartoon representation of EGFP (green) with tolerated cyt b_{562} insertion positions indicated by blue spheres and non-tolerated insertions indicated by red spheres.

structure would not be able to fold around another domain, therefore inhibiting chromophore maturation. Non-tolerated domain insertions were also identified in the longest loop of EGFP, which has been shown to play an important structural role for EGFP. Disruption of this loop by insertion of another whole domain may therefore reduce the structural stability or impede correct folding of EGFP.

5.2.3 Sequence analysis of fluorescent and non-fluroescent cyt b_{562}-EGFP fusion variants.

Secondary substitution mutations due to *cybC* insertion into ORF 2 and 3 of *egfp* (Fig 5.1) were identified in 87% of the fluorescent variants. Single substitution mutations were sampled in 12 (52%) of the variants, three of which were at the N-terminal side of cyt b_{562} with the other nine being at the C-terminal side (Table 5.1). Double substitution mutations, occurring at both N and C-terminal sides of cyt b_{562}, took place in eight (35%) of the variants (Table 5.1). In 74% of the cases (21) the substitution mutations were conservative (44%) or semi-conservative (30%). All of the non-conserved substitution mutations (6) were at positions where the residues are solvent accessible and in all but 2 cases the substitution mutations were from glycines or hydrophobic residues to charged or polar residues.

CybC insertions within *egfp* resulting in non-fluorescent variants were sampled in ORF 1 (13%) but less than expected, one-third, possibly indicating that this reading frame is under represented in the library. Secondary point mutations, due to *cybC* DNA cassette insertion into *egfp*, sampling ORF 2 and 3, were identified in 27 (84%) of the non-fluorescent variants. In contrast to the fluorescent variants much lower percentages of the substitution mutations were conserved or semi conserved (29% and 21% respectively). A high proportion of the substitution mutations (26%) involved the substitution of proline residues. Substitution of two solvent exposed proline residues in three different variants (NFCG4, NFCG15, NFCG21) to charged or polar residues were identified and six different variants had residues replaced by a proline. Due to the restricted backbone torsion angles inherent to proline, efficient and correct folding of the chimeric proteins may be affected.

One unique insertion site was sampled by both a fluorescent variant (CG5) and a non-fluorescent variant (NFCG17) (Table 5.1 and 5.2) the only difference being a single substitution mutation, S30Y, in the non-fluorescent variant. The side chain of

S30 is solvent exposed therefore substitution for the bulky side chain of tyrosine may hinder correct folding and potentially promote aggregation during folding.

DNA sequence analysis revealed that CG6 was the only variant not to contain the full GlyGlySer linker at the C-terminus of cyt b_{562}; the C–terminal Gly-Ser are not present leaving only a single glycine as the C-terminal linker between cyt b_{562} and EGFP.

5.2.4. Spectral characterisation of selected cyt b_{562}-EGFP integral fusion variants.

Cyt b_{562}-EGFP domain insertion variants CG1-CG12 and CG15 (Fig 5.5 and Table 5.1) were selected for more detailed spectral characterisation to investigate their fluorescence and absorbance properties. The properties of EGFP fluorescence and cyt b_{562} absorbance spectra are related to their function and any potential coupling between the two domains.

To characterise the spectral properties of EGFP and the selected cyt b_{562}-EGFP chimeras, protein was produced in *E. coli* TUNER™ (DE3) cell cultures grown in M9 minimal media (Section 2.5.4). A limited growth media is required for expression of the cyt b_{562}-EGFP chimeras to produce protein in the apo form, as when produced by *E. coli* in rich media (LB) a high proportion of the protein is in the holo haem bound form, which hinders haem titration studies (Section 2.6.1.2). Excitation and emission spectra of the relevant proteins in crude cell lysates were then measured to identify if cyt b_{562} domain insertions have altered the spectral properties of the EGFP domain.

The majority of the cyt b_{562}-EGFP variants had excitation and emission spectra comparable to that of EGFP with excitation and emission maxima of ~488 nm and ~511 nm respectively (Figure 5.6 and Table 5.3). Cyt b_{562}-EGFP chimeras CG8 and CG15 both exhibited altered spectral characteristics with excitation maxima (λ_{ex}) at two different wavelengths: ~400 nm and ~488 nm (Fig 5.7). They also had a slightly red shifted emission maxima of ~ 513 nm (Fig 5.7 and Table 5.3), when excited at 400 nm. The altered spectral properties are due to a change in the proportion of the chromophore being in the protonated state. The 400 nm excitation maximum is thought to be due to the phenol group of Tyr66 being protonated whilst the 488 nm excitation maximum is thought to be due to deprotonation of the phenol group (Fig 5.7) [62].

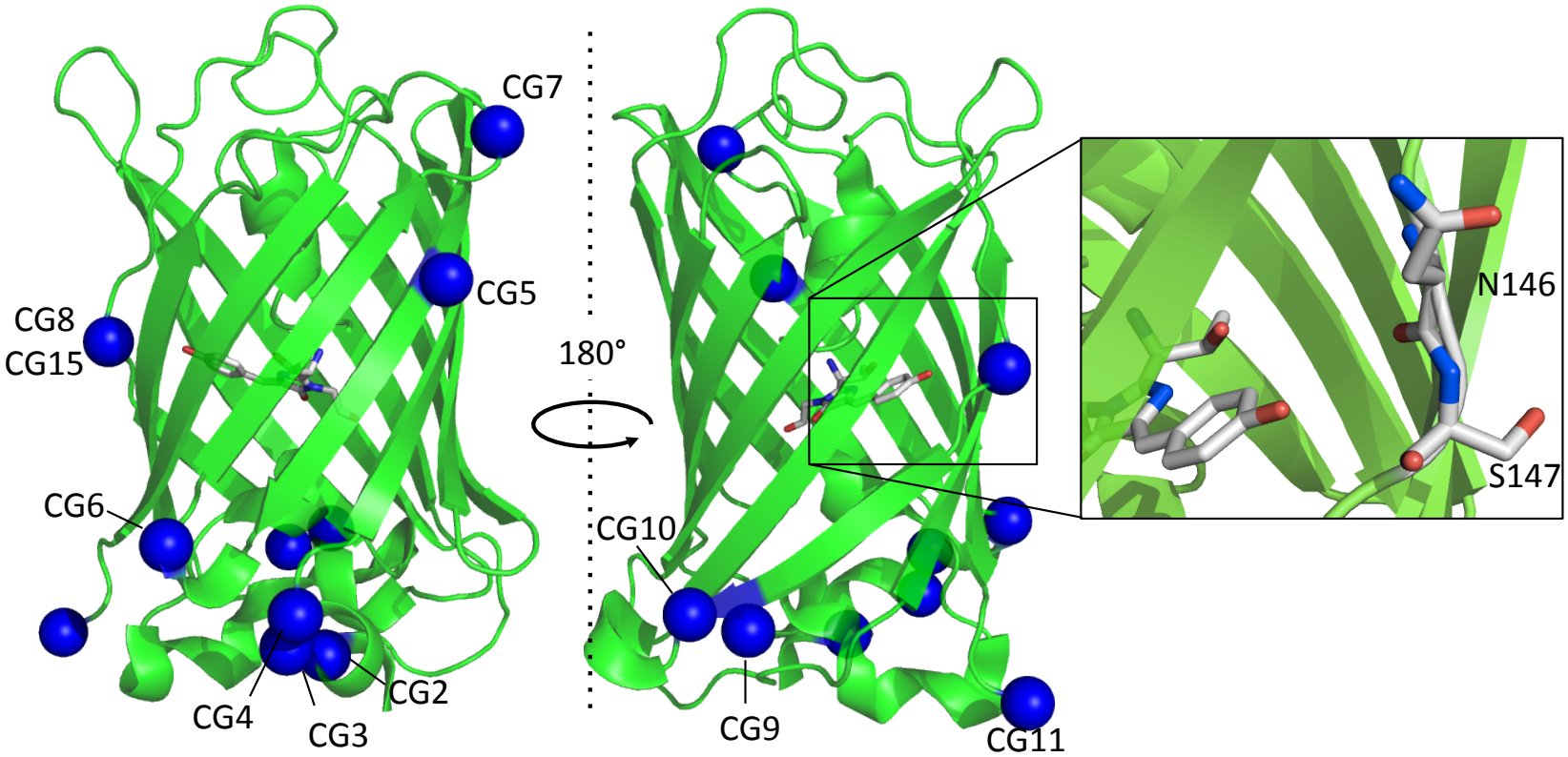


Fig 5.5 Selected cyt b insertion positions within EGFP. Selected positions within the tertiary structure of EGFP (green cartoon) tolerant to cyt b_{562} insertion (blue spheres) are highlighted. Variants CG8 and CG15 have cyt b insertions tolerated between residues N146 and S147, in very close proximity to the chromophore (sticks in CPK).

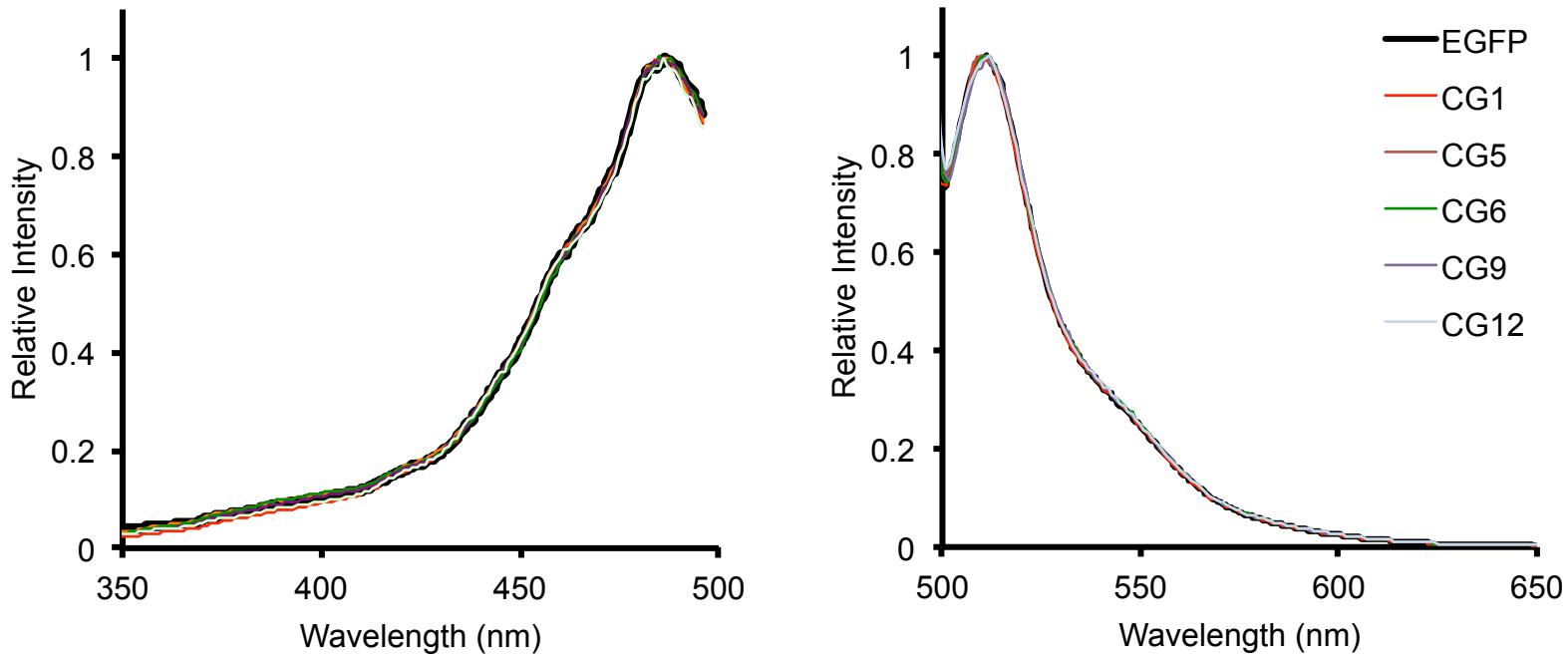


Fig 5.6 Excitation and emission spectra for EGFP and selected cyt b_{562}-EGFP chimeras. Excitation spectra (left panel) were measure while monitoring emission at 510 nm and emission spectra (right panel) were measure after excitation at 488 nm.

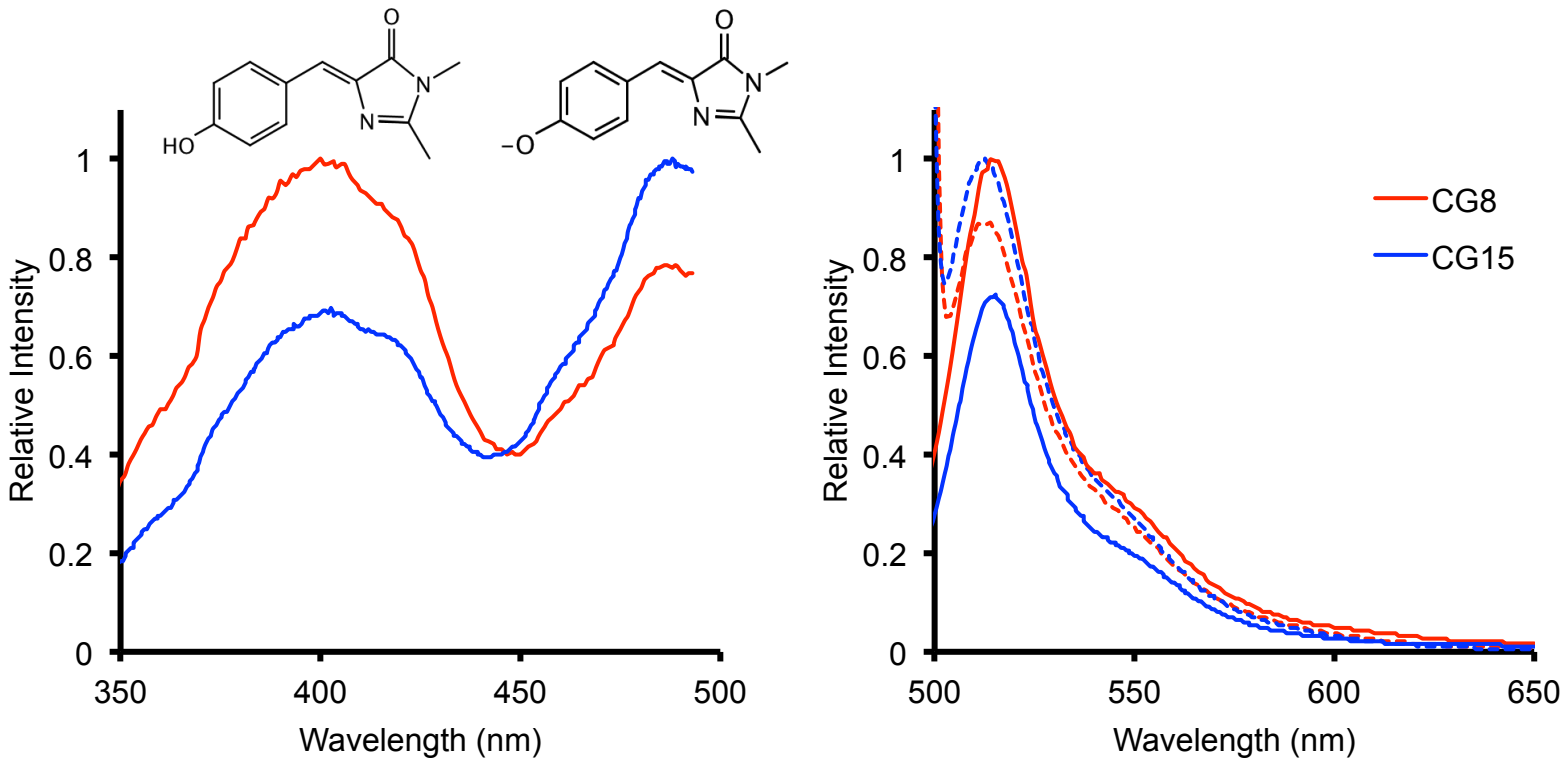


Fig 5.7 Excitation and emission spectra for cyt b_{562}-EGFP chimeras CG8 and CG15. Excitation spectra (left panel) were measure while monitoring emission at 510 nm and emission spectra (right panel) were measure after excitation at either 400 nm (solid lines) or 488 nm (dotted lines). Chemical structures for the chromophore in the protonated and deprotonated state are shown above the excitation maximum they are thought to be responsible for.

In order to establish if cyt b_{562} domain insertion had affected the fluorescent properties (molar extinction coefficient, quantum yield) of EGFP, measurements on purified protein samples were performed. EGFP and the cyt b_{562}-EGFP integral fusion scaffolds were produced in *E. coli* TUNER™ cells as described above and in section 2.5.4. The cell cultures containing the different cyt b_{562}-EGFP integral fusion scaffolds were lysed by French press and clarified by centrifugation (Section 2.5.5). The cyt b_{562}-EGFP integral fusion scaffolds were deemed to be in the soluble fraction after cell lysis by their observable green colour. The samples were then purified using a combination of techniques; (1) fractionation by ammonium sulphate precipitation, (2) anion exchange chromatography, (3) gel filtration and (4) another round of anion exchange. A detailed description of each purification step has been covered in section 2.5.8.

Molar absorption extinction coefficients and quantum yields were determined for selected cyt b_{562}-EGFP chimeras (Table 5.3) to see if cyt b_{562} domain insertion had altered the EGFP derived spectral properties. The molar absorption extinction coefficients for the majority of the variants were comparable to that of EGFP (55,000 $M^{-1}cm^{-1}$) (Table 5.3). Both CG6 and CG12 showed a small decrease in molar extinction coefficient to ~46,500 $M^{-1}cm^{-1}$ and ~46,600 $M^{-1}cm^{-1}$ respectively (Table 5.3). Given that the molar extinction coefficients for these variants are very similar to that of EGFP indicates that the cyt b_{562} domain insertion has not affected the overall structure of EGFP.

In contrast both CG8 and CG15 showed decreased molar absorption extinction coefficients at ~400 and ~ 488 nm (Table 5.3). The molar absorption extinction coefficients for CG8 are ~27,100 $M^{-1}cm^{-1}$ and ~18,900 $M^{-1}cm^{-1}$ at 400 or 488 nm respectively whilst for CG15 they are ~26,500 $M^{-1}cm^{-1}$ and ~19,800 $M^{-1}cm^{-1}$. It is not surprising that the molar absorption extinction coefficients have been affected given that the cyt b_{562} domain is inserted between residues N146 and S147 of EGFP, adjacent to the phenolate group of the chromophore (Fig 5.5).

For the majority of the cyt b_{562}-EGFP integral fusion scaffolds the quantum yield was very similar to that of EGFP (0.6) (Table 5.3). Therefore, brightness values for CG1, CG4, CG5 and CG10 were comparable to that of EGFP (33,000 $M^{-1}cm^{-1}$) showing that domain insertion had not affected the fluorescence properties in these

Table 5.3 Spectral characteristics for EGFP and cyt b_{562}-EGFP chimeras8

Variant	λ_{ex} (nm)[a]	λ_{em} (nm)[a]	ε ($M^{-1}cm^{-1}$)[b]	ϕ[c]	Brightness ($M^{-1}cm^{-1}$)[d]	Fluorescence lifetime (ns)[e]		Fold difference[f]	Holo-chimera λ_{max} (Ox/Red) (nm)[g]	$Haem^{Ox}$ Kd (nM)[h]
						τ_{apo}	τ_{holo}			
eGFP	488	510	55000	0.6	33000	2.47	2.48	-	-	-
CG1	487	508	54550±3950	0.61	33275	2.38	1.64	1.5	417/426	11.5±1.4
CG2	486	510	-	-	-	-	-	2.5	-	105.7±16.5
CG3	487	510	-	-	-	-	-	1.3	-	17.0±2.8
CG4	486	510	54290±3180	0.59	32030	2.5	1.57	1.9	418/426	16.5±3.2
CG5	487	509	54350±5990	0.63	34240	2.51	1.67	1.8	418/426	49.3±11.7
CG6	488	510	46510±2790	0.64	29770	2.47	-	>25	422/426	11.0±1.8
CG8	400 (487)	514 (513)	27140±3310 (18920±5210)	0.34	6820	-	-	1.7	418/426	38.2±5.5
CG9	486	510	-	-	-	-	-	1.5	-	13.7±0.5
CG10	486	510	53700±2190	0.59	31680	2.37	0.6	1.6	418/426	12.1±0.5
CG11	486	511	-	-	-	-	-	1.4	-	8.6±0.2
CG12	486	511	46570±2590	0.61	28410	2.39	1.8	1.7	415/426	11.8±1.0
CG15	488 (402)	512 (515)	19810±2980 (26537±1750)	0.37	8050	-	-	1.7 -	417/426	-
Cyt b_{562}	-	-	-	-	-	-	-	-	418/427[141]	~10[76]

[a] λ_{ex} and λ_{em} determined from fluorescence spectra. Red values in brackets refer to the less intense maxima

[b] Extinction coefficient determined from minimum of 3 absorbance measurements at 488 nm, error is a single standard deviation. Red values in brackets are extinction coefficients determined at the corresponding λ_{ex} in brackets.

[c] Quantum yield determined from integrated fluorescence emission after excitation at 488 nm against a fluorescein standard.

[d] Brightness = extinction coefficient x quantum yield,

[e] Fluorescence lifetimes are mean values with errors calculated from the standard deviation of 3 measurements where τ_{apo} and τ_{holo} are the fluorescence lifetimes in the absence and presence of equimolar haem respectively

[f] Fold difference in fluorescence emission at 510 nm after excitation at 488 nm with equimolar concentrations of protein:haem under oxidising or reducing conditions.

[g] λ_{max} values determined from UV-visible absorption spectra

[h] K_d determined from haem mediated fluorescence quenching data under oxidising conditions

variants. CG6 and CG12 had slightly reduced brightness (29,770 and 28410 $M^{-1}cm^{-1}$ respectively). due to a decrease in the EGFP chromophore molar absorbance coefficient (Table 5.3).

The quantum yields determined at 488 nm for variants CG8 and CG15 were decreased by almost two-fold with respect to EGFP (Table 5.3). This is probably due to disruption of the local structure around residues N146 and S147 in the EGFP domain due to cyt b_{562} domain insertion, consequently allowing solvent to access the interior of the β-barrel partially quenching the excited state chromophore. Therefore, variants CG8 and CG15 have much lower brightness compared to that of EGFP (6820 and 8050 $M^{-1}cm^{-1}$ respectively) due to greatly decreased molar absorption extinction coefficients and reduced quantum yields upon excitation at 488 nm (Table 5.3). In order to establish the brightness for CG8 and CG15 at 400 nm, quantum yield determination would be required against a 9-aminoacridine standard, which absorbs maximally at 400 nm, rather than the fluorescein standard used here.

5.2.3 Confirming cyt b_{562} domain integrity in cyt b_{562}-EGFP integral fusion scaffolds

The observed fluorescence for the selected cyt b_{562}-EGFP integral fusion scaffolds indicates that the EGFP domain was correctly folded; the β-barrel structure is a prerequisite for fluorescence [142]. However, to determine if the inserted cyt b_{562} domain was correctly folded and if insertion into EGFP alters the haem binding properties, UV-Visible spectroscopy was used to analyse the variants in the presence and absence of haem. Wild type cyt b_{562} has characteristic UV-Visible absorption spectra when haem is bound under oxidising conditions with a λ_{max} of 418 nm (Table 5.3). When haem is bound to cyt b_{562} under reducing conditions there is an increased absorption and red shifted λ_{max} to 427 nm (Table 5.3) with two smaller absorption maxima at 531 and 562 nm [141].

Insertion of cyt b_{562} into EGFP has not had an effect on the cyt b_{562} domains ability to bind haem and all variants display classic absorption characteristics, for both domains, when under oxidising and reducing conditions (Table 5.3). The UV-visible absporption spectrum for EGFP shows the classic absorbance maximum at 488 nm (Fig 5.8). Addition of haem to EGFP at an equimolar concentration under oxidising

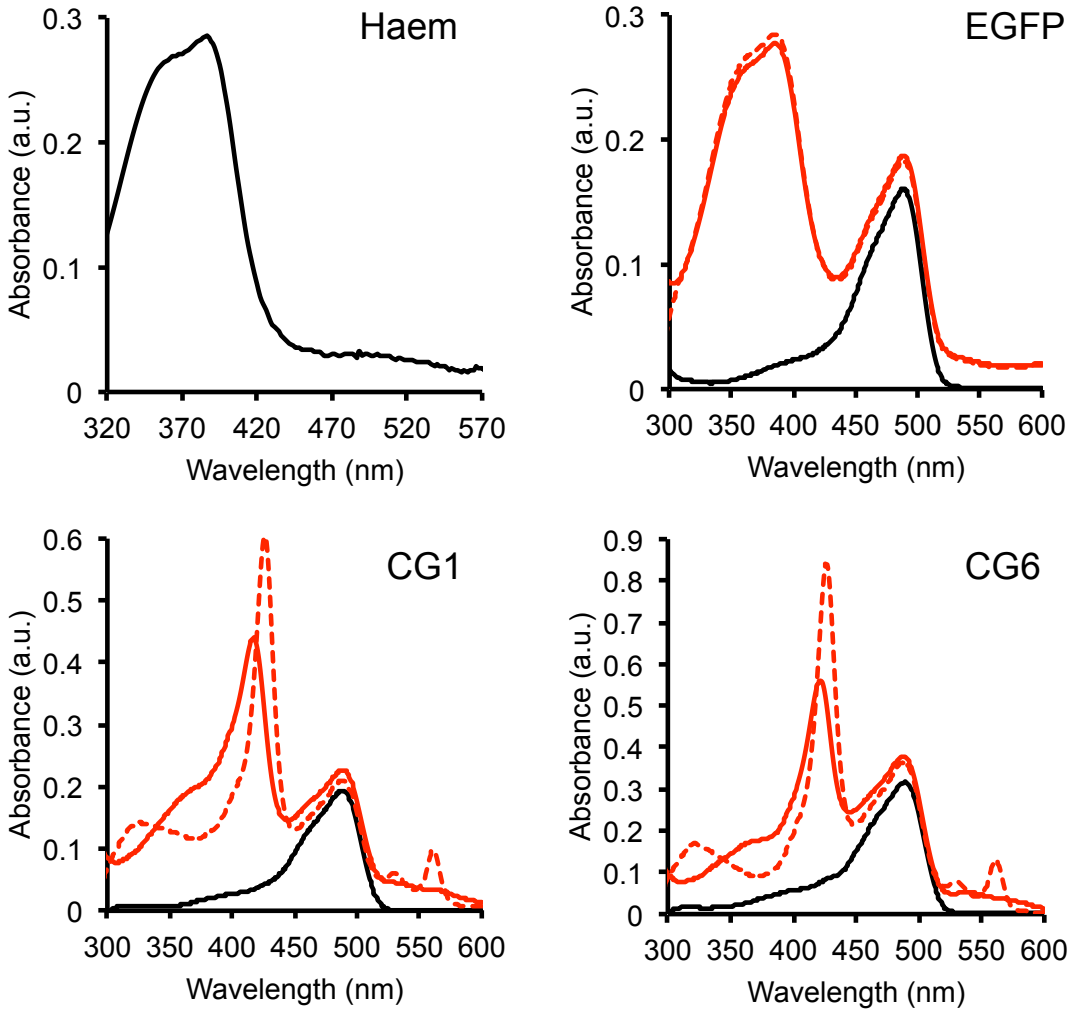


Fig 5.8 UV-visible absorption spectra for EGFP and cyt b_{562}-EGFP chimeras. Absorption spectra were measured for 5 μM of purified apo-protein samples (black lines) and in the presence of equimolar haem under oxidising (solid red line) or reducing (dotted red line) conditions. Titration of haem with EGFP presents the two characteristic absorption spectra: EGFP and free haem in solution. The terminal fusion, between cyt b_{562} and EGFP (CG1) and the integral fusion CG6 exhibit characteristic absorption spectra for EGFP and for haem bound to cyt b_{562} under reducing and oxidising conditions.

and reducing conditions presented a mixture of the EGFP spectra and the absorbance spectra for free haem in solution, showing that haem does not bind to EGFP alone (Fig 5.8).

All of the selected cyt b_{562}-EGFP chimeras bind haem as is evident from the characteristic absorption spectra of haem bound to cytochrome (Fig 5.8) with λ_{max} values of ~417 nm and ~427 nm for oxidising and reducing conditions respectively (Table 5.3). Variant CG6 had a slightly red shifted λ_{max}, under oxidising conditions, of 5 nm with respect to wild-type cyt b_{562} and the other cyt b_{562}-EGFP chimeras (Table 5.3, Fig 5.8). This may imply that the local environment around the haem is altered in CG6 with respect to wild type cyt b_{562} and the other cyt b_{562}-EGFP chimeras.

5.2.4 Effect of haem binding on EGFP derived fluorescence.

Haem binding to cyt b_{562} should result in quenching of EGFP fluorescence through energy transfer [82]. To investigate the extent and efficiency of haem binding to the cyt b_{562} domain on EGFP fluorescence, haem was titrated into crude cell lysates containing the chimeric proteins and the fluorescence monitored at 511 nm on excitation at 488 nm. Titration of haem resulted in a general decrease in fluorescence intensity (Fig 5.9). However, there were differences in the extent of decrease between different variants and whether reducing or oxidising conditions were used.

The extent of quenching of EGFP fluorescence on haem titration was dependent on the chimera (Fig 5.9). Under reducing conditions, the terminal fusion chimeras CG1 and CG12 retained 37% and 28% of their fluorescence in a five-fold excess of haem. Thus, total energy transfer was not possible in these chimeras, which represent the traditional approach to the construction of GFP fusions. CG11, with cyt b_{562} inserted close to the EGFP C-terminus, showed similar results to the terminal fusion proteins with its fluorescence maximally quenched by 58% when 15 nM haem was added under reducing conditions (Fig 5.9).

CG7 has cyt b_{562} inserted within a loop connecting β-strand 3 to the central α-helix containing the chromophore and also showed similar results to the terminal fusion proteins. In the presence of ~1.5 fold more haem than the terminal fusions CG7

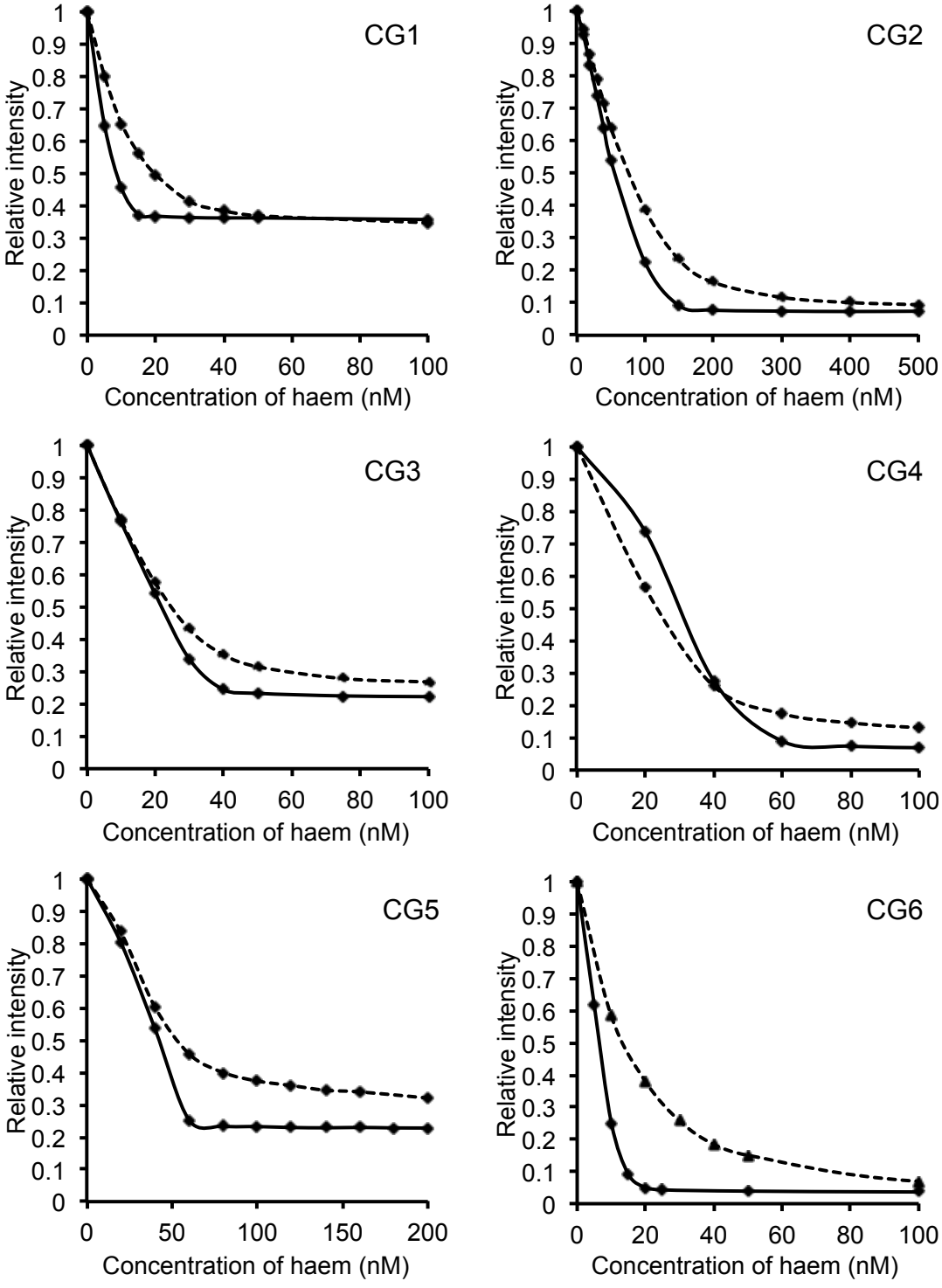


Fig 5.9 Haem-dependent fluorescence changes of cyt b_{562}-EGFP chimeras. Decrease in fluorescence on haem addition under reducing (solid line) and oxidising (dashed line) conditions. Haem mediated fluorescence quenching results for variants CG7 - CG12 are shown on the next page.

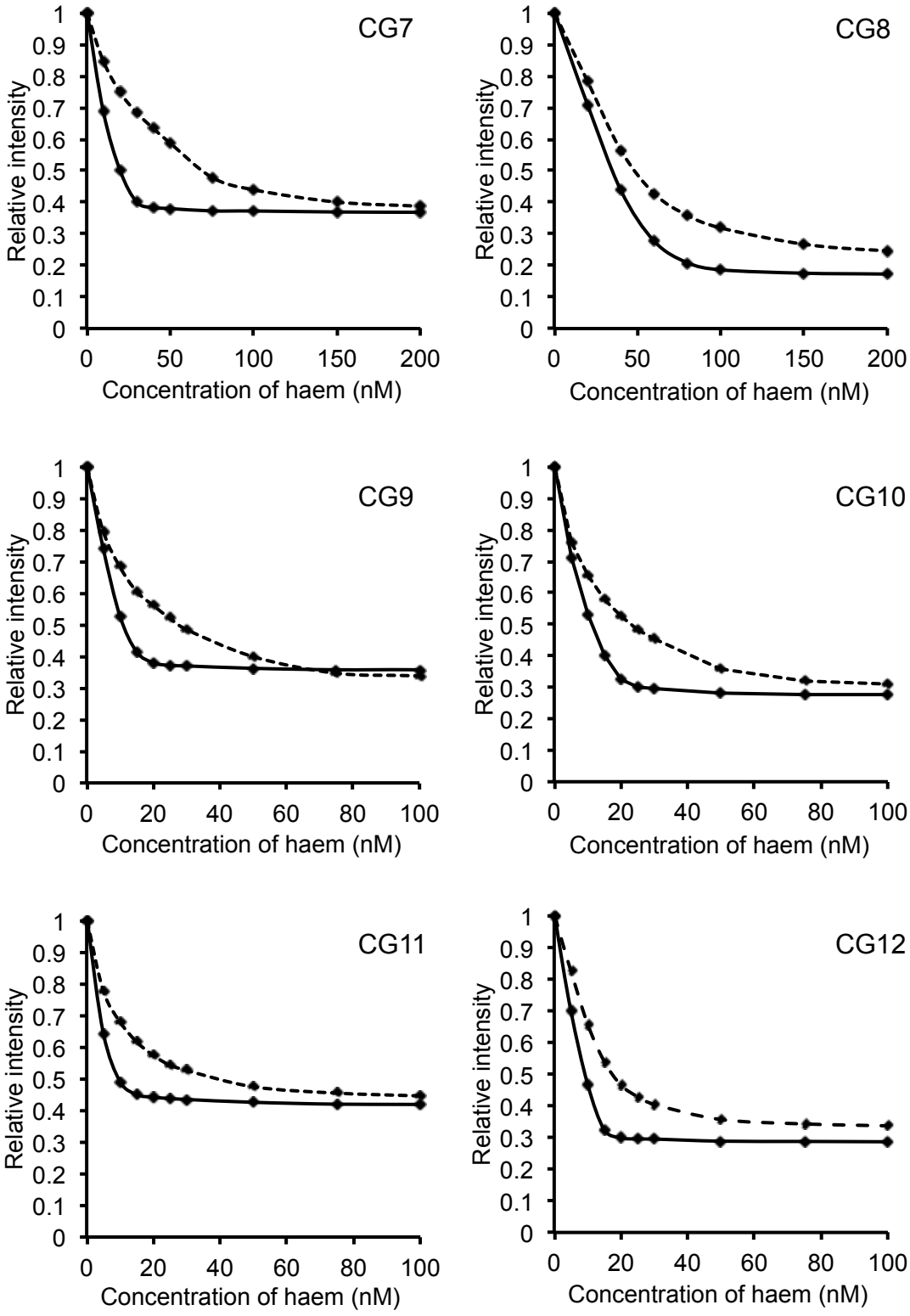


Fig 5.9 continued. Haem-dependent fluorescence changes of cyt b_{562}-EGFP chimeras. Decrease in fluorescence on haem addition under reducing (solid line) and oxidising (dashed line) conditions.

showed maximal quenching of 60% under reducing conditions (Fig 5.9). CG9 and CG10 have cyt b_{562} inserted only a single amino acid away from one another (Table 5.1) at the C-terminal end of β-strand 7. Both were maximally quenched in the presence of similar concentrations of reduced haem as the terminal fusion proteins, 20 nM and 25 nM for CG9 and CG10 respectively, retaining 38% and 30% of their fluorescence respectively (Fig 5.9).

Variants CG3 and CG5 were maximally quenched by 76% and 75% respectively although required two-fold and three-fold more haem to achieve full quenching with respect to the terminal fusions CG1 and CG12, implying that the haem binding characteristics for these variants have been altered by cyt b_{562} domain insertion into EGFP (Fig 5.9).

Variant CG6 showed the most promising haem mediated switching characteristics with only 3% of its total fluorescence remaining on the addition of 20 nM reduced haem (Fig 5.9). CG6 reaches maximal fluorescence quenching at similar concentrations of reduced haem to that of the terminal fusions, implying that insertion of cyt b_{562} at Y39 of EGFP has not affected haem-binding affinities.

The fluorescence of variant CG2, with cyt b_{562} positioned within the first α-helix of EGFP (Fig 5.5), was also quenched to a much greater extent than the other integral fusion variants. Only 7% of fluorescence intensity remained in the presence of excess haem (Fig 5.9). However, ten-fold more haem was required to reach this maximal quenching with respect to CG1. CG4 was another variant that displayed a far greater extent of fluorescence quenching than the other variants, with only 7% of fluorescence intensity observed in excess haem (Fig 5.9). However, as with CG2 a higher concentration (three-fold) of haem was required to reach full quenching. This suggests that haem binding to the cyt b_{562} domain or the fluorescent properties of the EGFP domain have been altered by domain insertion and are compromised in CG2 and CG4 compared to CG6 and the terminal fusion variants.

5.2.5 Redox mediated fluorescence quenching.

All variants exhibited a difference in concentration-dependence haem mediated quenching under reducing and oxidising conditions (Fig 5.9). Under reducing conditions, the decrease in fluorescence at 511 nm on increasing haem concentration did not follow that of a typical equilibrium binding curve profile (Fig

5.9). There is a rapid decrease in fluorescence during titration of reduced haem with the chimeras exhibiting a sharp transition at the point where it is estimated that protein and haem concentration are at a 1:1 ratio (Fig 5.10 a). At this ratio all binding sites are considered to be occupied by haem. Given that all protein samples were normalised to a starting fluorescence intensity of 100 a.u, confirmed to be ~20 nM for the majority of variants, the high affinity of reduced haem for the wild type cyt b_{562} (~10 pM) appears to have been retained by most of the cyt b_{562}-EGFP integral fusion scaffolds.

Under oxidising conditions the haem mediated quenching profiles resemble those under equilibrium (Fig 5.9). As mentioned previously, observation that some of the variants required higher concentrations of haem to reach maximal quenching implies that the haem binding characteristics of these variants have been compromised by insertion of cyt b_{562} into EGFP. This was confirmed by calculating K_d values for oxidised haem binding to the cyt b_{562}-EGFP variants from the fluorescence titration data (Table 5.3) (Section 2.6.1.13). CG1, CG12 and CG6 all had comparable K_d values (~11 nM) to that of wild type cyt b_{562} (~10 nM) confirming these variants had maintained the high affinity for oxidised haem.

Variants CG5 and CG8 showed a marked decrease in haem affinity for the cyt b_{562} as was apparent from their K_d values of ~49 nM and ~38 nM respectively. In particular variant CG2 exhibited an almost ten-fold increase in its K_d value (~105 nM) with respect to wt cyt b_{562} (Table 5.3).

Comparison of the remaining fluorescence intensity of each variant under reducing and oxidising conditions, when the apparent [protein]:[haem] = 1:1 (Fig 5.10 a), was performed to assess the potential for the chimeras to act as redox sensors. All variants showed between a ~1.5 to ~1.9 fold difference in the levels of remaining fluorescence (Fig 5.10 b, Table 5.3), apart from variant CG6 which showed up to a >25-fold difference in remaining fluorescence between oxidising and reducing conditions.

Variants CG2, CG4 and CG6 all have insertion positions clustered together in the tertiary structure of EGFP (Fig 5.5) and all result in almost 100% quenching of fluorescence when titrated with haem to saturation. However, the near 100% quenching upon haem binding, the >25 fold difference in remaining fluorescence between oxidising and reducing conditions, whilst maintaining similar haem binding

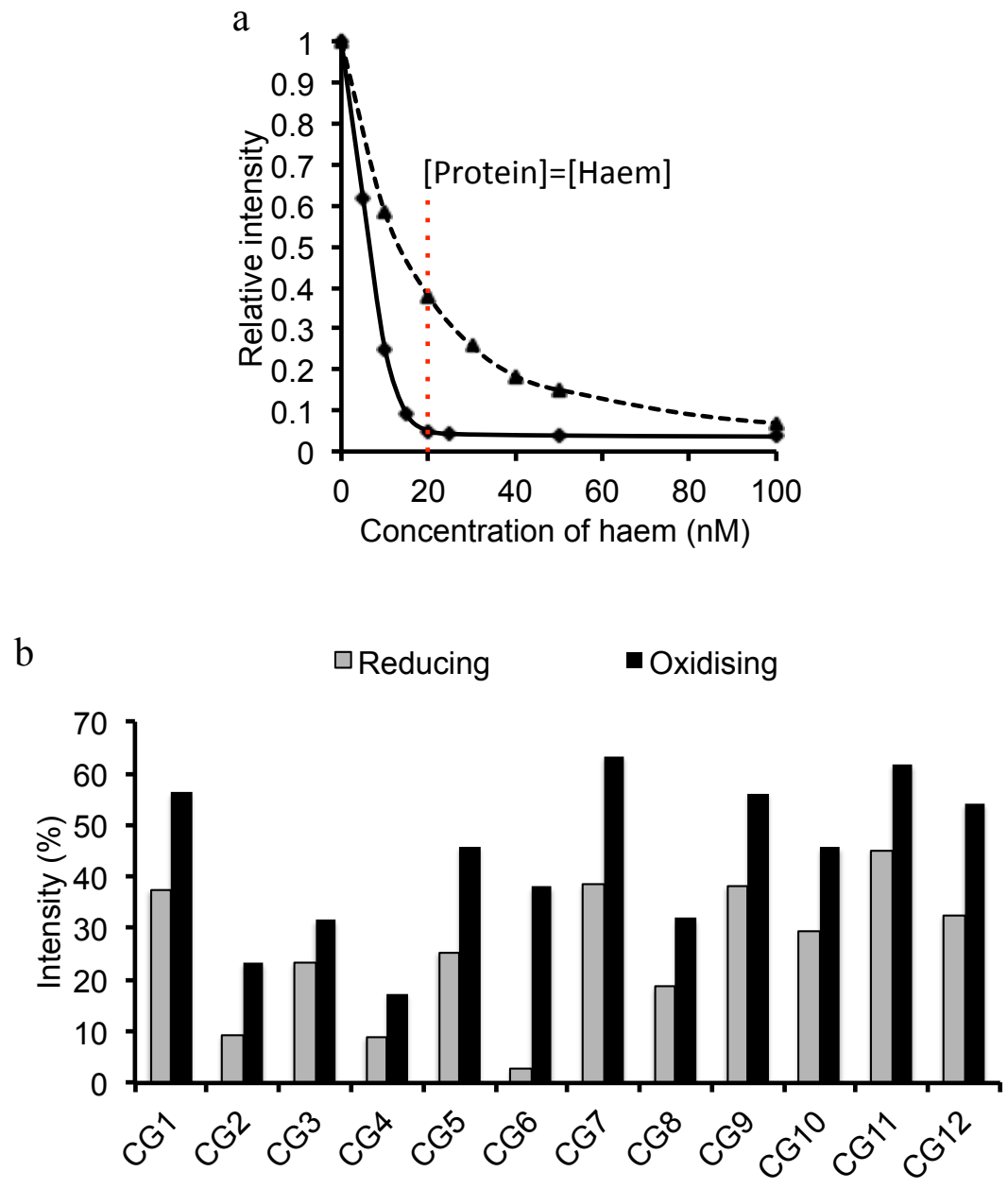


Fig 5.10 Differential haem mediated quenching under oxidising and reducing conditions. **a.** Comparison between the remaining fluorescence under oxidising and reducing conditions was carried out when an apparent [protein]:[haem] of 1:1 was reached. The example titration curves used are for the cyt b_{562}-EGFP chimera CG6 **b.** Remaining fluorescence intensity under oxidising and reducing conditions for all cyt b_{562}-EGFP variants.

affinities to wt cyt b_{562} makes CG6 an attractive variant for the use as a tailor made energy transfer component with the potential to act as a redox sensor.

5.2.6 Oxidant induced fluorescence switching.

Given that variant CG6 exhibited the largest difference (>25 fold) between fluorescence under oxidising and reducing conditions (Fig 5.10 b, Table 5.3), it could be considered as a sensor for redox conditions. Therefore, the ability of CG6 to monitor changes to redox conditions was assessed *in vitro*. A slight excess of haem (30 nM) was added to CG6 (20 nM) under reducing conditions (1 mM ascorbic acid) inducing rapid almost total quenching of fluorescence (Fig 5.11 a). Different oxidants were then applied to holo-CG6 to assess its ability to sense changes in redox conditions by haem dissociation and monitoring gain in signal through relieving fluorescence quenching (Fig 5.11 a and b). KNO_3 or NaOCl had no effect on fluorescence intensity (Fig 5.11 b) but the common biologically important reactive oxygen species (ROS), H_2O_2, relieved quenching eventually restoring the original level of fluorescence (Fig 5.11 a).

When no oxidant was added the fluorescence intensity essentially remained zero for CG6 (Fig 5.11 a). As the other oxidants used, KNO_3 and NaOCl, did not relieve quenching, CG6 could potentially act as a H_2O_2 specific sensor. UV-Visible absorption spectra confirmed that haem was bound to CG6 before the addition of H_2O_2 (Fig 5.11 c). Absorption spectra measured 2 hrs and 20 hrs after the addition of H_2O_2 confirmed the dissociation of haem from CG6 (Fig 5.11 c). As haem showed little dissociation from CG6 in the presence of KNO_3 or NaOCl, H_2O_2 may be exerting its effect by chemical modification of haem rather than altering the iron moiety redox state. Oxidative cleavage of haem has been seen in previous studies [77].

In comparison, CG1 only lost 54 % of fluorescence on haem addition in the presence of 1 mM ascorbate. On the addition of H_2O_2 a rapid recovery of fluorescence (Fig 5.11 d) was observed. Fluorescence data obtained from the oxidant induced fluorescence switching of CG1 and CG6 on the addition of H_2O_2 were fit to a single exponential function (Section 2.6.1.3) to determine the rate of fluorescence gain. CG6 exhibited a rate constant for the gain in fluorescence signal after the addition of H_2O_2 of 0.52 min^{-1} where as CG1 showed a rate constant >14-fold faster of 7.47 min^{-1}.

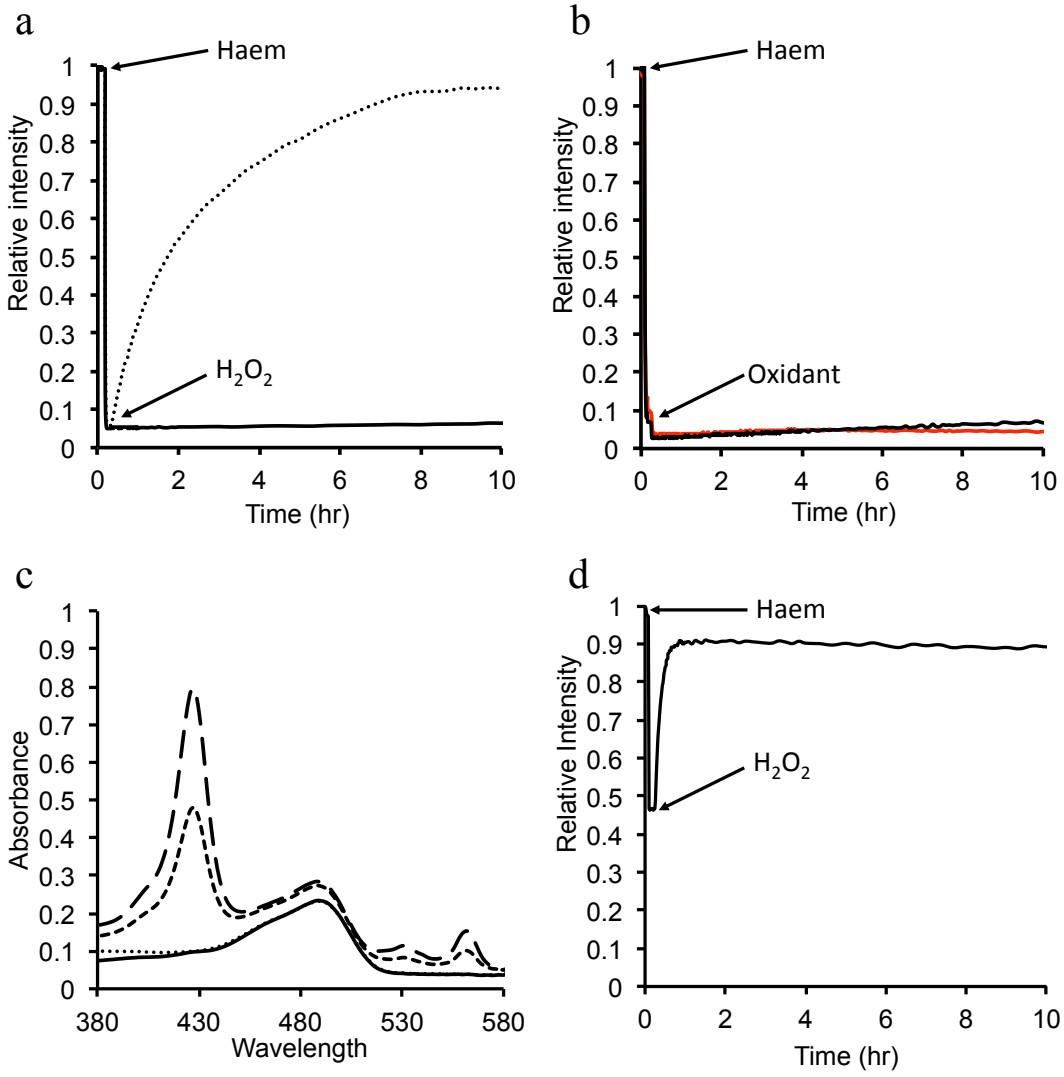


Fig 5.11 Oxidant induced fluorescent switching. **a,** Addition of a slight excess of haem (30 nM) to CG6 (20 nM) under reducing conditions (1 mM ascorbic acid) induced rapid fluorescence quenching followed by the addition of 0.02% (v/v) H_2O_2 to induce oxidising conditions (dotted line). CG6 fluorescence in the absence of added oxidant added is shown as a black line. **b,** addition of oxidants, KNO_3 (black line) or NaOCl (red line). **c,** UV-visible absorption spectra for CG6 measured before (solid line) and 10 minutes after the addition of haem (long dashed line). Spectra were then measured 2 hrs (short dashed line) and 20 hrs (dotted line) after the addition of H_2O_2. **d,** Addition of a slight excess of haem (30 nM) to CG1 (20 nM) under reducing conditions (1 mM ascorbic acid) induced rapid fluorescence quenching followed by the addition of 0.02% (v/v) H_2O_2 to induce oxidising conditions (dotted line).

This could be due to the different orientations of the cyt b_{562} domains in the CG1 terminal fusion and the CG6 integral fusion to EGFP; in the terminal fusion the two domains are probably highly dynamic with respect to one another allowing the solvent to freely access the bound haem moiety, where as in CG6 the cyt b_{562} domain may be positioned such that the bound haem moiety is less solvent exposed.

Control titrations of ascorbic acid or H_2O_2 with apo-CG6 were carried out to confirm that these reagents had no effect on CG6 fluorescence emission (Fig 5.12) and that the changes in fluorescence seen in the oxidant induced switching experiments was due to haem dissociation.

5.2.7 Effect of haem binding on EGFP derived fluorescence lifetimes

To further investigate the effects of haem binding on the EGFP derived fluorescence properties of the cyt b_{562}-EGFP integral fusion scaffolds, fluorescence lifetimes were measured in the absence and presence of equimolar concentrations of haem under reducing conditions. In the absence of haem, all cyt b_{562}-EGFP chimeras have comparable fluorescence lifetimes to that of EGFP (~2.47 ns) (Table 5.3). This implies that the cyt b_{562} domain insertions are well tolerated in EGFP and there has been little disruption of the chromophores local environment.

In the presence of haem, the fluorescence lifetime for EGFP was not altered (Fig 5.13), agreeing with previous results (Fig 5.8) and control titrations (Section 5.2.8, Fig 5.14) that haem does not bind to EGFP alone or quench EGFP fluorescence when in solution. Binding of haem to the cyt b_{562}-EGFP chimeras resulted in a decrease in fluorescence lifetime (Table 5.3 and Fig 5.13). The majority of the chimeras displayed ~1.6 fold decrease in their fluorescence lifetimes with CG10 showing a marked decrease of almost four-fold (Fig 5.13 and Table 5.3). The fluorescence lifetime for holo-CG6 could not be measured as the fluorescence signal is essentially zero in the presence of haem due to near 100% quenching.

However, variant CG4, which also showed >95% quenching on haem binding (Fig 5.9), only showed a decrease in its fluorescence lifetime of ~1.5 fold. These results agree with the notion that the haem binding properties of the cyt b_{562} domain in CG4 have been affected by insertion into EGFP, as a fluorescence lifetime was still detectable in the presence of equimolar reduced haem.

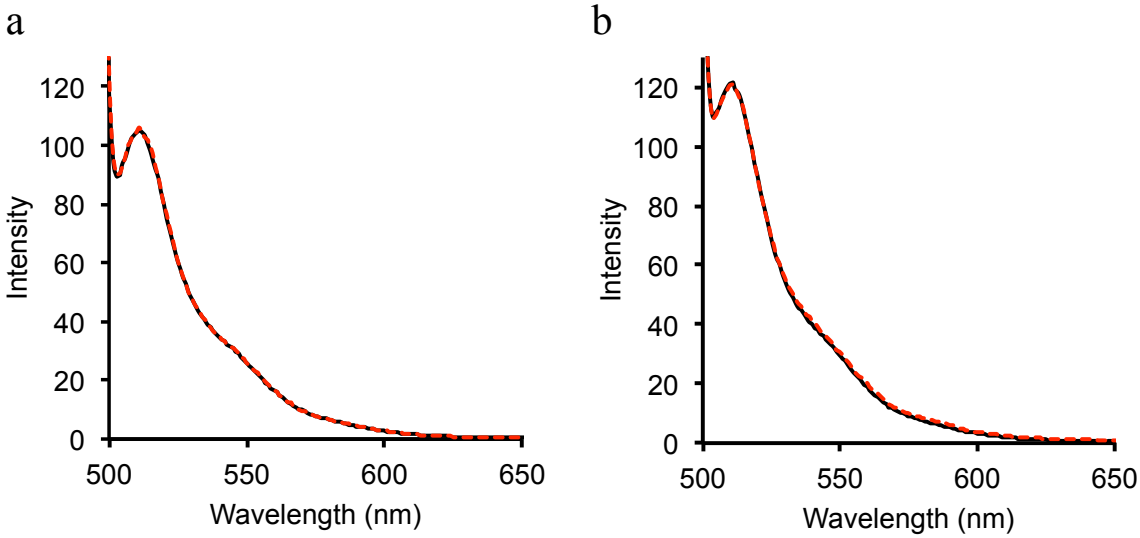


Fig 5.12. Effect of ascorbic acid or H_2O_2 on the fluorescence emission of apo CG6. a, CG6 fluorescence emission spectra in the absence (black line) and presence (red dashed line) of 1 mM ascorbic acid. **b,** CG6 fluorescence emission spectra in the absence (black line) and presence (red dashed line) of 0.02% (v/v) H_2O_2. Emission spectra were measured after excitation at 488 nm.

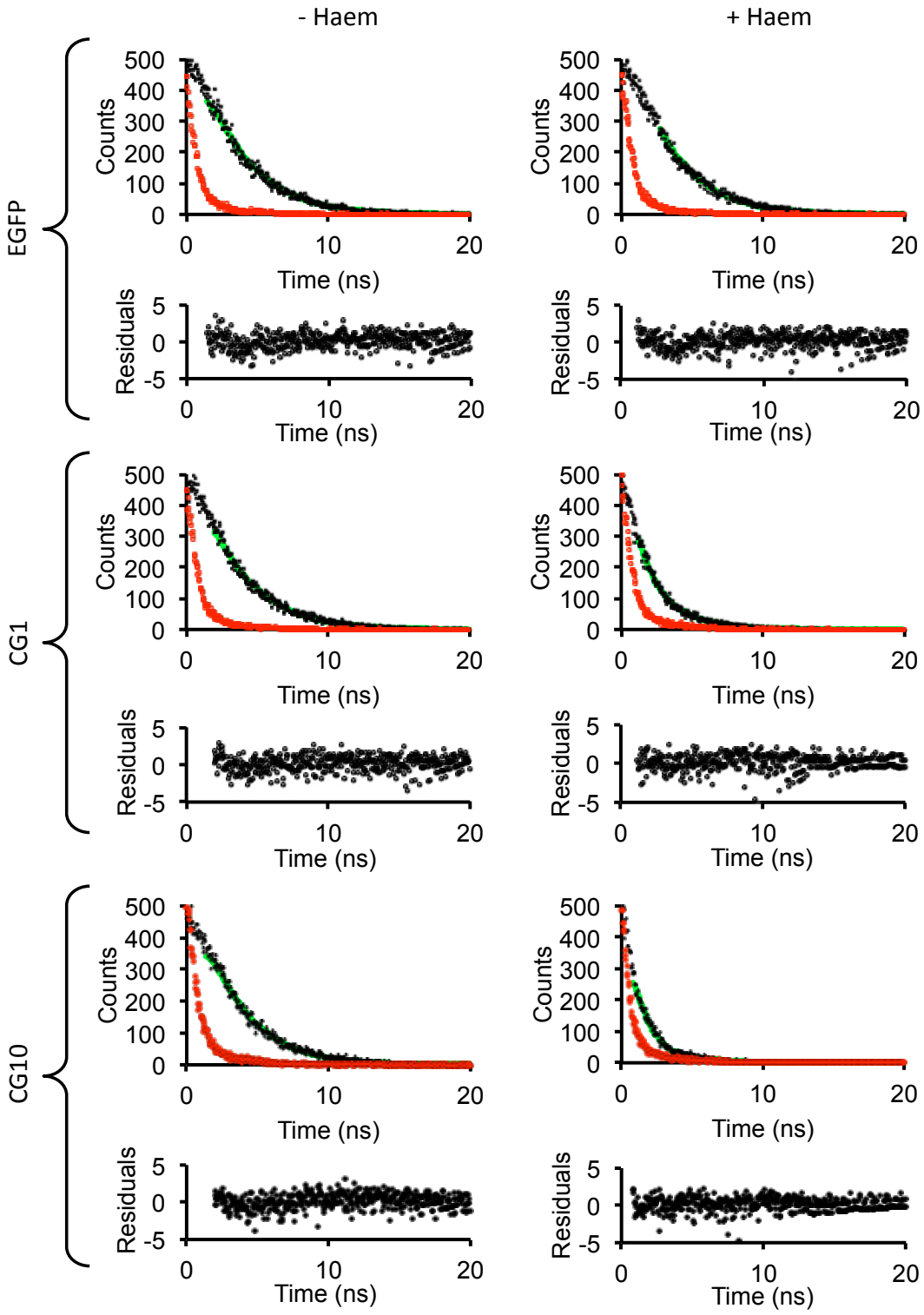


Fig 5.13 Fluorescence lifetime measurements of apo and holo cyt b_{562}-EGFP variants. Fluorescence lifetime measurements were made by fitting single exponential functions (green) to fluorescence lifetime data (black dots) after data reconvolution with the instrument response function (red dots). Fluorescence decays for EGFP, CG1 and CG10 are shown here in the absence and presence of equimolar concentrations of haem. Haem had no effect on the fluorescence lifetime of EGFP concurrent with previous haem effects on EGFP.

Given that CG10 had 30% remaining fluorescence in the presence of an excess of reduced haem (Fig 5.9), similar to that of other variants (Fig 5.9), the observation that its fluorescence lifetime decreases four-fold with equimolar reduced haem (Table 5.3 and Fig 5.13) implies there may be a structural change upon haem binding coupled to the quenching effects.

The observation that CG6 fluorescence intensity is quenched by almost 100% in the presence of equimolar concentrations of reduced haem with an undetectable fluorescence lifetime, under the same conditions, identifies it as an artificial protein scaffold capable of 100% energy transfer between its two chromophores. This artificial protein scaffold exhibits similar energy transfer efficiencies to naturally occurring light harvesting systems.

5.2.8 EGFP control titrations.

To show that the observed changes in fluorescence seen in all titration experiments are due to haem binding to the cyt b_{562} domains of the chimeric proteins and not any other reagents, control titrations were performed with EGFP. The addition of haem to wild-type EGFP resulted in no change in fluorescence (Fig 5.14 a). Addition of NaOH (used to dissolve haem) (Fig 5.14 b), KNO_3 (Fig 5.14 c) or ascorbic acid (Fig 5.14 e) (used to give oxidising or reducing conditions, respectively) to EGFP resulted in no change in fluorescence. Therefore any changes seen in chimera fluorescence can be attributed to haem binding to the cyt b_{562} domains.

5.3 Discussion.

Integral to synthetic biology is the construction of artificial protein scaffolds for the use as genetically encoded biomolecular switches [129, 130]. The functional coupling of two disparate proteins would be most advantageous for the creation of a diverse number of artificial protein scaffolds for uses in natural and synthetic applications [132-134]. Biomolecular switches are sampled by nature during evolution, quite often with integral domain fusion architecture (~9% of known structures in the PDB) [24, 25].

Insertion of one protein domain into another creates an intimate structural linkage by which changes in one domain can be communicated through to another.

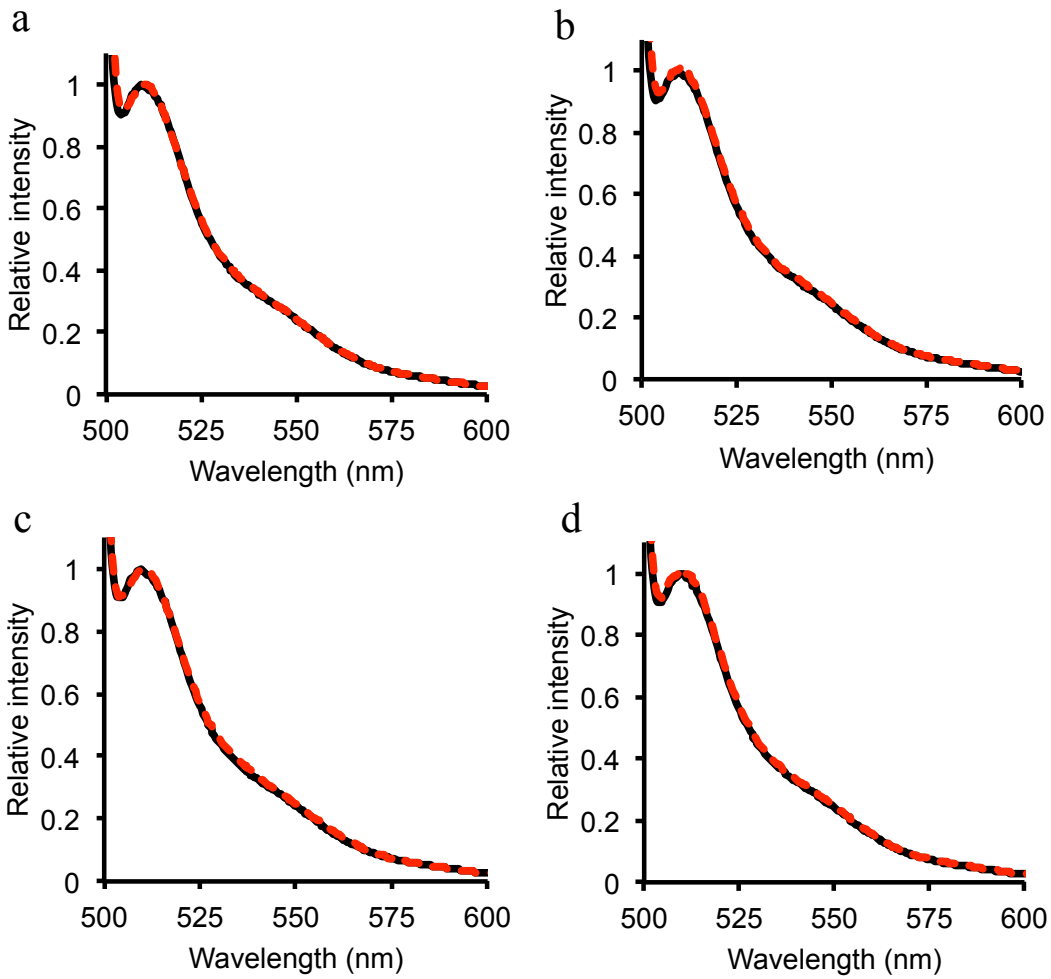


Fig 5.14. Emission spectra for EGFP under various conditions. Emission spectra for EGFP were measured in the absence (black line) and presence (red dashed line) of a, haem (1 uM), b, NaOH (50 uM), c, ascorbic acid (1 mM) or d, KNO_3 (1 mM) to confirm any changes in fluorescence seen during haem titration is due to haem binding to the cyt b_{562} domains and not due to other reagents used in the experiments. Fluorescence was monitored at 510 nm after excitation at 488 nm for EGFP alone (black line) or in the presence of reagent (red line).

The nature of integral fusion architecture also can fix the position of one domain with respect to another in contrast to domains terminally fused together, which have the potential to move and rotate freely with respect to one another. The process of identifying sites within a target protein that will not only tolerate the insertion of another protein domain but also will functionally link the two is very difficult.

Rational design of new protein scaffolds by domain insertion has seen some success but with resulting chimeras exhibiting only modest switching magnitudes [32, 34, 35]. Directed evolution approaches to creating artificial protein scaffolds allows for chimeras with desired characteristics to be identified out of diverse libraries of constructs. Previous studies linking the haem binding properties of cyt b_{562} to the antibiotic resistance protein TEM-1 β-lactamase (TEM-1) using a directed evolution approach not only identified sites within TEM-1 tolerant to cyt b_{562} domain insertion that may not have been identified using a rational approach but also identified variants with impressive switching magnitudes of up to 128-fold [31, 37].

Using a directed evolution approach here we have described the construction and characterisation of integral fusion protein scaffolds, between cyt b_{562} and EGFP, which act as efficient energy transfer scaffolds and redox-dependent switches. A diverse number of insertion sites within EGFP tolerant to cyt b_{562} insertion were identified. As expected the majority of the tolerated insertion positions were located in the loop regions connecting the secondary structures thereby not disrupting the β-barrel fold of EGFP. However, several insertion sites were identified in β-strands, α-helices and a tight turn that may not have been chosen by rational design approaches.

The power of the directed evolution approach described earlier (Chapter 3) to implement domain insertion is the ability to design peptides that link the two domains at the site of insertion. Where as other available transposon based techniques that involve carrying the DNA cassette encoding the insert domain within the transposon results in the linker sequences being encoded by the transposase recognition elements and the target site duplication [41, 42]. Previous work has shown that length, flexibility and hydrophilicity of the linking peptides can play a major role in the tolerance to domain insertion and switching magnitude obtained from functionally coupled variants [31, 37, 143]. Short linkers have been shown to reduce the tolerance of EGFP to cyt b_{562} insertion (Chapter 3, Section 3.2.8) probably due to reduced conformational flexibility for the correct and efficient folding of the EGFP domain.

However, short linkers were also shown to have a positive effect on the switching magnitudes obtained from cyt b_{562}-TEM-1 integral fusion chimeras [31, 37].

DNA sequence analysis identified one variant, CG6, which had the GlyGlySer tripeptide linker at the N-terminal end of the cyt b_{562} domain but just a single Gly residue at the C-terminal end of cyt b_{562} (Table 5.1). The difference in linker length separating the two domains of this integral fusion scaffold, with respect to the other cyt b_{562}-EGFP chimeras, could account for the increased levels of fluorescent quenching (up to 100%) upon haem binding. This is potentially due to the shorter linker reducing the conformational flexibility between the two domains and orienting the cyt b_{562} domain in an optimal position such that the two chromophores (EGFP chromophore and haem) are in close proximity for optimal energy transfer. The nature and consequence of this shorter linker is described in more detail in Chapter 6.

5.3.1 Tolerance of EGFP to cyt b_{562} domain insertion

Unlike with the TND library variants there is less of a clear boundary between tolerated and non-tolerated insertion sites when viewed with respect to the secondary structure topology (Fig 5.4) and tertiary structure (Fig 5.5) of EGFP. In the TND library (Chapter 4) there was a very obvious divide between amino acid deletions being tolerated in loops and towards the ends of the β-barrel structure, with the non-tolerated deletions concentrated to the β-strands (Chapter 4, Fig 4.4 and Fig 4.5).

Tolerated domain insertions were identified in ordered secondary structures (Fig 5.3), and non-tolerated domain insertions identified in the loops towards the end of the β-barrel. Domain insertion can be considered a major mutational event that breaks up the continuity of the peptide backbone of the accepting domain. Without a very detailed knowledge of the accepting domain it is almost impossible to identify sites that will or wont tolerate a domain insertion, retain the functions of both domains and result in functional coupling of the two domains. Although dogma suggests loops are more tolerant to mutational events due to their inherent conformational flexibility they may also be crucial in forming on pathway folding intermediates, or stabilizing the parent domain.

For these reasons it is important to use a directed evolution approach to domain insertion so that variants can be identified that do retain functionality and

exhibit functional coupling. Screening through libraries of proteins also allows for the identification of variants that exhibit the highest switching magnitudes.

5.3.2 Effect of cyt b_{562} domain insertion on EGFP spectral characteristics

Apart from cyt b_{562} domain insertions adjacent to the chromophore (CG8 and CG15) of EGFP, the spectral properties remained unaltered for all variants with only minor decreases in brightness due to reduced molar absorption coefficients or quantum yields. This implies that cyt b_{562} domain insertion is well tolerated in the variants studied here with little disruption to the overall fold of the EGFP domain or the chromophores local environment.

In the case of CG8 and CG15, cyt b_{562} domain insertion between residues N146 and S147 in β-strand 7 of EGFP (Fig 5.5) resulted in altered spectral properties (Fig 5.7). There is a complex hydrogen bonding network involving residues in β-strands adjacent to the chromophore and the chromophore itself (Chapter 1, Fig 1.12 b) [62]. The hydrogen-bonding network in EGFP maintains the chromophore in the deprotonated state resulting in a single excitation maximum at 488 nm. Disruption of the hydrogen-bonding network can promote the protonation of the phenolate of Tyr66 giving rise to the 400 nm excitation maximum [62]. Given that residues N146 and H148 are important for stabilizing the phenolate in its anionic form it wouldn't be surprising if cyt b_{562} domain insertion between N146C and S147 had affected the hydrogen-bonding network and therefore altered the spectral properties.

5.3.3 Effect of haem binding on EGFP derived fluorescence properties

The functionality of the cyt b_{562} domain was maintained in all variants, confirmed by UV-visible absorption spectroscopy (Fig 5.8), with functional coupling also observed for all variants (Fig 5.9). In the presence of haem, EGFP derived fluorescence for all of the cyt b_{562}-EGFP integral fusion scaffolds was quenched (Fig 5.9). The extent to which the fluorescence was quenched for each variant was dependent on the position of cyt b_{562} domain insertion.

Binding of haem to the cyt b_{562}-EGFP integral fusion scaffolds also resulted in a decrease of the EGFP derived fluorescence lifetimes (Fig 5.13). The majority of the variants exhibited a ~1.5 fold decrease in their fluorescence lifetime in the presence of equimolar haem (Table 5.3), except for variant CG10 which showed a ~four-fold

decrease in its fluorescence lifetime (Fig 5.13, Table 5.3). Given that variant CG10 was quenched to a similar extent as other cyt b_{562}-EGFP chimeras in the presence of equimolar haem (Fig 5.9) it stands to reason that the increased fold change in fluorescence lifetime upon haem binding may be mediated through structural effects as well as quenching via energy transfer.

The cyt b_{562} domain of CG10 is inserted between residues A154 and D155 at the end of β-strand 7 (Fig 5.15). There are two residues, N146 and H148 that are a part of β-strand 7 and form part of the extensive hydrogen bonding system with the fluorophore (Fig 5.15 and Fig 1.12 b). Possible structural changes upon binding of haem to the cyt b_{562} domain of CG10 may be propagated down β-strand 7 disrupting the structural integrity of the β-barrel, and affecting the hydrogen bonding network between residues in β-strand 7 and the chromophore (Fig 5.15). Alterations in the hydrogen bonding nework upon haem binding may be apparent by changes in the spectral characteristics of CG10 however given that haem quenches fluorescence these may not be detectable.

Previous studies on changes to a cyan fluorescent proteins (CFP) fluorescent lifetime upon interaction with ATP highlighted the importance of the residue H148 on maintaining the fluorescence lifetime properties [144]. Therefore similar mechanisms may be taking effect in CG10 altering its fluorescence lifetime. It would be interesting to investigate the fluorescence lifetimes of variants CG8 and CG15, which also have a cyt b_{562} domain inserted within β-strand 7 but in closer proximity to H148 than in CG10, to see if similar decreases in fluorescence lifetimes are detected for these variants.

5.3.4 CG6 redox sensing properties

The results from haem mediated effects on fluorescence intensity (Fig 5.9) and fluorescence lifetime (Section 5.2.7) identified variant CG6 as an optimal energy transfer scaffold with respect to the other cyt b_{562}-EGFP chimeras. Previous attempts to construct artificial energy transfer scaffolds [33, 82, 138], by terminal fusion of fluorescent proteins to cyt b_{562} domains, have only resulted in systems with a modest 65% energy transfer efficiency. The energy transfer scaffold, CG6, identified here can

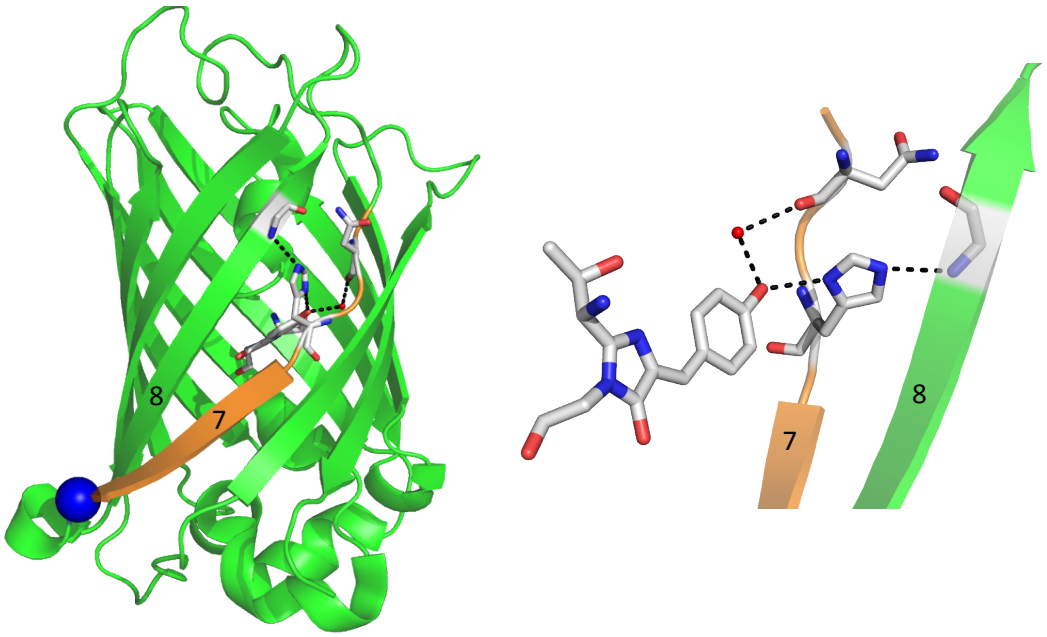


Fig 5.15 CG10 cyt b domain insertion position in eGFP. a, Variant CG10 has a cyt b_{562} domain inserted at D155 (blue sphere) of EGFP (cartoon). D155 is the last residue in β-strand 7 (orange), which contains important residues (sticks, CPK) for maintaining the hydrogen bonding network with the chromophore. **b,** Part of the hydrogen bonding network (black dashed lines) between residues in β-strands 7 and 8, a highly conserved water molecule (red sphere) and the chromophore. The chromophore and residues are shown as sticks in CPK colouring, whilst β-strands 7 and 8 are shown in orange or green cartoon mode respectively.

mediate up to 100% energy transfer from the excited state chromophore of EGFP to the haem bound moiety in the cyt b_{562} domain (Fig 5.9 and section 5.2.7), similar efficiencies to that of naturally occurring light harvesting systems.

Apart from the CG6 scaffold having the potential to be used as a sensor of haem or as an efficient energy transfer scaffold, it has the potential to act as a sensor for the natural cellular ROS, H_2O_2. In the presence of equimolar or an excess of haem essentially 100% of CG6 fluorescence is quenched (Fig 5.9 and Fig 5.11 a). On the addition of 0.02% (v/v) H_2O_2 fluorescence recovery of CG6 was observed at a rate of ~0.5 min^{-1}, until 100% of CG6 starting fluorescence was recovered (Fig 5.11 a). Addition of other oxidising reagents to holo-CG6 elicited no gain in fluorescence making CG6 potentially H_2O_2 specific.

However CG6 did display a ~14 fold slower rate of fluorescence recovery with respect to the terminal fusion CG1. This is potentially due to differences in the solvent accessibility of the bound haem moiety to the solvent and therefore the oxidising effects of H_2O_2. The structural and biophysical characterisation of CG6, to help determine the mechanisms of functional coupling at the molecular level, have been performed and the results discussed in Chapter 6.

5.3.5 Conclusion

Using a novel transposon based directed evolution approach to construct integral domain fusion scaffolds between cyt b_{562} and EGFP it has been possible to identify a variant CG6 with a 100% energy transfer efficiency that retains the high haem binding affinities of wt cyt b_{562} whilst maintaining EGFP derived fluorescence properties. The power of the directed evolution approach allows for the linking peptides between the two domains to be designed giving greater control over the resulting constructs and the large switching magnitudes observed between functionally coupled integral fusion scaffolds.

Unlike with the survey of tolerated single amino acid deletions throughout EGFP there is a less clear boundary between where in GFP will be tolerable or intolerable to domain insertion or which positions will give the best functional coupling between the two domains used. Therefore establishing a set of rules for identifying potential targets within a parent domain for domain insertion is more

difficult. It can be said that domain insertions are less likely to be tolerated within ordered secondary structures and more likely to be tolerated within loop regions.

Given the disruptive nature of a domain insertion on a parent domain even insertion positions within loops can potentially be detrimental to parent domain function. Very small differences of insertion position (one amino acid apart) can also have a profound effect on the efficiency of functional coupling observed, as was evident between variant CG3 and CG4 where CG4 resulted in 95% maximal quenching and CG3 only achieved ~75% maximal quenching in the presence of haem (Fig 5.9). This highlights the necessity and power of a directed evolution strategy over a rational design approach, paving the way for advancements in the construction of novel artificial biomolecular switches.

Chapter 6: Structure and biophysical characterization of the energy transfer scaffold, CG6

6.1 Introduction

In Chapter 5 an integral fusion protein termed CG6 was constructed in which cyt b_{562} was inserted between residues Y39 and G40 of EGFP. This particular variant displayed high efficiency energy transfer between the EGFP chromophore and the haem bound to cyt b_{562}, as evident by virtually total fluorescence quenching in the holo-protein. Haem is a known efficient quencher of fluorescence [81], including that of EGFP [33, 82], with quenching occurring *via* energy transfer to the haem moiety. Energy transfer efficiency between two chromophores is dependent on their proximity and orientation with respect to one another [145].

Domain insertion can generate intimate structural and spatial linkage between domains to facilitate functional coupling, as demonstrated in Chapter 5 with the generation of CG6, and is an emerging protein engineering strategy for linking unrelated protein function. However, our molecular understanding of functional linkage in engineered domain insert protein scaffolds is hindered by the lack of detailed structural information. This in turns limits our ability to construct these potentially useful protein scaffolds and understand how functional coupling occurs. To date, most success has been achieved using directed evolution mutagenesis approaches to sample diverse domain insert positions and linking sequence followed by selection for coupled functionality.

To understand the molecular basis for increased energy transfer efficiency and the oxidant induced fluorescence switching properties observed for CG6, this Chapter will provide a detailed biophysical and structural characterisation of CG6. This will include determination of the 3D atomic resolution structure by X-ray crystallography. Such an analysis of CG6 will further our knowledge at the molecular level on the functional coupling of two disparate proteins and the effect domain insertion has on protein stability.

6.2 Results

6.2.1 Biophysical charaterisation of CG6

6.2.1.1 Haem-dependent structural changes probed by CD spectroscopy

To determine if there was any structural change upon haem binding to apo-CG6, circular dichroism (CD) spectra of CG6 were measured in the presence and absence of haem. There were no major structural changes on haem binding to CG6 (Fig 6.1 a). The CD spectrum of CG6 is dramatically different to EGFP, with CG6 containing an increased helical signature content (as indicated by the troughs at 210 and 222 nm), as to be expected upon insertion of the largely helical cyt b_{562} into the predominantly β-sheet EGFP (Fig 6.1 a). The troughs around 208 and 220 nm were slightly deeper for holo-CG6 compared to apo-CG6 suggesting a small increase in helical character but on the whole, the two forms of CG6 appeared to have very similar structures.

CD spectra for CG1, essentially a head-to-tail fusion between cyt b_{562} and EGFP (Chapter 5), were also measured in the absence and presence of haem (Fig 6.1 b). CG1 is a control to assess if there was any difference in the overall nature of secondary structure character between a head-to-tail fusion scaffold or a domain insertion fusion scaffold. As with CG6 there were no major structural changes upon haem binding to CG1. The CD spectrum for CG1 was also dramatically different to EGFP, with CG1 containing an increased helical signature content (as indicated by the troughs at 210 and 222 nm), due to the fusion of the largely helical cyt b_{562} to the N terminus of the predominantly β-sheet EGFP (Fig 6.1 b). The CD spectrum for holo-CG1 also suggested a small increase in helical character, evident from deeper troughs around 208 and 220 nm. As with CG6, the apo- and holo- forms of CG1 appeared to have very similar structures.

Comparison of the CD spectra for CG1 and CG6 (Fig 6.1 c) shows there is very little difference in the secondary structure signature between the two cyt b_{562}-EGFP chimeras. Given that the CD spectra are very similar between the terminal fusion (CG1) and the integral fusion (CG6) of cyt b_{562} with EGFP implies that the overall structures of both domains have not been significantly altered.

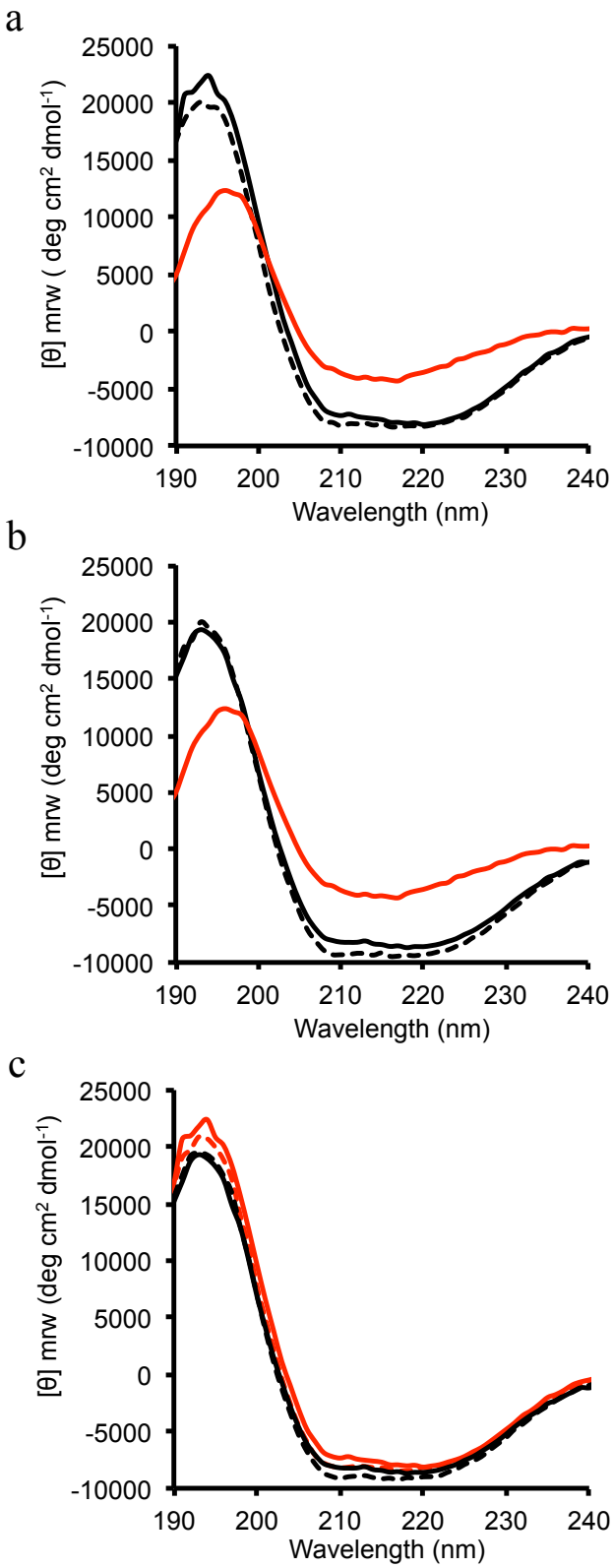


Fig 6.1 Effect of haem on the secondary structure characteristics of CG1 and CG6. Circular dichroism spectroscopy was performed for **a,** CG6 and **b,** CG1 (5 μM samples) in the absence (solid black line) and presence (dashed black line) of equimolar haem. CD spectra for EGFP (red line) was also measured as a comparison to the fusion scaffolds CG1 and CG6. **c,** Overlay of CD spectra for CG1 (red) with CG6 (black) in the absence (solid lines) and presence (dashed lines) of equimolar haem.

6.2.1.2 Analytical size exclusion chromatography

Analytical size exclusion chromatography was performed on CG1 and CG6 to determine if cyt b_{562} fusion to the N-terminus of EGFP (CG1) or integral fusion within EGFP (CG6) has an effect on the quaternary structure of EGFP. The effect of haem binding on the oligomeric nature of the chimeras was also investigated.

A standard curve for the relationship between molecular weight and elution volume from the column was determined using the BioRad gel filtration standards (Section 2.6.2.1), so that the elution volume for EGFP, CG1 or CG6, at different concentrations, could be related to the molecular weight of that species under native conditions (Appendix C).

The elution profile for CG6 (Fig 6.2 c) confirmed that it is monomeric in both its apo and holo forms. The apo and holo CG6 eluted at similar volumes (~15.2 ml), equivalent to an apparent molecular weight of 38 kDa, close to that of its theoretical monomeric molecular weight (38.9 kDa) (Table 6.1). There is a very slight increase in elution volume for holo-CG6 with respect to apo-CG6 (up to ~0.04 ml), equivalent to a decrease in calculated molecular weight of ~ 500 Da, even though the theoretical molecular weight is increased due to the bound haem (~616 Da). This slight decrease in hydrodynamic volume implies that holo-CG6 has a slightly more compact structure with respect to apo-CG6.

The elution profile for CG1 showed two elution peaks; a major peak at ~15.9 ml and a minor peak at ~16.06 ml (Fig 6.2 b). The protein in the CG1 samples eluting at ~16.06 ml is similar to that of the elution volume for EGFP (~16.2 ml) (Fig 6.2 a) and is equivalent to a molecular weight of ~25.7 kDa, very close to the theoretical molecular weight for EGFP (26.9 kDa) (Table 6.1). This protein species corresponding to the minor peak for CG1 is most probably a degradation product of CG1 and represents the EGFP domain without the cyt b_{562} domain fused to its

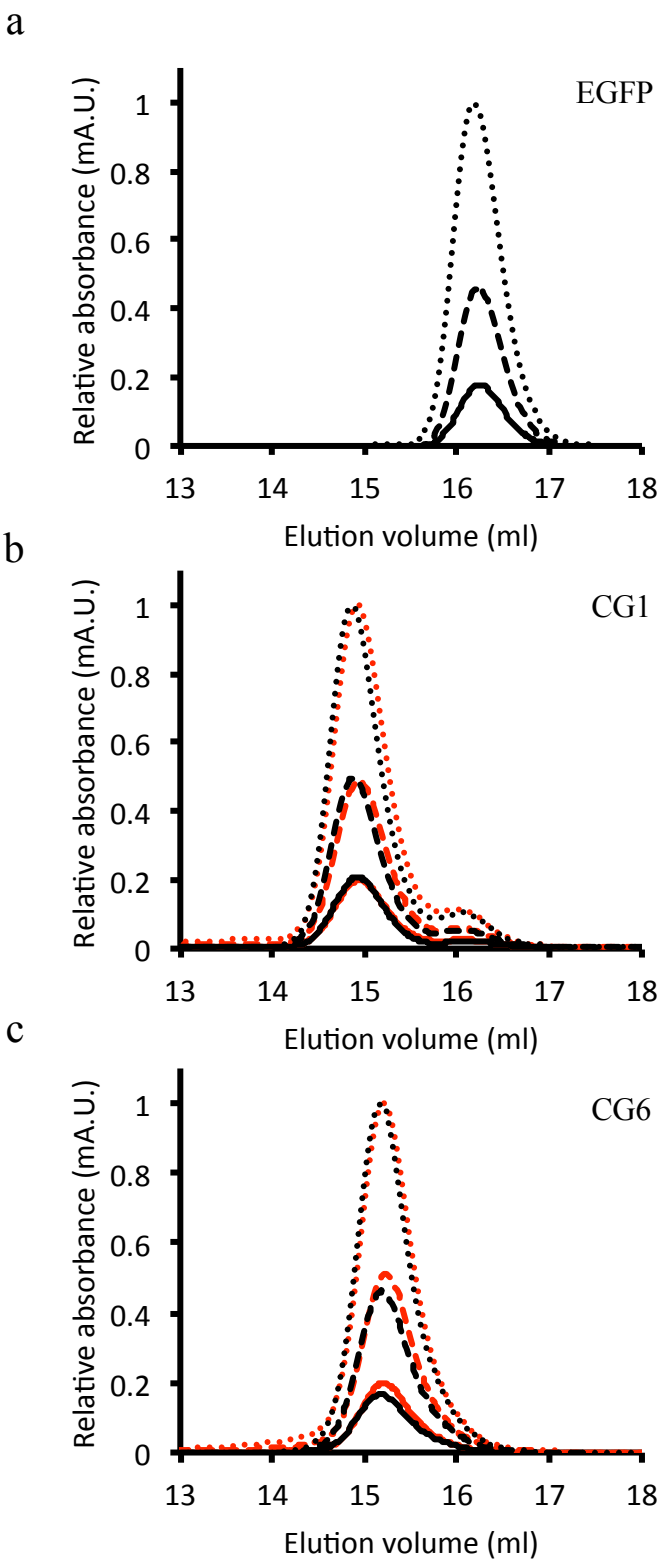


Fig 6.2 Size exclusion chromatography of EGFP, CG1 and CG6. Samples of **a,** EGFP **b,** CG1 or **c,** CG6 were applied to a Superdex^{TM} 200 gel filtration column and the elution of the protein samples monitored by absorbance at 488 nm. Protein concentrations of 10 μM (solid black line), 25 μM (dashed line) or 50 μM (dotted line) were applied to the column in the absence (black) and presence (red) of equimolar haem.

Table 6.1 Size exclusion chromatography analysis

Variant	Concentration (μM)	Elution volume (ml)[a]	Calculated Mw (Da)[b]	Average Mw (Da)[c]	Theoretical Mw (Da)[d]
EGFP	10	16.23	23700	24000	26941
	25	16.23	23800		
	50	16.18	24400		
Apo-CG1	10	14.93	43800	44800	39138
	25	14.86	45100		
	50	14.84	45500		
Holo-CG1	10	14.93	43700	43800	39754
	25	14.93	43700		
	50	14.92	44000		
Apo-CG6	10	15.19	38700	38900	38945
	25	15.18	38800		
	50	15.16	39200		
Holo-CG6	10	15.20	38500	38500	39562
	25	15.19	38500		
	50	15.19	38500		

[a]Elution volumes determined from peak absorbance at 488 nm (Fig 6.2)
[b]Molecular weights calculated using standard curve (Appendix C)
[c]Average calculated molecular weights over the different concentrations
[d]Estimated molecular weight from primary sequence

N-terminus. The major absorption peak for CG1, with an elution volume of ~15.9 ml, is equivalent to a protein species with an estimated molecular weight of ~44.3 kDa, up to 5.5 kDa larger than the theoretical molecular weight for CG1 (~39.1 kDa) (Table 6.1). Given that a CG1 dimer would have a theoretical molecular weight of ~78.3 kDa and an estimated elution volume of ~13.68 ml the major absorption peak observed for CG1 (~14.9 ml) is likely to represent a monomeric form with an increased hydrodynamic volume.

As with CG6, the elution volumes between apo-CG1 and holo-CG1 differed by up to ~0.09 ml, with an average estimated molecular weight for apo-CG1 being 1 kDa heavier than holo-CG1 (Table 6.1). This could be indicative of a more compact CG1 and CG6 structure when in the holo-form.

6.2.1.3 Guanidinium chloride induced equilibrium unfolding

To assess the effect of cyt b_{562} domain insertion on EGFP stability in the CG6 construct, guanidinium chloride (GdmCl) induced equilibrium unfolding experiments were performed (Fig 6.3). Furthermore, the effect of haem binding on stability was also probed through comparison of unfolding of apo- and holo-protein. CG1 was used as a control to investigate the role of a standard head-to-tail fusion arrangement. EGFP was also denatured in the absence and presence of haem as a control to confirm that free haem in solution does not affect the stability of the EGFP domains.

Unfolding of EGFP, CG1 and CG6 had to be monitored by absorbance at 488 nm, rather than fluorescence emission at 520 nm, as in the holo forms both CG1 and CG6 have their fluorescence quenched (Chapter 5, Fig. 5.9). The λ_{max} for EGFP in its native form is at ~488 nm. Upon unfolding of the β-barrel structure in GdmCl the absorbance at 488 nm decreases whilst an absorption maxima increases at ~375 nm [142]. Given that fully denatured EGFP essentially has no absorbance at 488 nm it is possible to monitor its unfolding *via* the chromophore absorbance.

Protein samples (4 μM) were incubated in a range of GdmCl concentrations (0–6 M) at 37 °C in 96-well plates. Absorption at 488 nm was measured after 2.5, 5, 22, 48, 120 and 250 hrs. The fraction of denatured protein was calculated from the absorption data at 488 nm using Equation 13 (Section 2.6.2.2). To make sure

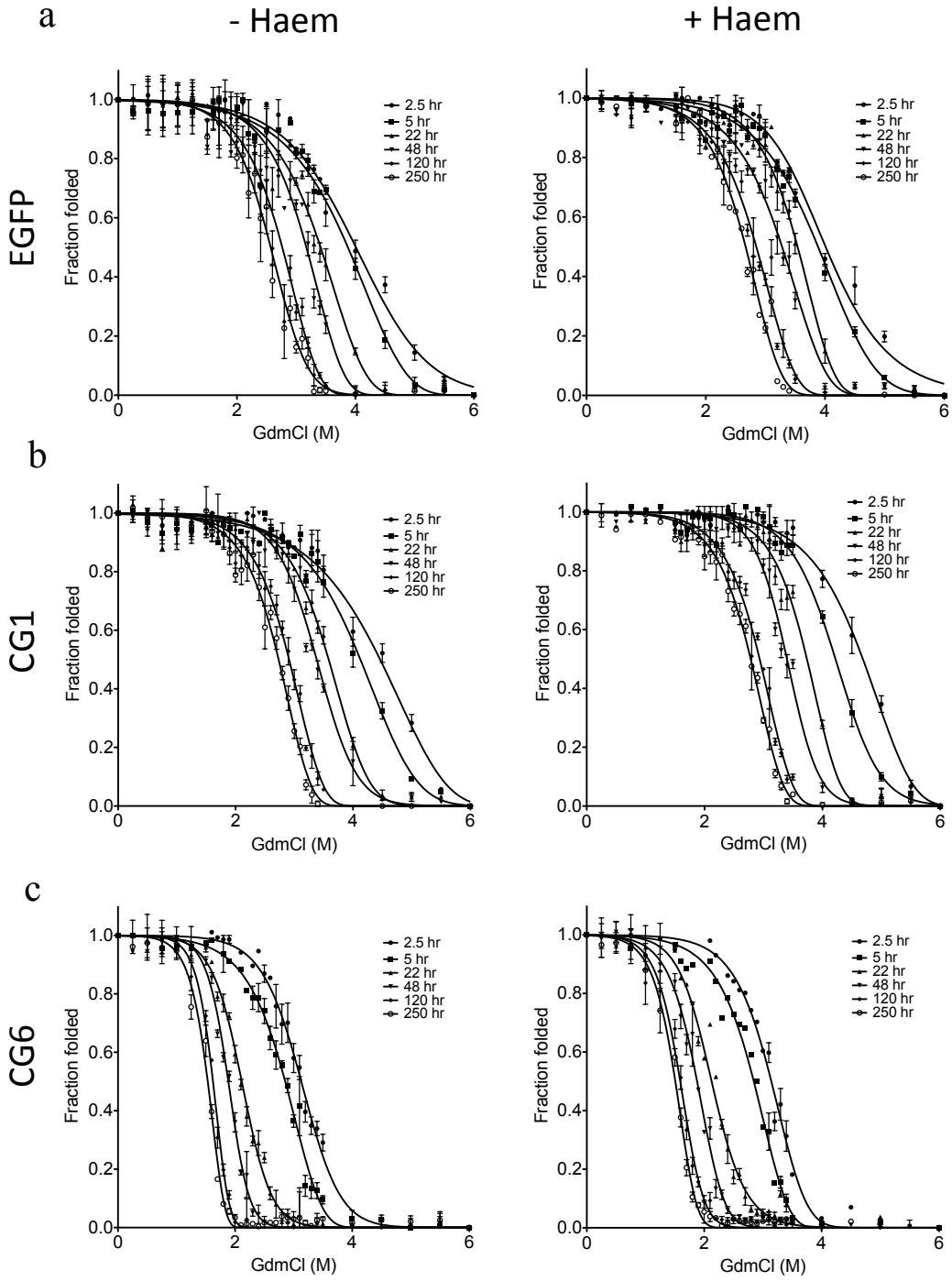


Fig 6.3 Guanidinium chloride induced equilibrium unfolding. Equilibrium unfolding data for **a,** EGFP, **b,** CG1 and **c,** CG6 at different time points as indicated in the figure. Data was collected in the absence (left panels) and presence (right panels) of equimolar haem. The data were fit to a 5-parameter asymmetric equation (Section 2.6.2.2).

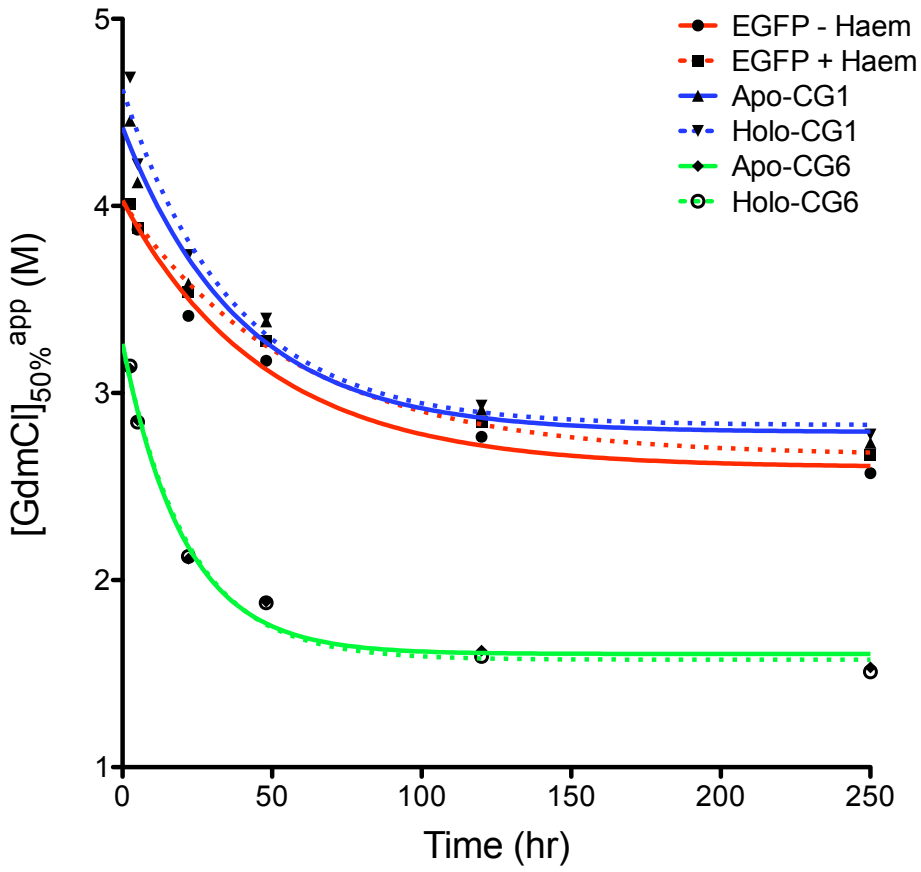


Fig 6.4 Equilibrium unfolding rate determination for EGFP, CG1 and CG6. Apparent $[GdmCl]_{50\%}$ values determined from the 5-parameter asymmetric fit to the equilibrium unfolding data (Fig 6.3) were plot against incubation time and fit with a single exponential decay.

Table 6.2 Equilibrium unfolding parameters and melting temperature

Variant	k_{eq} [a] (10^{-2} hr^{-1})	$[GdmCl]_{50\%}$ [b] (M)	$\Delta G_{N-D}^{H_2O}$ [c] (kcal mol^{-1})	m-value[d] (kcal mol^{-1} M^{-1})	T_m^{app} (°C)
EGFP - Haem	2.08±0.36	2.68±0.02	5.05±0.49	1.957±0.185	84
EGFP + Haem	1.74±0.20	2.68±0.01	5.16±0.50	1.953±0.190	-
Apo-CG1	2.55±0.72	2.75±0.01	5.64±0.41	2.070±0.152	84
Holo-CG1	2.73±0.77	2.79±0.01	5.49±0.38	1.980±0.141	-
Apo-CG6	4.83±0.94	1.55±0.01	5.67±0.33	3.873±0.195	74
Holo-CG6	4.55±0.90	1.54±0.01	5.56±0.50	3.742±0.303	-

[a]Equilibrium unfolding rate constant determined from a single exponential decay (Fig 6.4)
[b]Concentration of GdmCl at which 50% of the protein sample is in the native and denatured state at equilibrium, determined from a 2 state model (Section 2.6.2.2).
[c]Free energy of denaturation ($\Delta G_{N-D}^{0} = \Delta G_{N-D}^{H_2O} - m[GdmCl]$)
[d]Measure of dependence of ΔG on denaturant concentration determined from the slope of the plots in Fig 6.5 b.

equilibrium had been approached the apparent [GdmCl] at which half of the protein was in the denatured state ($[GdmCl]_{50\%}^{app}$) was determined from an asymmetric 5 parameter fit (Section 2.6.2.2) (Fig 6.3) and plotted against incubation time (Fig 6.4).

The $[GdmCl]_{50\%}^{app}$ values against time were fit to single exponential decay curves (Section 2.6.2.4, Equation 17) and show that, like other fluorescent proteins, both EGFP and CG1 took an extended period of time (up to 250 hr) to reach equilibrium, independent of whether haem was present or not (Fig 6.4). In contrast CG6 had reached equilibrium ~two-fold quicker (by ~100 hr) than EGFP and CG1 (Fig 6.4 and Table 6.2). The $[GdmCl]_{50\%}$ at 250 hr for CG6 (1.5 M) was also lower than that of EGFP (~2.7 M) and CG1 (~2.7 M) by ~1.2 M (Table 6.2).

Haem appeared to have no effect on $[GdmCl]_{50\%}$ and very little effect on the equilibrium rate constants. As was expected fusion of cyt b_{562} to the N-terminus of EGFP (CG1) does not appear to alter the stability of EGFP, as the $[GdmCl]_{50\%}$ values are almost identical to that of EGFP alone (Table 6.2) and both EGFP and CG1 reach equilibrium after similar incubation times in GdmCl (~250 hr).

As EGFP, CG1 and CG6 all appeared to have approached equilibrium after 250 hr (Fig 6.4) the equilibrium unfolding data from the 250 hr incubation were fit to a two-state model (Section 2.6.2.2) (Fig 6.5 a). It must be stressed that although it appears that all samples have approached equilibrium that it is not assumed that the system is at equilibrium. Nevertheless, thermodynamic parameters can be established for comparison between individual samples used here. The data for EGFP, CG1 and CG6 all fit to a two-state model ($R^2 > 0.98$) despite previous studies showing that equilibrium unfolding data for GFPs is best fit by a three-state model. A three-state model was tested here but did not converge on the data. As explained in Chapter 4 for the equilibrium unfolding of the single amino acid deletion variants of EGFP (Section 4.2.6.2), an unfolding intermediate may be indistinguishable from the native or denatured states under the conditions at which the equilibrium unfolding experiments were run (pH 8.0 at 37 °C).

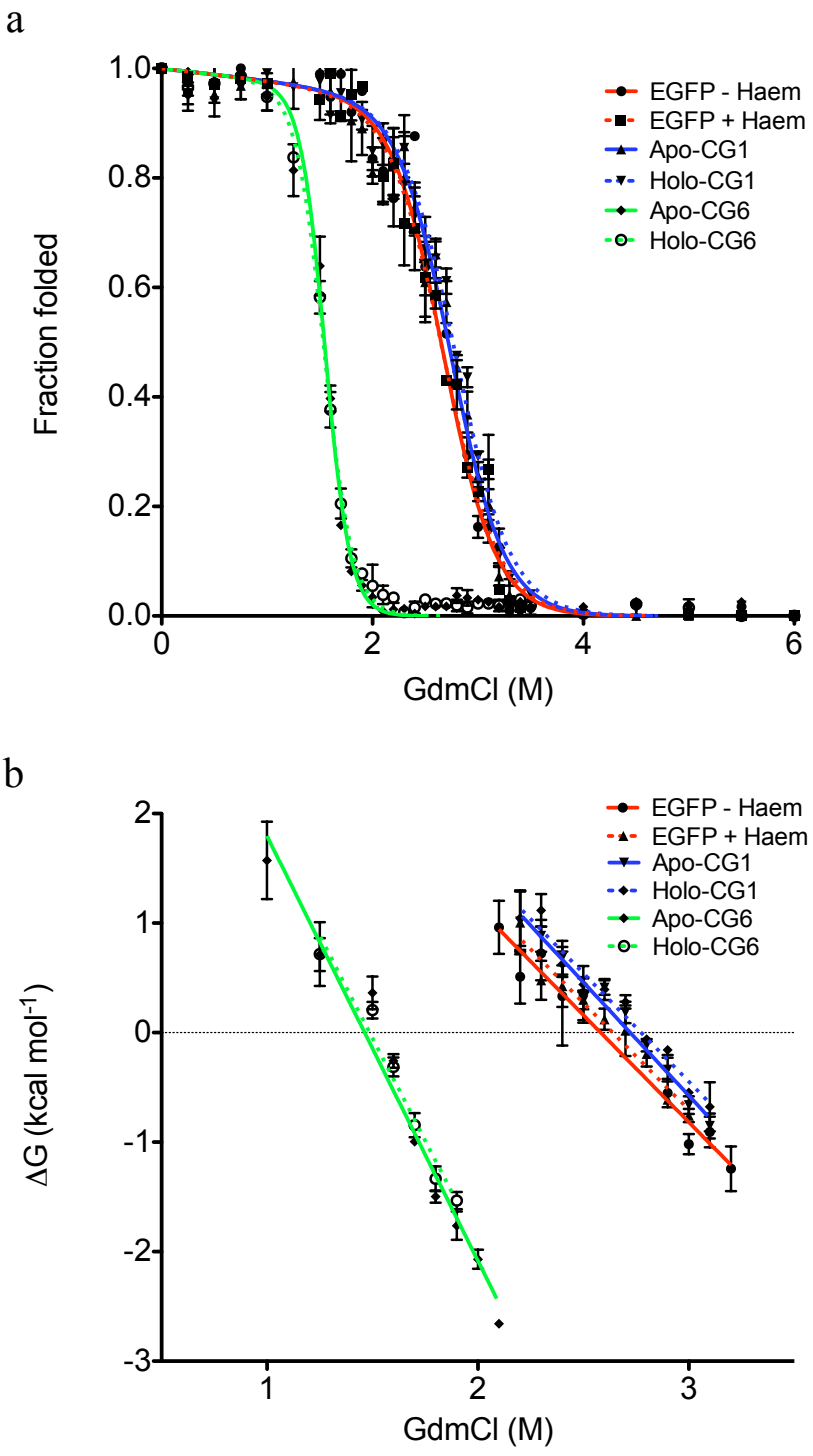


Fig 6.5 Equilibrium unfolding and linear dependence of ΔG on [GdmCl]. a, Equilibrium unfolding data, after 250 hr incubation, were fit to a 2 state model taking into account sloping baselines inherent to spectroscopic measurements of protein samples in GdmCl. **b,** The linear dependence of ΔG on [GdmCl] (m-value) was used to calculate free energies of denaturation ($\Delta G_{N-D}^{H_2O}$).

Although the $[GdmCl]_{50\%}$ at equilibrium for CG6 is ~1.2 M less than EGFP and CG1 the $\Delta G_{N-D}^{H_2O}$ for EGFP (~5.1 kcal mol^{-2}), CG1 (~5.5 kcal mol^{-1}) and CG6 (~5.5 kcal mol^{-1}) were similar. The free energy of denaturation calculated for EGFP agrees closely to previous equilibrium unfolding experiments (5.16 kcal mol^{-1}) [111] and with the value determined by monitoring fluorescence during unfolding (5.28 kcal mol^{-1}) (Chapter 4, section 4.2.6.2). Under oxidizing conditions haem appeared to have no affect on the $\Delta G_{N-D}^{H_2O}$ values for EGFP, CG1 or CG6.

The observation that CG6 has a similar $\Delta G_{N-D}^{H_2O}$ to EGFP and CG1, given its decreased $[GdmCl]_{50\%}$, is due to an increase in the dependence of ΔG_{N-D}^{0} on [GdmCl] (m-value) (Table 6.2). The m-value is a measure of the dependence of ΔG on [GdmCl]. Both EGFP and CG1 have m-values of ~1.9 kcal $mol^{-1}M^{-1}$ (Table 6.2) whilst the ΔG_{N-D}^{0} for CG6 has a much higher dependence on [GdmCl], as is evident by its increased m-value of ~3.8 kcal $mol^{-1}M^{-1}$. This two-fold increase in m-value indicates that denatured CG6 is less compact and structured than denatured EGFP and CG1 or the degree to which an intermediate state is populated has been changed.

As discussed in Chapter 4 (Section 4.2.6.2. and 4.3.3) analysis of equilibrium unfolding data with a 2-state model does not take into account the presence of an intermediate state and can therefore result in an underestimation of the m-value if an intermediate state is populated. Further equilibrium unfolding studies would be required to ascertain 3-state unfolding and a more accurate representation of the impact of domain insertion on EGFP stability and folding.

6.2.1.4 Thermal denaturation of EGFP, apo-CG1 and apo-CG6

To assess the effect of cyt b_{562} domain insertion on the thermostability of the EGFP domain, protein samples (1 μM) were slowly heated (1 °C/min) from room temperature (25 °C) to 98 °C whilst monitoring fluorescence using a qPCR machine (Section 2.6.2.5) (Fig 6.6). There are two phases to the thermal denaturation curves (Fig 6.6), the first slow phase is due to a negative dependence of EGFP fluorescence with increasing temperature, as seen in previous studies [146, 147]. The second phase is a sharp transition from the native to denatured state, from which melting temperatures are calculated.

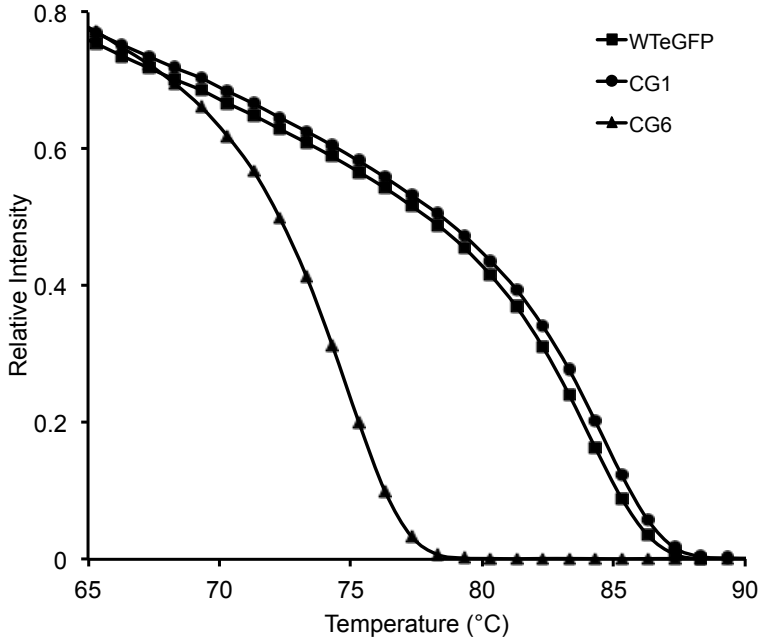


Fig 6.6 Thermal denaturation of EGFP, apo-CG1 and apo-CG6. Fluorescence emission was monitored during temperature ramping from 25-98 °C at 1 °C/min using a qPCR thermal cycler. Melting temperatures (T_m^{app}) are listed in Table 6.3.

Apparent melting temperatures (T_m^{app}) calculated from the fluorescence data (Table 6.2) show that the apo form of the cyt b_{562} fusion to the N-terminus of EGFP (CG1) does not alter the melting temperature of the EGFP domain (Fig 6.6). Both EGFP and apo-CG1 had a calculated melting temperature of 84 °C. In comparison, insertion of cyt b_{562} into EGFP (CG6) reduced the melting temperature of the apo-form by 10 °C to 74 °C, implying that cyt b_{562} domain insertion into EGFP at position Y39 reduces the thermal stability of EGFP.

Given that the thermal denaturation of EGFP, CG1 and CG6 was irreversible (data not shown) accurate thermodynamic parameters cannot be determined. Therefore any potential formation of aggregates may result in the underestimation of T_m^{app} values.

6.2.2 Crystallographic structure of holo-CG6

To understand the molecular basis of the efficient fluorescence quenching in the CG6 cyt b_{562}-EGFP integral fusion variant, described in detail in Chapter 5, the holo form was crystallized. Holo-CG6 protein samples (10 mg/ml in 50 mM Tris-HCl, pH 8.0, 150 mM NaCl and 256 μM haem) were screened for crystal formation by the sitting drop vapour diffusion method with incubation at 4 °C. The crystal of holo-CG6 was obtained from 0.1 M MES/NaOH, pH 6.4, 200 mM magnesium acetate and 20% (w/v) PEG 8000.

X-ray diffraction data were collected on the Diamond Light Source beamline I02 thanks to Dr Anna Piasecka. The crystallographic statistics are shown in Table 6.3. Structure determination was performed in collaboration with Dr. Matthias Bochtler and Dr. Honorta Czapińska. The full methods for structure determination are outlined in Section 2.7.2.

Crystals grew in space group P2(1) and contained three molecules of the fusion protein in the asymmetric unit. Two of molecules were very well ordered. However, the third had high temperature factors (Chain C), particularly for the EGFP domain. Routine molecular replacement was used to orient and position the EGFP and cyt b_{562} domains for the two well-ordered copies of the molecule in the asymmetric unit. The poorly ordered third molecule in the asymmetric unit did not give significant molecular replacement signals and was found during refinement (Section 2.7.2).

A R-Free value of 27.7 % indicates that the models determined for the three molecules in the asymmetric unit are a good representation of the experimental data. Given that the third molecule in the asymmetric unit was poorly ordered and has high temperature factors, the R-Free value would not reduce any more with further refinement.

The EGFP and cyt b_{562} domain sit next to each other in the integral fusion scaffold (Fig 6.7) connected by a shorter single glycine 'inner linker' and a longer Gly-Gly-Ser 'outer linker' (Fig 6.7). The difference in the length of the linkers are important for the arrangement of the two domains with respect to one another and are discussed in more detail in Section 6.2.2.2. The individual EGFP and cyt b_{562} domains in the three molecules were very similar to previously solved structures of EGFP and cyt b_{562} (Fig 6.8) as was evident from RMSD values over the backbone atoms of 0.31-0.38 Å or 0.78-0.82 Å respectively (Table 6.4). This showed that insertion of cyt b_{562} into EGFP in the tight turn between β-strands 2 and 3 has not affected the overall structures of the EGFP or cyt b_{562} domains.

Due to the third molecule (chain C) in the asymmetric unit having high temperature factors detailed analysis of the CG6 structure has only been performed using chain A or chain B.

6.2.2.1 EGFP and cyt b_{562} domain arrangement in the CG6 crystal structure

In all three molecules in the asymmetric unit the crystal structures indicated that the two domains are arranged in a side-by-side manner adopting a "V" shape (Fig 6.9 a). Structural alignment of the EGFP domains from each molecule in the asymmetric unit showed some variation in the hinge angle between the domains (up to 23° for different pairs (Fig 6.9 a), but with the inter-domain contacts largely remaining the same (Section 6.2.2.3). The RMSD across all backbone atoms for the three different molecules (2.5-4.8 Å) highlights the difference of the position of the cyt b_{562} domains with respect to the EGFP domain (Table 6.5). The difference in hinge angle between the two domains seen in the three molecules might suggest some flexibility between the two domains. In principle, the crystal could select the most compact conformation from a continuum that might range from closed to fully open.

Table 6.3 Crystallographic statistics for CG6

Space group	P1 2(1) 1
a (Å)	64.8
b (Å)	125.2
c (Å)	89.3
β (°)	90.4
Resolution range (Å)	30 - 2.75
Total reflections	136754
Unique reflections	36793
Completeness (%) (last shell)	99.3 (100.0)
I/σ (last shell)	19.9 (2.4)
R(sym) (%) (last shell)	3.4 (32.4)
B(iso) from Wilson ($Å^2$)	80.5
Protein atoms excluding H	7996
Solvent molecules	98
R-factor (%)	23.4
R-free (%)	27.7
Rmsd bond lengths (Å)	0.017
Rmsd angles (°)	1.8
Ramachandran core region (%)	92.7
Ramachandran allowed region (%)	6.6
Ramachandran additionally allowed region (%)	0.1
Ramachandran disallowed region (%)	0.6

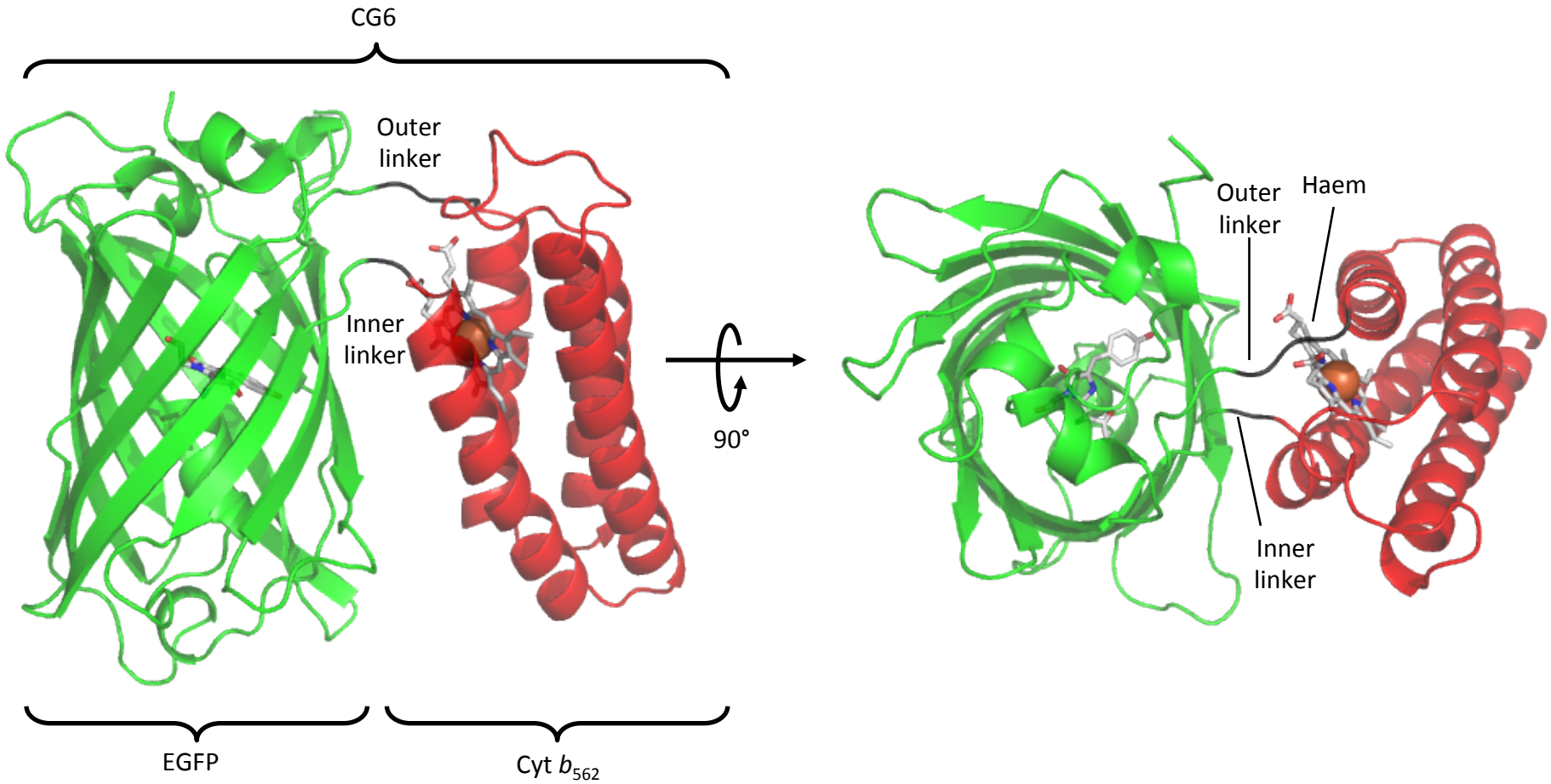


Fig 6.7 EGFP and cyt b_{562} domain arrangement in the crystal structure of CG6. Cartoon representation of CG6 is shown from a side on view (left) and top down view (right). The EGFP domain is shown in green with the cyt b_{562} domain shown in red. The single glycine 'inner linker' and Gly-Gly-Ser 'outer linker' are shown in black. The EGFP domain chromophore and haem bound to the cyt b_{562} domain are shown in stick representation coloured by CPK.

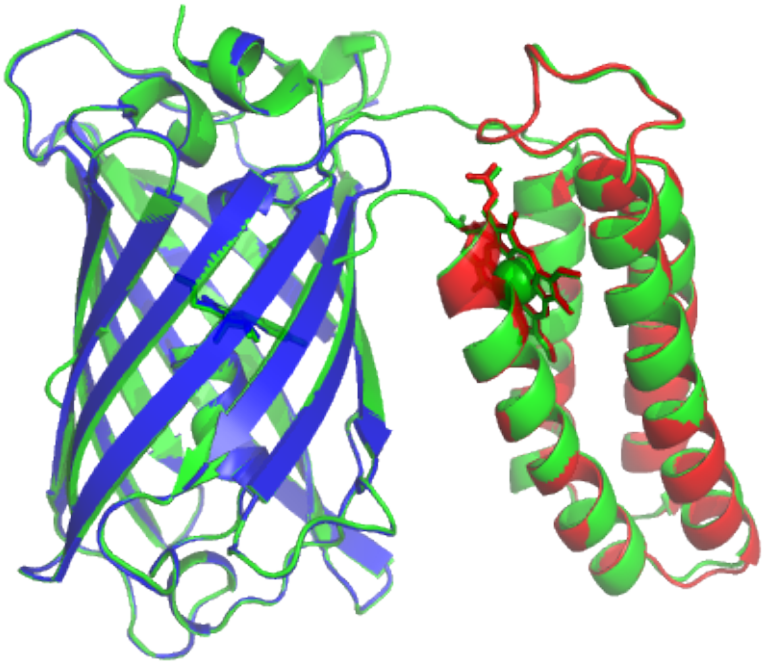


Fig 6.8 Superposition of CG6, EGFP and cyt b_{562} structures. Cartoon representations of CG6 (green) superposed on wild type EGFP (blue, structure determined in Chapter 5) and wild type cyt b_{562} (red, PDB:256B).

Table 6.4. RMSD measurements between EGFP or cyt b_{562} and the 3 molecules on CG6 in the asymmetric unit.

		EGFP[a]	Cyt b_{562}[a]
CG6	Chain A	0.31	0.78
	Chain B	0.32	0.82
	Chain C	0.38	0.78

[a] All RMSD values are in Å

Table 6.5 RMSD measurements between all three molecules of CG6 in the asymmetric unit

Molecule	A[a]	B[a]	C[a]
A	0	4.82	2.50
B	4.82	0	3.14

[a] All RMSD values are in Å

The V-shape of the fusion protein positions the chromophores of EGFP and cyt b_{562} close together (Fig 6.9 b). The distance between them is approximately 17 Å, and thus smaller than the distance that could be expected for a more open arrangement of the domains. The two chromophores lie in the same plane with the angle between them being approximately 45° (Figure 6.1 b).

6.2.2.2 Importance of linker length to CG6 domain arrangement

The different length of the linkers plays a critical role towards fixing the V-shape of the fusion protein. The short inner linker between R148 ($R106^{cytb}$) and G150 ($G40^{GFP}$) consists of a single glycine that is located in a loop between the N-terminal end of β-strand 3 in EGFP and the C-terminal α-helix of cyt b_{562} (Fig 6.10). The shorter inner linker is positioned essentially the same in all three molecules in the asymmetric unit (Fig 6.10) and is likely to have less flexibility than the outer linker, therefore forming a turn allowing the two domains to align side by side (Fig 6.10).

The longer outer linker (Fig 6.10 a) between F39 ($Y39F^{GFP}$) and A43 ($A1^{cytb}$) consisting of the Gly-Gly-Ser tripeptide, has increased flexibility as is evident from the three different cyt b_{562} domain arrangements (Fig 6.9). The longer outer linker is positioned differently in all three molecules, and provides the flexibility required for the two domains to arrange side-by-side (Fig 6.10 b). A shorter linker in its place would result in reduced conformational flexibility between the two domains potentially producing a more open conformation.

6.2.2.3 EGFP and cyt b_{562} domain arrangement of CG6 in solution

To ascertain the relevance of the CG6 domain arrangement observed in the crystal structure small angle X-ray scattering (SAXS) analysis was performed on CG6 in solution to determine the domain arrangement. The experimental intensity data (Fig 6.11) were compared to predicted fits to data for three different domain arrangement models (performed by Dr Michal Gajda at EMBL): The two domains aligned side-by-side, at right angles to one another or aligned to form a long rod (Fig 6.11). An excellent fit was observed for the side-by-side domain arrangement ($\chi^2 = 0.98$) as also

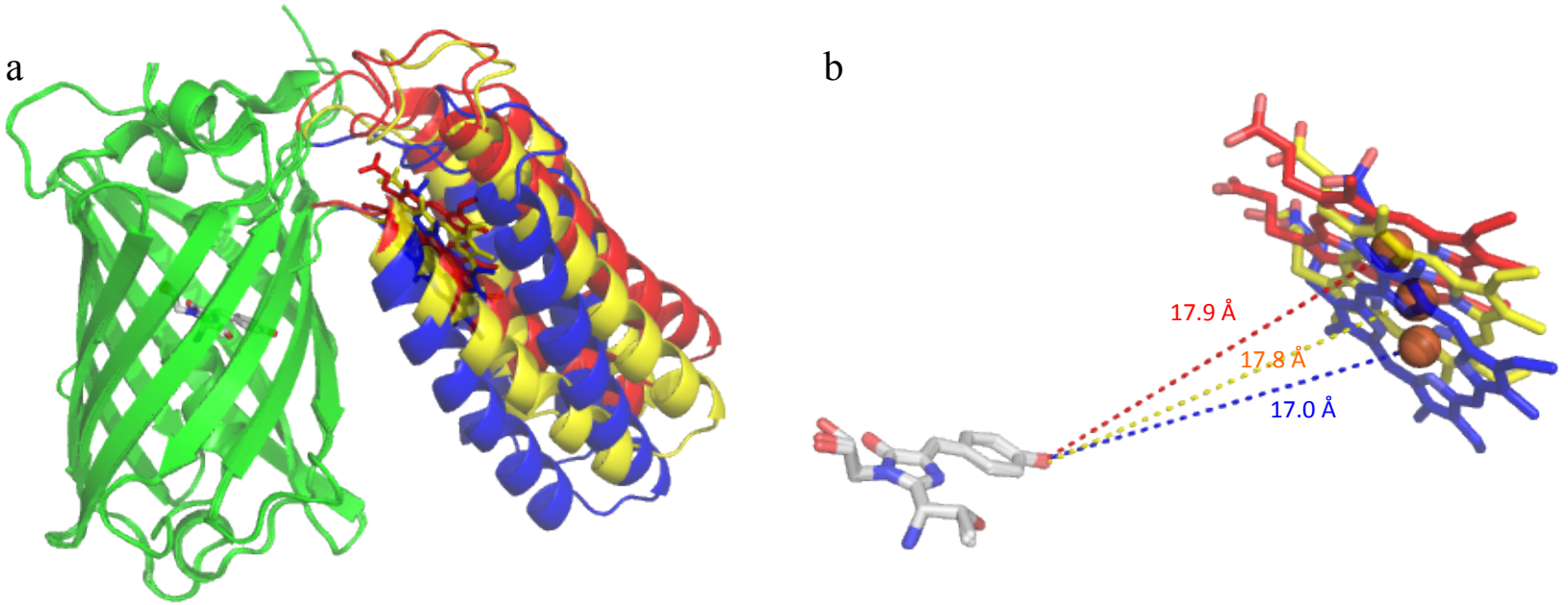


Fig 6.9 Superposition of the 3 CG6 molecules in the asymmetric unit. a, The 3 molecules were aligned by the EGFP domains (green cartoon) with the chromophore shown in stick representation. The cytochrome domains are shown in cartoon representation coloured red, yellow and blue. Unless stated otherwise the structure used for representation of CG6 has the cyt b_{562} domain coloured red. **b,** Interchromophore distances between the EGFP domain chromophore and the cyt b_{562} bound haem in the three molecules of the asymmetric unit.

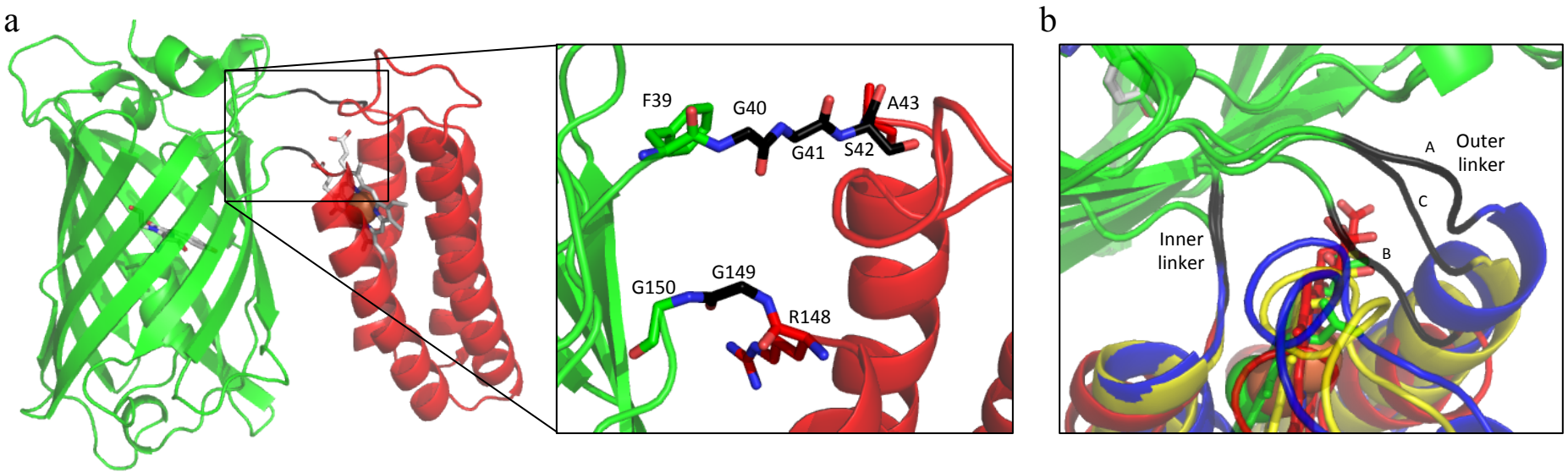


Fig 6.10 Structure of the inner and outer linkers joining the EGFP domain to the cyt b_{562} domain in CG6. a, Cartoon representation of CG6 with the EGFP domain in green, the cyt b_{562} domain in red and the linkers in black. **Inset,** Close up view of the inner and outer linkers. The inner linker comprises of a single glycine residue (G149) whilst the longer outer linker comprises of a Gly-Gly-Ser tripeptide (sticks: black). **b,** Close up view of the linker region for all three CG6 molecules in the asymmetric unit highlighting the difference in the conformation of the linkers between the different molecules (A, B and C).

observed in the crystal structure (Fig 6.7) but not for the other models (χ^2 = 1.92 or 3.03). The SAXS data support the relevance of the crystallographic data in solution. Therefore, the V-shaped domain arrangement appears to be at least predominant (or even exclusively present) in solution.

6.2.2.4 EGFP-cyt b_{562} domain interface in CG6

The side-by-side domain arrangement of EGFP and cyt b_{562} in CG6 brings the two domains in close proximity to one another creating a domain-domain interface. The interface between the cyt b_{562} and EGFP is remarkably hydrophilic (Fig 6.12) and does not look like a typical naturally evolved protein interaction surface. Altogether, the interface between the cyt b_{562} and EGFP domains extends over an area of approximately 700-800 Å^2 and buries about 1400-1600 Å^2 of solvent accessible surface.

Many favourable contacts are observed between the two domains (Fig 6.12). Among these, at least one salt bridge is consistently formed between one of the haem carboxylate groups and the guanidino group of R181 (R71^{GFP}) (Fig 6.12 a and b). The guanidino group of R181 also forms hydrogen bond interactions with the hydroxyl group of S42 (linker), which is also in hydrogen bonding distance of a haem carboxylate group (Fig 6.12 a). In the second molecule (Chain B) there is a hydrogen bond between the guanidino group of R181 to the main chain O atom of G40 (linker) and from the main chain N atom of S42 (linker) to one of the haem carboxylate groups (Fig 6.12 b). Again depending on the molecule in the asymmetric unit, one of the haem carboxylate groups forms a hydrogen bond with T333 (T223^{GFP}) (Fig 6.12 a). In both molecules there is consistently a hydrogen bond interaction between the guanidino group of R148 (R106^{cytb}) and the amide O of Q312 (Q202^{GFP}) (Fig 6.12 a and b). In both molecules there is consistenly a hydrogen bond formed between the K151 side chain ε-amino group and the main chain O of Q145.

Although there are a high proportion of hydrophilic interactions between EGFP and cyt b_{562} there are also more conventional protein-protein contacts. For example, R148 (R106^{cytb}) stacks favourably against F331 (F221^{GFP}) (Fig 6.12 a and b).

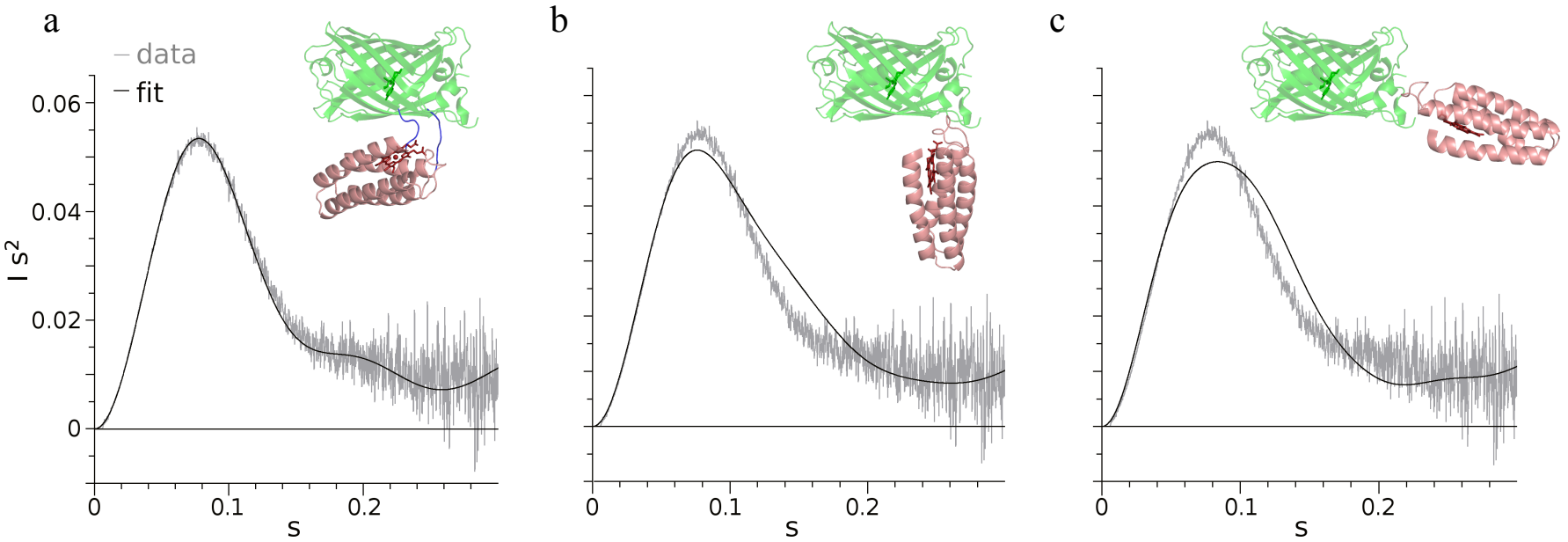


Fig 6.11 Domain arrangement of CG6 in solution determined by SAXS analysis. SAXS experimental intensity data were compared to predicted fits for domain arrangement models of **(a)** a side-by-side arrangement, **(b)** the two domains at right angles to one another and **(c)** for a fully elongated rod arrangement.

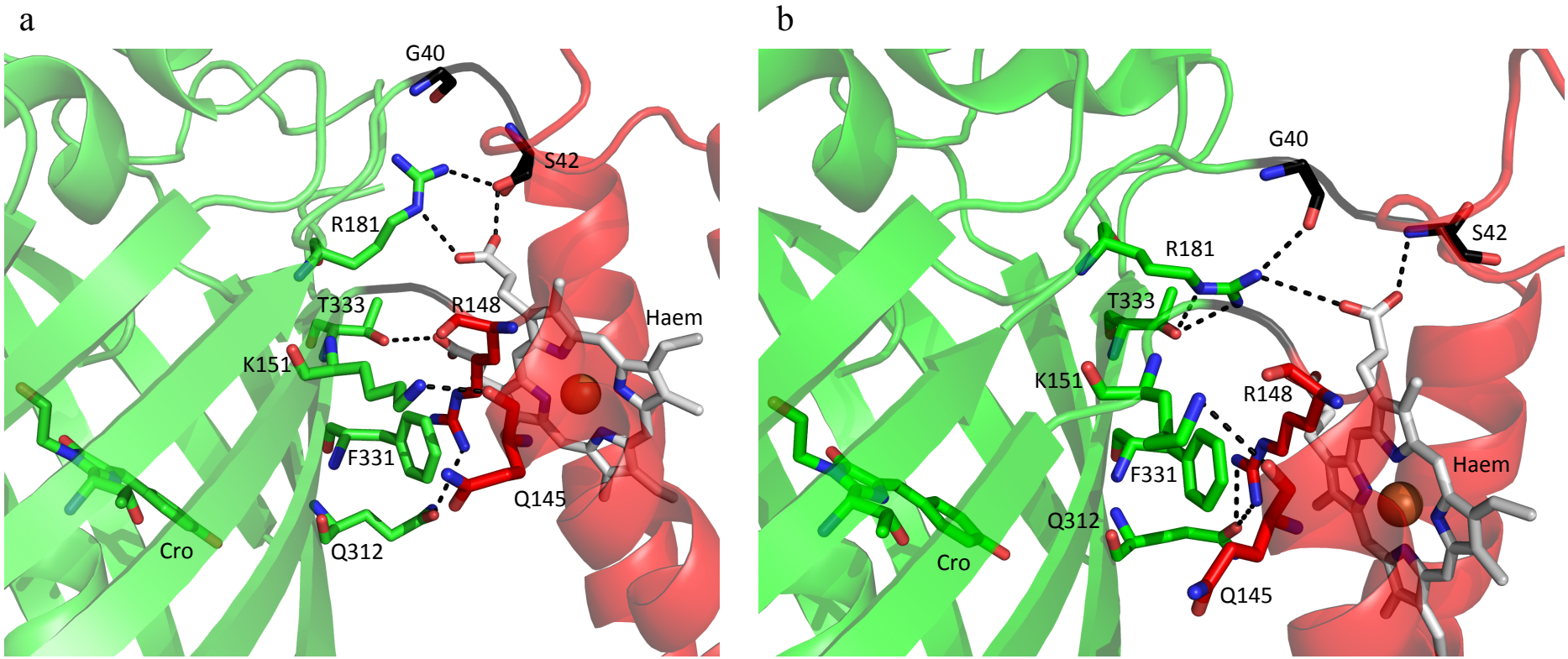


Fig 6.12 Interdomain and linker interactions. The polar interactions (black dashed lines) between residues from the EGFP domain (green sticks), cyt b_{562} domain (red sticks) and the linking residues (black sticks) within **(a)** molecule A or **(b)** molecule B of the asymmetric unit for CG6. Haem is shown as sticks with CPK colouring with the coordinated iron molecule shown as an orange sphere.

6.2.2.5 Effect of cyt b_{562} domain insertion on EGFP chromophore local environment

The crystal structure determined for EGFP (Chapter 4) identified an altered hydrogen-bonding network between the chromophore and residues in its immediate vicinity. In particular a tridental electron density was observed for residue E222 that was successfully modeled by two side chain conformations (Fig 4.3). In both conformations a hydrogen bond between S65 and Y66 observed in wt GFP, was interrupted between the corresponding T65 toY66 of EGFP (Fig 4.3).

Similarly in CG6 the hydrogen bond chain between T175 ($T65^{EGFP}$) and Y176 ($Y66^{EGFP}$) through E330 ($E222^{EGFP}$), S313 ($S205^{EGFP}$) and a conserved water molecule is interrupted between E330 and S313 (Fig 6.13 a). The Fo-Fc electron density map generated after molecular replacement and refinement of CG6 did not exhibit the tridental density observed for E222 in EGFP (Fig 6.13 b). This is probably due to the lower resolution of the CG6 structure (2.75 Å) with respect to the EGFP structure (1.35 Å) determined in Chapter 4. Instead E330 is modeled by a single side chain conformation with hydrogen bonds between its carboxylate group to a conserved water molecule and to the hydroxyl group of S313 (Fig 6.13 a). The hydroxyl group of S313 in turn forms a hydrogen bond chain to Y176 through a conserved water molecule also coordinated *via* a hydrogen bond to the main chain O of N254 ($N146^{EGFP}$) as seen in EGFP.

The same hydrogen-bonding network observed in EGFP (Fig 4.3) between residues N146, H148, R168 and T205 to the chromophore are also present between the corresponding residues in CG6 (Fig 6.13). In CG6 the hydroxyl group of T311 ($T205^{EGFP}$) forms a direct hydrogen bond to the phenolate group of Y176 ($Y66^{EGFP}$) and also forms a hydrogen bond to the main chain N of H256 ($H148^{EGFP}$). A hydrogen bond between the side chain of H148 and the main chain O of R168 seen in EGFP is also seen in CG6 with further interaction by a hydrogen bond formed between the amide group of N254 ($N146^{EGFP}$) and the main chain O of R276 ($R168^{EGFP}$) (Fig 6.13 a). A side chain N of H256 ($H148^{EGFP}$) also forms a direct hydrogen bond interaction with the phenolate of the chromophore. All of the polar interactions between the chromophore phenolate group and surrounding resdidues is sufficient to stabilize a charge on the phenolate, given that E330 ($E222^{EGFP}$) is neutral.

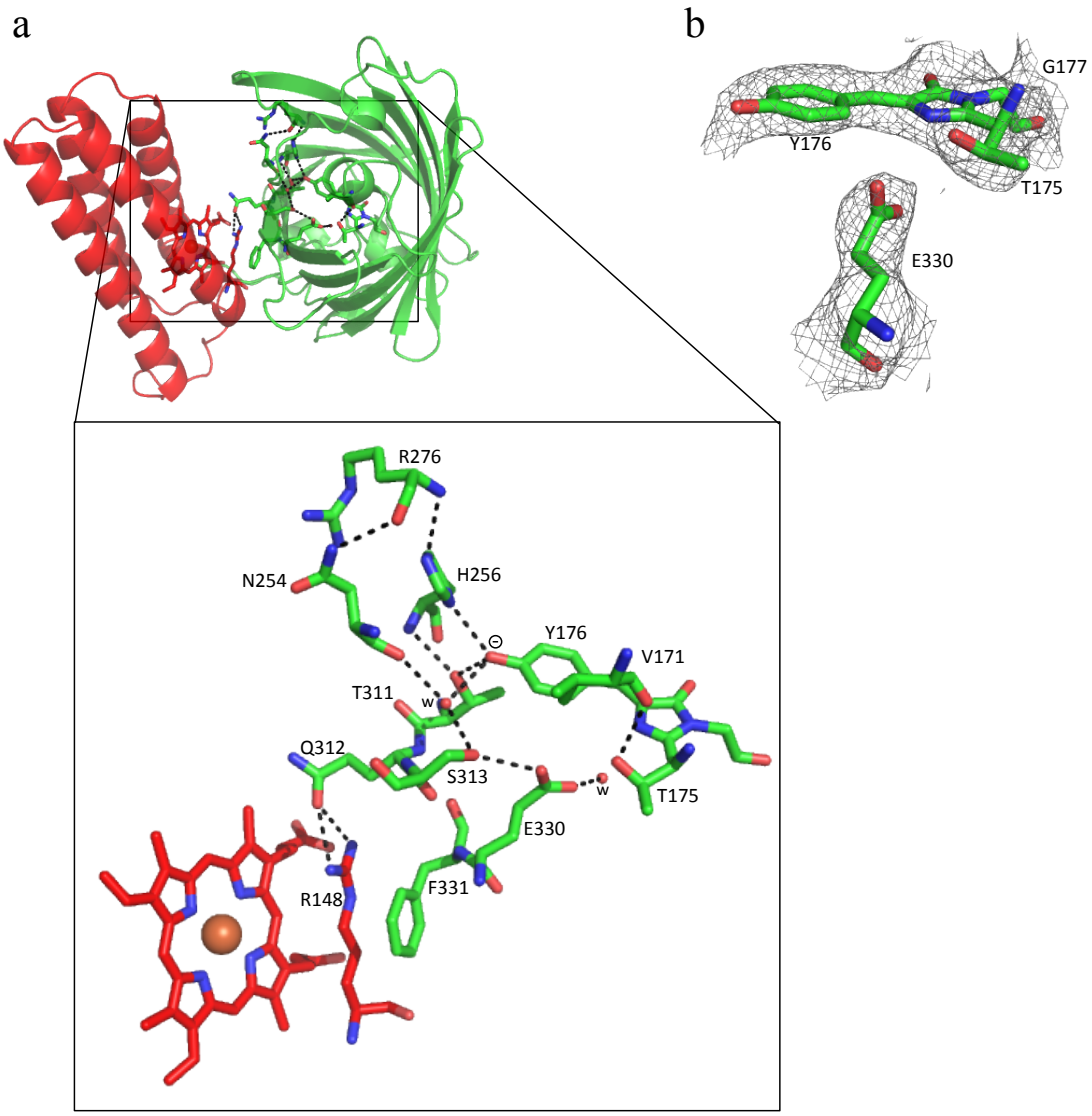


Fig 6.13 EGFP chromophore local environment in CG6. a, The crystal structure for CG6 (molecule B) is shown in cartoon representation with the EGFP domain (green), cyt b_{562} domain (red). Residues making polar interactions (black dashed lines) are shown as sticks coloured corresponding to the domain they originate from. **Inset:** Close up view of residues participating in polar interactions. Conserved water molecules are labeled as w. **b,** F_o-F_c electron density map for the chromophore and residue E330 (sticks) contoured to 1σ.

The residues that contribute to the hydrogen bond network around the chromophore of CG6 are in close proximity to the EGFP cyt b_{562} domain-domain interface with some adjacent residues (Q312 and F331) contributing to the domain interface (Fig 6.13 a). The interaction between residues Q312 ($Q204^{EGFP}$) and F331 ($F223^{EGFP}$) with R148 ($R106^{cytb}$) may be the reason for only a single conformation of the E330 carboxylate group. However, a higher resolution structure would be required to identify if there are two conformations for the E330 side chain or not.

The hydrogen bonding network in both of the well ordered molecules in the asymmetric unit (chain A and chain B) were the same and although there are slight differences in the domain-domain interface between the two molecules (Fig 6.12) it does not effect the chromophore environment.

6.2.2.6 Effect of domain insertion on haem binding to the cyt b_{562} domain.

As described in Section 6.2.1.1 cyt b_{562} domain insertion has very little effect on the overall structure of the EGFP domain or the cyt b_{562} domain in CG6 (Fig 6.8, Table 6.4). In wild-type cyt b_{562} the haem moiety is bound non-covalently, with the haem iron coordinated between the sulphur atom of the M7 side chain and a N atom from the H102 side chain (Fig 6.14 a). In CG6 the haem moiety is coordinated in the same manner between the sulphur atom of M49 and the N atom of H144 (Fig 6.14 b). The only difference between the haem molecule bound to wt cyt b_{562} and CG6 is a slight difference in the position of one of the haem propanoate groups (Fig 6.14).

The major difference between the haem binding environment in wt cyt b_{562} and CG6 is that haem is more surface exposed when bound to wt cyt b_{562} (Fig 6.14 c). In CG6 one side of the haem binding pocket is covered by the EGFP domain, therefore making the haem less solvent accessible (Fig 6.14 d).

6.3 Discussion

Domain insertion provides a general approach for structurally and spatially linking normally disparate proteins so that the function of one protein is coupled to another; it provides a mechanism to artificially construct novel biomolecular switches for use in synthetic biology, biosensing or even as novel energy transducers in bionanotechnology. From a protein engineering perspective, the challenge lies in

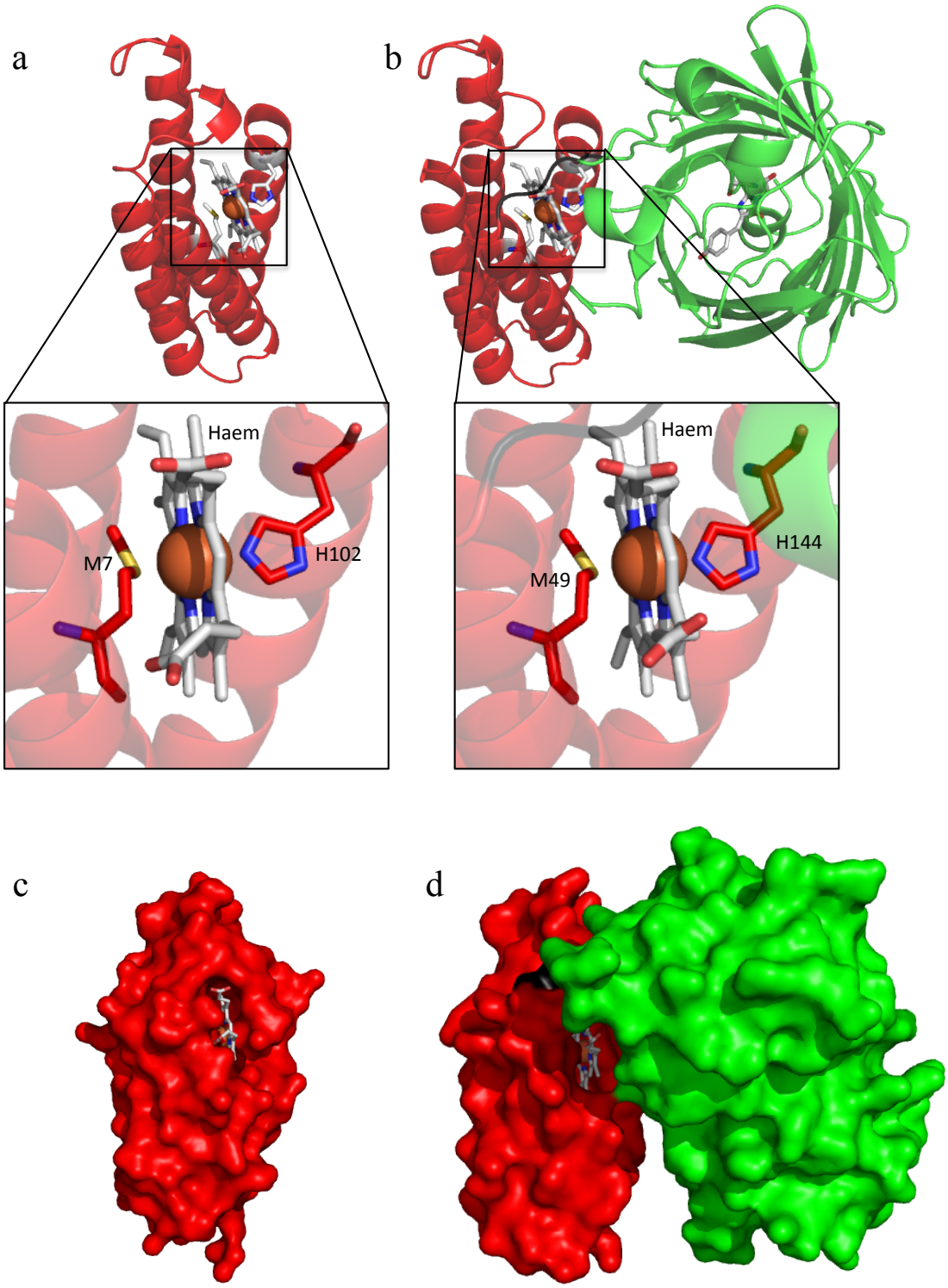


Fig 6.14 Haem binding environment in wild type cyt b_{562} and CG6. a, cartoon representation of wild type cyt b_{562} (red, PDB:256B) with haem and its axial ligands, M7 and H102, shown as sticks. **b,** Cartoon representation of CG6 with the EGFP domain in green, the cyt b_{562} domain in green and the linkers in black. Haem and its axial ligands, M49 and H144, are shown as sticks. **c,** Molecular surface representation for wild type cyt b_{562} with haem shown as sticks. **d,** Molecular surface representation for CG6 (coloured the same as in b), with haem shown as sticks.

predicting sites within the accepting protein that not only tolerate insertion of a whole domain but also couple the functions of the two proteins. The use of directed evolution to sample a diverse range of domain insertion sites and domain linking sequences coupled with screening for linked functionality has provided a useable approach that has met with some success [31, 32, 37]. However, little is known how in these constructed protein scaffolds that the key facet of functional coupling is achieved at the molecular level and how such a domain insertion can influence stability and folding. This in turn hinders our ability to generate suitable scaffold models that could form the basis of designing optimal protein switches. Therefore, retrospective structural analysis coupled with folding studies of directly evolved protein constructs is very valuable.

6.3.1 EGFP and cyt b_{562} domain arrangement in CG6

Domain insertion could be considered to be a largely disruptive mutational event, given that the continuity of the polypeptide chain for the protein accepting the domain insert is broken. Dogma dictates that insertion positions are likely to be restricted to inherently flexible regions such as loops and traditionally rational sequence insertion approaches have focused on such loops [121, 147, 148]. The majority of the cyt b_{562}-EGFP integral fusion proteins identified previously (Chapter 5) conform to this dogma. CG6 on the other hand has cyt b_{562} inserted within a more structurally constricted site; a β-turn linking strands 2 and 3 of EGFP (Chapter 5, Fig 5.5). Insertion at this position did not appear to disrupt the functions of the individual proteins significantly, with CG6 having comparable brightness to EGFP and affinity for haem to cyt b_{562} (Table 5.3).

Three possible EGFP-cyt b_{562} domain arrangements could be envisaged (Fig 6.11 a, b and c): a linear arrangement, one domain lying perpendicular to other or the side-by-side arrangement. The former two would be considered most likely to form as the linkers require a less constrained structure and the two domains are still essentially independent of each other. The side-by-side arrangement would require a defined hinge point structure and even interaction between the domains to maintain this configuration. However, the side-by-side arrangement is the most desirable as the structures and hence functions of the domains are more likely to be intimately linked. The determined structure of CG6 reveals that the EGFP and cyt b_{562} domains take up

the side-by-side arrangement through a molecular hinge at the linker regions (Fig 6.7, Fig 6.9 and Fig 6.10). The observed CG6 domain arrangement is not a crystal artefact as SAXS data confirmed that the V-shape domain arrangement is relevant in solution (Fig 6.11).

Critical to the side-by-side placement of the domains are the linker sequences separating the EGFP and cyt b_{562} domains. The differential sequence length at the two interdomain linking points (Fig 6.10) introduces a constrained molecular pivot point. The single amino acid linker forms the acute inner turn while the longer linker forms the outer turn (Fig 6.10). Without the differential linker lengths it is unlikely that the two domains could assume their relative positions. Gly40 occupies a confined environment and replacement with a longer linker may not be tolerated sterically in the current conformation and introduce additional flexibility between the two domains. Several favourable interactions are also formed between the residues in the longer outer linker to residues from both the EGFP domain and the cyt b_{562} domain, further stabilizing the domain arrangement (Fig 6.12).

The side-by side configuration is reinforced by interdomain interaction. The domain interface is largely comprised of hydrophilic contacts (Fig 6.12), atypical of naturally evolved domain/subdomain interfaces. Residues outside the linking sequences are important contributors to the domain interface. For example, one of these residues, R181 ($R71^{EGFP}$) from the EGFP domain, forms a salt bridge to the haem propionate group. While some limited focus is given to arbitrary linking sequences when constructing linked domains very little consideration is given to domain interactions, as these are generally difficult to anticipate. From a design perspective, we have shown by retrospective analysis of a directed evolved domain insert scaffold that domain placement is essential for maximised functional coupling (in this case energy transfer); linker sequences together with domain interactions are essential in achieving this.

6.3.2. Domain arrangement facilitates energy transfer

Haem quenches fluorescence through energy transfer but critical to transfer efficiency is the positioning of the two chromophores; addition of haem alone to EGFP (Chapter 5, Fig 5.14) or the simple fusion of cyt b_{562} to the termini of EGFP is not sufficient for maximal energy transfer (Chapter 5, Fig 5.9) [33, 82]. The structure

of CG6 provides an explanation for the high efficiency energy transfer between the protein-bound haem and the EGFP chromophore.

The side-by-side domain arrangement forms the characteristic V-shape conformation (Fig 6.7). This domain arrangement results in an interchromophore distance of ~17 Å (Fig 6.9 b). The previously calculated R_0 (the Förster radius at which energy transfer is 50% efficient) for the chromophores of cytochrome b_{562} and EGFP is 46 Å [33]. As energy transfer efficiency (E) is related to the interchromophore donor-acceptor distance (r) through the inverse 6^{th} power by the equation $E= 1/1+(r/R_0)^6$ and r is determined to be ~17 Å (Fig 6.9 b), E is 99.7%. This is in line with observed near total quenching of fluorescence of holo-CG6 (Chapter 5, Fig 5.9).

High energy transfer efficiency is confirmed through analysis of the fluorescence lifetimes in the presence and absence of haem through the relationship $E= 1-(\tau_{holo}/\tau_{apo})$, where τ_{holo} and τ_{apo} represent the lifetimes in the presence and absence of haem. As the τ_{holo} for CG6 is essentially 0 (Chapter 5, Table 5.3), E is 100%. The energy transfer efficiency and the distances between chromophores in this new scaffold are comparable to natural energy transfer systems, such as the light harvesting complexes. EGFP is known to undergo photoinduced electron transfer [149-151] and cyt b_{562} is capable of facilitating conductance over distances of 10-30Å between gold electrode [85, 86]. Thus, the new CG6 protein scaffold may act as a starting point for a simple bio-based light harvesting system that could be integrated with inorganic devices [150].

6.3.3 Effect of cyt b_{562} domain insertion on the chromophore local environment

As mentioned previously extensive hydrogen bonding networks exist between the chromophore and residues in its immediate vicinity in GFP and its structural and functional relatives (Chapter 4) [62, 99]. For green fluorescent proteins this hydrogen-bonding network is responsible for maintaining the protonation state of the chromophore tyrosyl group and is therefore intimately linked to the spectral properties [62, 152].

Crystal structure determination of EGFP (Chapter 4) showed two altered hydrogen bonding networks between the chromophore and surrounding residues with respect to wt GFP (Chapter 4, Fig 4.3). The differential hydrogen bond networks are

formed by two conformations of the E222 carboxylate side chain (Chapter 4, Fig 4.3). Analysis of the probable chromophore hydrogen-bonding network in CG6 was comparable to that of one of the networks observed for EGFP (Fig 6.13). Only a single conformation for the side chain of E330 is modelled in CG6 as the resolution is not high enough to determine if two conformations exist, therefore there appears to be a single predominant hydrogen-bonding network formed, similar to one of the networks in EGFP (Fig 4.3 ai).

Given that residues adjacent to those involved in the chromophore environment hydrogen bond network are part of the domain-domain interface, may play a role in constricting the E330 carboxylate side chain into a single conformation. However, to confirm this notion a higher resolution structure would need to be determined for CG6.

6.3.4 Structural insight into CG6 oxidant sensing properties

The crystal structure of CG6 also provides an insight into some of its observed redox-dependent switching properties. Upon haem binding to the cyt b_{562} domain of CG6 very little structural change was observed (Fig 6.1), therefore the structures for the apo and holo forms could be considered equivalent. The oxidation state-dependent affinity of haem for cyt b_{562} in CG6 is maintained with fluorescence quenching occurring at lower haem concentrations under reducing conditions (Chapter 5, Fig 5.9). As the structure of the cyt b_{562} domain of CG6 is almost identical to that of wt cyt b_{562} (RMSD ~0.75Å over all backbone atoms, Fig 6.8) and haem binds to CG6 in a similar manner to cyt b_{562} (Fig 6.14) it is not unexpected that redox-dependent affinity is retained. The oxidation state dependent haem affinity coupled with the high signal gain on haem dissociation makes CG6 an attractive sensor of changes from the normally reducing conditions inside the cell to oxidising that accompanies several important biological events. Real-time analysis showed that while initial binding and fluorescence quenching under reducing conditions was rapid, on changing to oxidising conditions signal gain was slow (Chapter 5, Fig 5.11 a).

The CG6 structure provides a plausible explanation for the slow haem dissociation kinetics (Chapter 5, Fig 5.11 a). Haem lies at the domain interface making additional interactions with the protein *via* the propionate groups (Fig 6.12) and restricting access to the solvent (Figure 6.14 d). Thus, these structural features are

likely responsible for the low dissociation rate. In comparison, the head-to-tail fusion variant CG1 displays fast H_2O_2 dissociation kinetics (Chapter 5, Fig 5.11 d) that is likely due to the haem being freely accessible to the solvent and probable lack of contacts between the haem propionate groups and residues in either the linkers or the EGFP domain.

6.3.5 Effect of cyt b_{562} domain insertion on EGFP stability

Although several domain insertion scaffolds, with GFP as the accepting domain have been developed, studies on the effects of domain insertion on the stability of the integral domain fusion scaffolds have been neglected. Here, we have shown that despite the insertion of the cyt b_{562} domain into a tight β-turn of EGFP, the free energy of denaturation for the integral domain fusion scaffold, CG6, is comparable to that of EGFP (Table 6.2). The mechanism by which CG6 unfolds however appears to be altered with respect to EGFP as was evident from a 1.2 M decrease in the $[GdmCl]_{50\%}$ value at equilibrium (Table 6.2) and the >two-fold increase in m-value. The increased m-value indicates an increase in unfolding cooperativity with respect to EGFP, implying a single unfolding step, in contrast to previous work on the unfolding of GFPs which have suggested unfolding by a two step process [71, 113].

Chemical induced denaturation of CG6 showed there to be a decrease in the $[GdmCl]_{50\%}$ value (Table 6.2) with respect to EGFP and the terminal fusion CG1. Similarly thermal stability for EGFP appeared to be decreased by cyt b_{562} domain insertion. The melting temperature for CG6 was 10 °C lower than that for EGFP or the terminal fusion CG1. This data agrees with previous studies on the insertion of antibody binding loops into superfolder GFP (sfGFP) with the majority of insertions resulting in decreased thermal stability [147].

6.3.6 Conclusion

We have provided a molecular explanation of how the functions of two normally disparate proteins can be coupled through a directed evolution domain insertion process. The side-by-side domain arrangement results in high efficiency energy transfer from the chromophore of EGFP to the haem moiety of cyt b_{562} comparable to that of natural systems through optimisation of chromophore

arrangement. Critical to the domain organisation are the linker sequences and domain interactions. Whilst domain insertion appeared to have decreased the thermal stability of the EGFP domain and decreased the concentration of GdmCl required to denature the protein, an increase in the cooperativity of unfolding resulted in similar levels of stability to that of EGFP alone. Retrospective structural analysis coupled with folding studies has thus proved valuable in understanding the molecular features of a directly evolved protein scaffold not present in nature and provides the basis by which to engineer future scaffolds for use as novel and useful biological components.

Chapter 7: Rational design of cyt b_{562}-EGFP chimeras to generate novel fluorescent ratiometric redox sensors

7.1 Introduction

Free radicals and other reactive oxygen species (ROS) such as $HO^{\bullet -}$, $O_2^{\bullet -}$ and H_2O_2 are produced across a wide range of physiological processes [153]. An excess of ROS is harmful to cells as it can cause oxidative damage to DNA, lipids and proteins, which can result in cell death [154]. Oxidative stress can be caused endogenously by a build up of ROS as a by-product of aerobic respiration or exogenously caused by exposure to UV light [155] or extremes in environmental conditions such as temperature [155].

ROS also play vital cellular roles as signaling molecules modulating many important processes such as, activation of transcription factors (Apoptosis inducing factor [156], NF-κB and AP-1), protein tyrosine phosphatases, caspases [157], GTPases (Ras) [154] and chaperones [153]. ROS also play an important role in the immune response; upon phagocytosis of bacteria or other cellular debris (produced by apoptosis or necrosis) by macrophages a respiratory burst results in the over production of ROS ($O_2^{\cdot -}$ and H_2O_2) used by the cell to break down and digest the pathogen or cell debris and activate further signaling pathways controlling the inflammatory response [158]. Therefore, given the importance and general interest in cellular redox processes a biosensor for reporting cellular redox changes is highly desired.

Studies have shown that cysteine residues play a key role in redox sensing [159] and can be used as a biosensor trigger for cell redox state; under oxidizing conditions disulphide bonds form whereas the free thiol is retained under normal cellular reducing conditions. Switching between disulphide-thiol forms can cause conformational changes within proteins, which can modulate protein output (function) [160]. For example under oxidizing conditions the bacterial chaperone Hsp-33 undergoes a conformational change due to the formation of two disulphide bridges, causing a metal (Zn^{2+}) binding site to unfold, revealing a hydrophobic patch for unfolding proteins to bind to [161-163]. Likewise formation of intramolecular disulphide bonds in the tetrameric bacterial transcription factor OxyR triggers its ability to bind DNA [164].

The redox sensing properties of cysteine residues have been exploited in fluorescent proteins to generate redox state sensitive fluorescent probes such as roGFPs [165-167]. Two cysteine residues (S147C and Q204C) were incorporated into adjacent β-strands of the β-barrel structure in close proximity to the fluorophore (Fig 7.1) [165-168]. Under oxidizing conditions intramolecular disulphide bonds are formed altering the spectral characteristics of the parent fluorescent protein. Under reducing conditions roGFP2 has a single excitation maxima at 490 nm. Under oxidizing conditions there is a decrease in the amplitude of the 490 nm maxima and the appearance of an excitation maxima at 400 nm [154, 165]. The altered spectral properties upon disulphide formation is due to a shift in the position of the S205 hydroxyl group, altering the hydrogen bond network surrounding the chromophore, therefore modulating the chromophore phenolate protonation state (Fig 7.1).

The redox midpoint potential of roGFP2 was determined to be -272 mV [165]. Further mutations have generated several different variants with varying redox midpoint potentials, ranging from -229 to -287 mV [166], and increased response rates to oxidants and reducing agents [167]. Given that the redox midpoint potentials for proteins catalytically active in establishing the thiol-disulphide equilibrium range from -270 mV to -122 mV [166], there is a clear need to extend the range of midpoint potentials of roGFP like proteins to be used as quantitative reporters. Increased response rates to oxidants such as H_2O_2 are also a desired quality for improved time resolution of signaling events instigated by H_2O_2 bursts [167].

Using the directed evolution approach described in Chapter 3, domain insertion of cyt b_{562} into EGFP between residue N146 and S147 resulted in two variants, CG8 and CG15, having altered spectral properties with respect to EGFP (Chapter 5, Fig 5.7). Both variants had two excitation maxima at ~400 nm and ~488 nm, implying domain insertion had altered the ratio of the EGFP fluorophore in the protonated or deprotonated forms. CG8 and CG15 only differed in primary sequence by two amino acids due to the ORF-3 *cybC* DNA cassette used to generate them (Fig 7.2). CG15 contained the S147C mutation seen in the roGFPs (Fig 7.2). Therefore it was investigated if introduction of a second cysteine residue, in close vicinity to the S147C mutation site of CG15 instilled ratiometric redox sensitive properties. This chapter therefore describes the rational design approach to introducing secondary

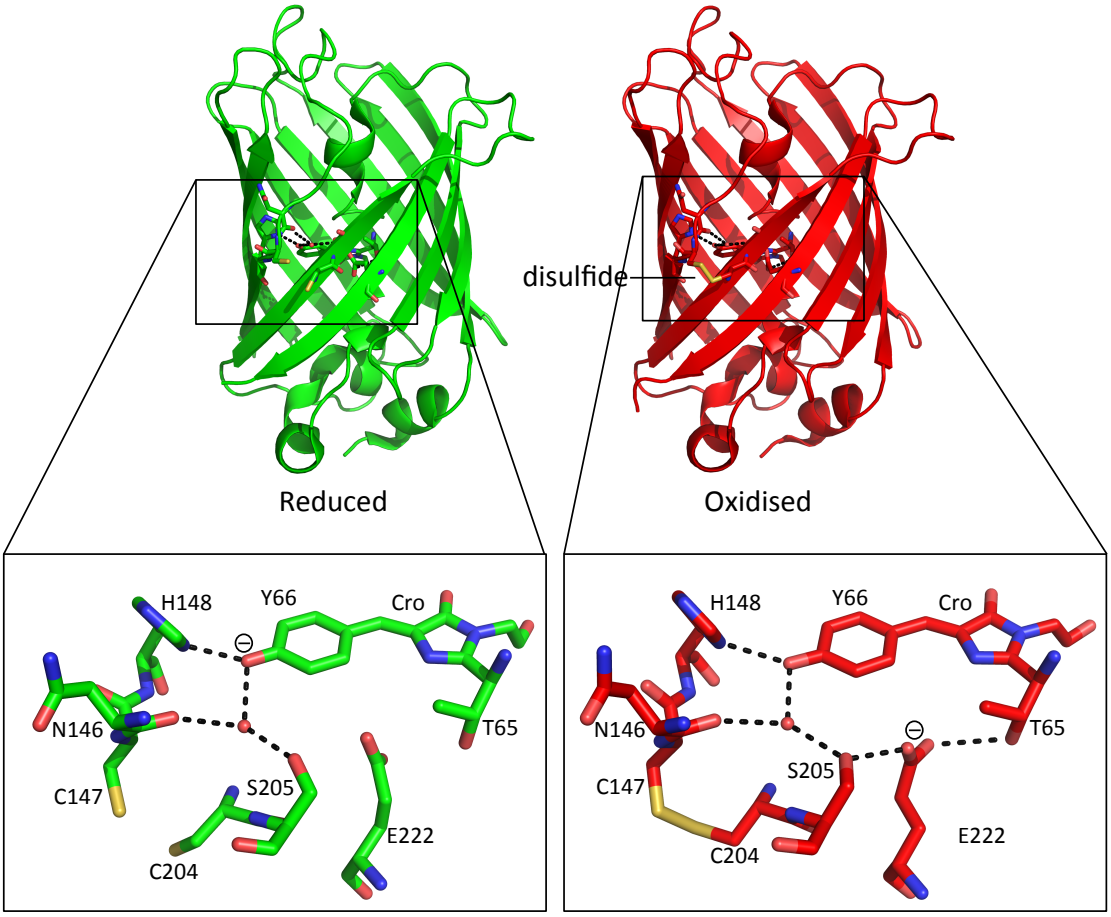


Fig 7.1 Crystal structures of reduced and oxidized roGFP2. Reduced roGFP2 (pdb:1JC0, green cartoon) and oxidized roGFP2 (pdb:1JC1, red cartoon) with residues responsible for forming the hydrogen-bonding network (black dashed lines) with the fluorophore shown in sticks.

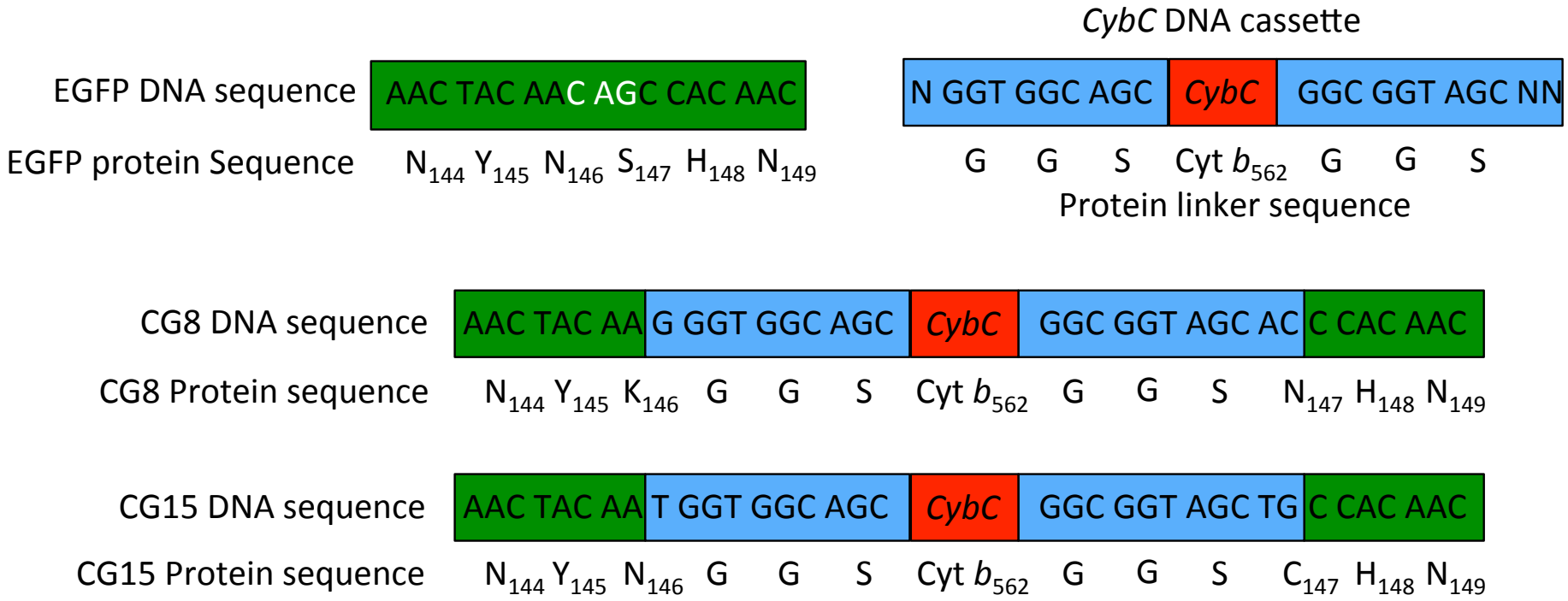


Fig 7.2 DNA and protein sequences for EGFP, inter domain linkers, CG8 and CG15. The triplet nucleotide removed from the *egfp* gene (white) crosses two codons coding for N146 and S147. The *cybC* DNA cassette used to keep the domain insert constructs in frame contains a single random nucleotide (N) and two random nucleotides (NN) at the 5' and 3' end of the cassette respectively. The random nucleotides reconstitute the codons coding for residues at positions 146 and 147 of EGFP, resulting in the substitution mutations N146K and S147N in CG8 and S147C in CG15.

cysteine mutations to the cyt b_{562}-EGFP integral domain fusion scaffold, CG15. Spectral and biophysical characterization has been performed to characterise the redox sensing properties of four different CG15 double cysteine mutants ($CG15^{CC}$).

7.2 Results

7.2.1 Rational design of CG15 double cysteine mutants

Rational design approaches to protein engineering often require that a prior knowledge of the protein structure and function is known to make informed decisions of which residues to mutate. Given that there is currently no determined crystal structure for CG15, identifying which residues to mutate to cysteines was performed using a molecular model. The crystal structure for the cyt b_{562} (pdb:256B) domain was inserted into the crystal structure for EGFP (determined in Chapter 4) between residues N146 and S147 (Fig 7.3). The two domains were separated by GlyGlySer linkers at the N- and C-terminal ends of the cyt b_{562} domain (Fig 7.3). Residue S147 was then mutated to a cysteine residue and the CG15 model was energy minimized using the in built auto-sculpture tool in MacPyMol.

Using the CG15 molecular model it was possible to design which secondary residues to mutate to cysteines with the potential of forming disulphide bonds with C147 (Fig 7.3). Residue C147 is positioned with its side chain on the solvent exposed side of the EGFP β-barrel. Therefore the criteria for residues selected as potential candidates for disulphide bond formation with C147 are: (1) they should be in close proximity to C147 and (2) their side chains must be on the solvent exposed side of the β-barrel. Although the CG15 structure is a molecular model and represents only a possible conformation and arrangement of the two domains with respect to one another it was possible to identify four residues that fit the required criteria. Ideally, the switch between the free thiol and disulphide bond should alter the hydrogen bond network associated with the chromophore and thus change the ratio of absorbance at 400 nm and 488 nm.

The first selected was N146, which is next to C147 in the primary sequence of EGFP but separated by the cyt b_{562} insertion in CG15. In the CG15 model, these two residues lie adjacent to one another (Fig 7.3). Mutation of N146 to a cysteine and

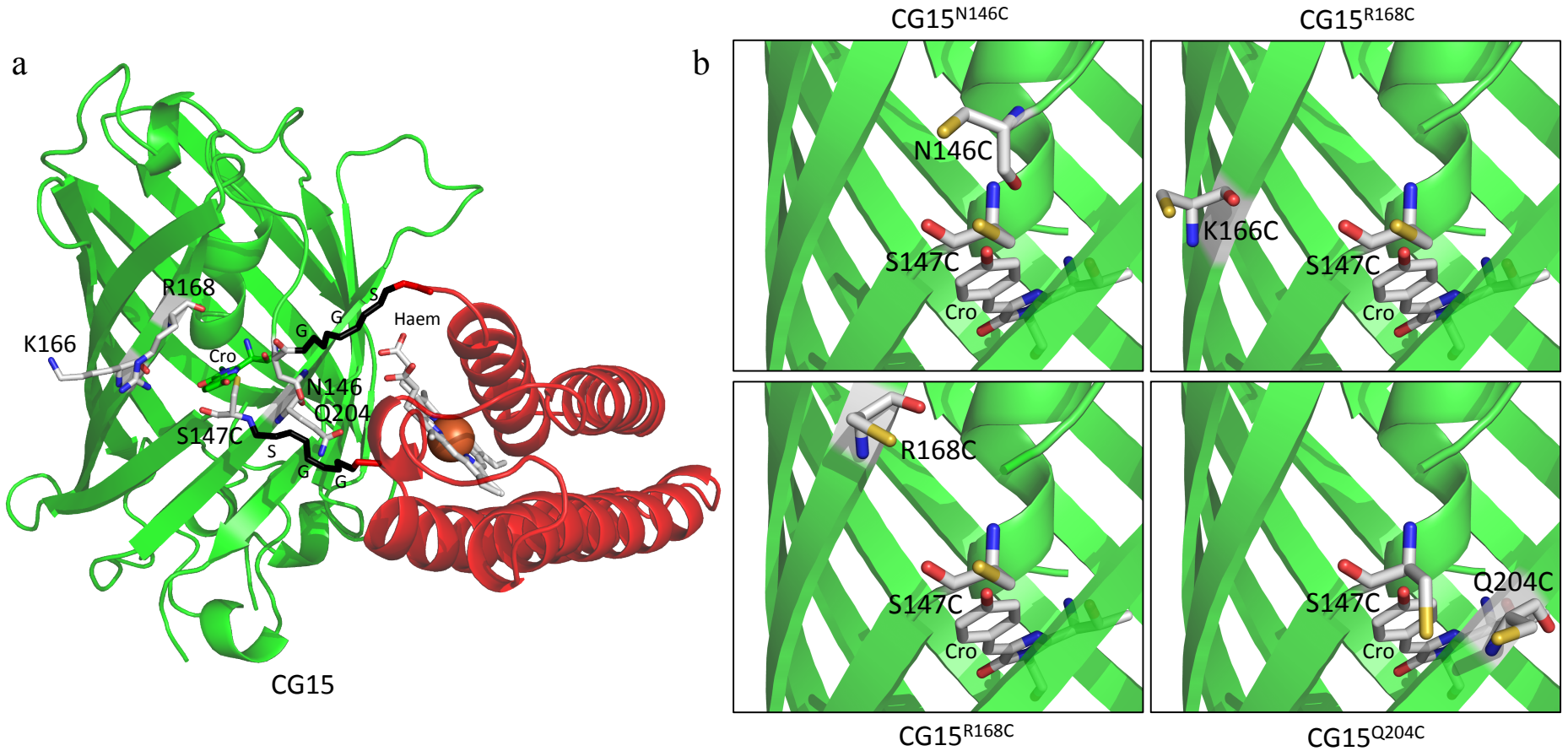


Fig 7.3 Model of CG15 structure and CG15 cysteine substitution positions. a, Model of CG15 with the EGFP domain (green cartoon) and the haem (CPK sticks) bound cyt b_{562} domain (red cartoon) linked together by GlyGlySer linkers (GGS: black sticks) between residues N146 and S147 of EGFP. CG15 with the substitution mutation S147C (CPK sticks) together with residues selected by rational design to be mutated to cysteines (CPK sticks) highlighted. Cro stands for chromophore. **b,** Close up view of CG15 double cysteine models with the cyt b_{562} domain and linkers removed for clarity. CG15 models were made using MacPyMol.

formation of a disulphide bond with C147 would bridge the gap between the N-terminal end of the GlyGlySer linker, at the N-terminus of the cyt b_{562} domain, and the C-terminal end of the GlyGlySer linker, at the C-terminus of cyt b_{562} domain. Formation of a disulphide bond between these two residues would potentially reconnect the break in β-strand 7 caused by the insertion of cyt b_{562}.

Two residues a single amino acid apart from one another in β-strand 8, adjacent to β-strand 7 containing residue C147, were also selected (K166 and R168). Both residues have side chains external to the β-barrel of the EGFP domain and are in close proximity (~8 Å) to C147 in the CG15 molecular model (Fig 7.3). The last residue selected was Q204, which is the same residue used in the roGFPs. The Q204C mutation in CG15 would act as a control and comparison to roGFP2, to see the effect a cyt b_{562} domain insertion has on the redox sensing ability of this double cysteine mutant.

The cysteine mutations were introduced into the gene encoding CG15 by site directed mutagenesis (Section 2.4.3). Combinations of primers were used to introduce TGC codons, coding for cysteine, at the required positions within the CG15 gene to produce variants $CG15^{N146C}$, $CG15^{K166C}$, $CG15^{R168C}$ and $CG15^{Q204C}$. DNA sequence analysis of $CG15^{CC}$ genes, isolated from *E. coli* BL21 (DE3) Gold cells (Table 2.1), confirmed the presence of the desired mutations.

The $CG15^{N146C}$, $CG15^{K166C}$, $CG15^{R168C}$ and $CG15^{Q204C}$ variants were expressed in *E. coli* TUNER™ (DE3) cells grown at 20 °C in M9 minimal media (Section 2.1.4) supplemented with 100 μg/ml ampicillin and 150 μM IPTG. Expressed protein was purified as previously described for cyt b_{562}-EGFP integral domain scaffolds (Chapter 5, section 5.2.4). The purified samples were used for all spectral characterization.

7.2.2 Effect of double cysteine mutations on CG15 spectral properties.

7.2.2.1 Analysis of the UV-visible absorption properties of CG15 and variants under reducing and oxidising conditions.

To determine if the new combination of cysteine residues in each of the variants introduced redox sensitive properties into the basic CG15 scaffold, UV-Vis absorbance spectra were measured in the presence of either an oxidant (H_2O_2) or reductant (DTT). Each of the mutants have altered absorption spectra with respect to

EGFP (For EGFP spectra see Chapter 4, section 4.2.3) and CG15 (Fig 7.4). All variants retained the two absorption maxima of ~400 nm and ~490 nm observed for CG15 but with differing intensity ratios at the two λ_{max} values, dependent on the presence of DTT or H_2O_2. Variants $CG15^{N146C}$, $CG15^{K166C}$ and $CG15^{R168C}$ displayed a similar change to their absorption spectra with the 400 nm peak decreasing and the ~490 nm peak increasing in intensity in the presence of H_2O_2 (Fig 7.4). The magnitude of change in the absorption spectra does, however, differ between $CG15^{N146C}$, $CG15^{K166C}$ and $CG15^{R168C}$ with the highest magnitude change observed for $CG15^{N146C}$ and the smallest observed for $CG15^{K166C}$ (Fig 7.4). The absorption spectra for $CG15^{Q204C}$ changes in the opposite direction to the other $CG15^{CC}$ mutants; in the presence of H_2O_2 the ~400 nm absorption peak increases in amplitude whilst its ~490 nm absorption peak decreases (Fig 7.4)

Molar absorption extinction coefficients were calculated for CG15, $CG15^{N146C}$, $CG15^{K166C}$, $CG15^{R168C}$ and $CG15^{Q204C}$ at both absorption maxima (~400 nm and ~488 nm) under reducing (10 mM DTT) and oxidizing conditions (0.02% (v/v) H_2O_2) (Table 7.1). Under reducing conditions the molar absorption extinction coefficients at ~400 and ~490 nm for all of the double cysteine mutants are very similar. This implies that under reducing conditions every $CG15^{CC}$ variant has a similar ratio of chromophore in the protonated and deprotonated state. However, the difference in molar absorption extinction coefficients, at ~400 and ~490 nm, under oxidizing conditions showed greater variation (Table 7.1). This is indirect evidence that formation of disulphide bonds between the two cysteines in each variant is modulating the ratio of chromophore in the protonated or deprotonated state, therefore altering the absorption spectra.

The most significant observed difference between the variants was under oxidising conditions (Table 7.1). Under oxidizing conditions the molar absorption extinction coefficient for $CG15^{N146C}$ at 490 nm increased to 19,000 $M^{-1}cm^{-1}$ from 14,200 $M^{-1}cm^{-1}$ and red shifted by 5 nm from 485 nm (Table 7.1). The molar absorption extinction coefficient at 403 nm decreased to 8,600 $M^{-1}cm^{-1}$ from 12,200 $M^{-1}cm^{-1}$ and blue shifted by 5 nm (Table 7.1). This is a change in ~400 nm:490 nm ratio of 2.6-fold (δ values, Table 7.1).

Under oxidising conditions, $CG15^{K166C}$ displayed a 6 nm blue shift in the λ_{max}

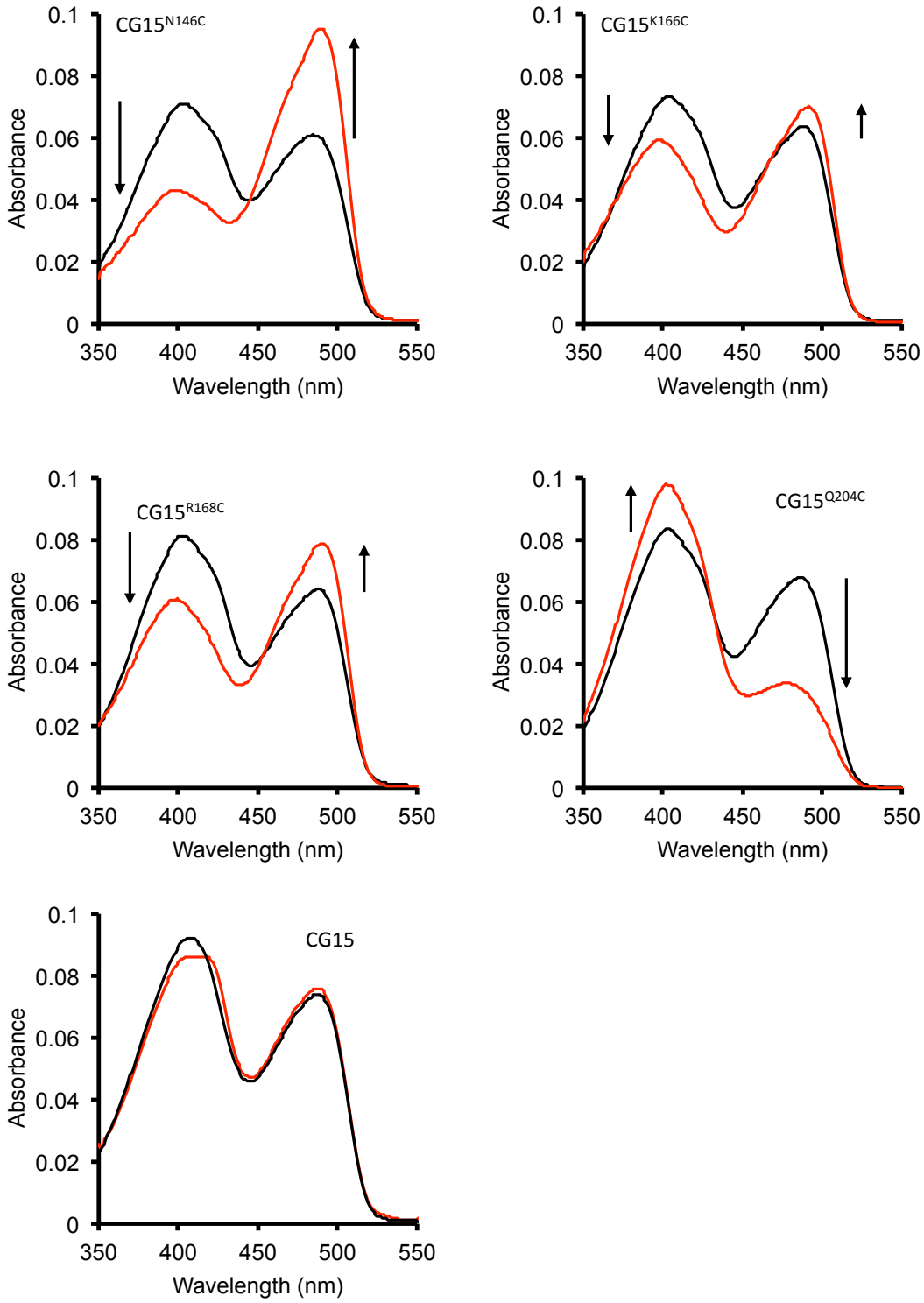


Fig 7.4 UV-visible absorption spectra for CG15 and variants. UV-visible absorption spectra for 5 μM CG15 and $CG15^{CC}$ varianst (as indicated in the figure) were measured in the presence of either 10 mM DTT (black line) or 0.02% (v/v) H_2O_2 (red line) to induce reducing or oxidizing conditions, respectively. Black arrows indicate the direction of change in the amplitudes of the absorption peaks from reducing to oxidizing conditions.

Table 7.1 UV-visible absorption spectral characteristics

Variant	λmax (nm)[a]		ε ($mM^{-1}cm^{-1}$)[b]		δ_{abs}[c]	Holo-chimera λmax (Ox/Red)[d] (nm)
	Red	Ox	Red	Ox		
CG15	416 (487)	408 (487)	17.3 (15.2)	18.4 (14.8)	1.1	417/426
$CG15^{N146C}$	403 (485)	490 (398)	14.2 (12.2)	19.0 (8.6)	2.6	417/426
$CG15^{K166C}$	403 (488)	492 (397)	14.6 (12.7)	14.0 (11.8)	1.4	416/426
$CG15^{R168C}$	402 (488)	490 (399)	16.2 (12.8)	15.8 (12.2)	1.6	416/426
$CG15^{Q204C}$	403 (486)	402 (478)	16.7 (13.6)	19.4 (6.5)	2.4	417/426
Cyt b_{562}	-	-	-	-	-	417/428

[a] Red and Ox stand for reduced and oxidized respectively and are shown on a white or grey background respectively. λmax for absorption maxima with lower amplitudes are shown in brackets and red text.
[b] Millimolar absorption extinction coefficients calculated at λmax values stated for each variant.
[c] Dynamic range value is the maximum δ-fold change in absorption peak ratio
[d] λmax values determined in the presence of equimolar haem

at 403 nm to 397 nm and a decrease in the molar absorption extinction coefficient from 14,600 $M^{-1}cm^{-1}$ to 11,800 $M^{-1}cm^{-1}$. Simultaneously the molar absorption extinction coefficient at 488 nm (12,700 $M^{-1}cm^{-1}$) increased to 14,000 $M^{-1}cm^{-1}$ and exhibited a red shift in the λ_{max} by 4 nm to 492 nm (Table 7.1). This is a change in the ~400 nm:490 nm ratio of 1.4-fold (δ values, Table 7.1).

Variant $CG15^{R168C}$ showed a change in molar absorption extinction coefficients in between that of $CG15^{N146C}$ and $CG15^{K166C}$, and showed only very minor red and blue shifts in the λ_{max} values at ~400 and ~490 nm between oxidizing and reducing conditions (Table 7.1). Its molar absorption extinction coefficient at ~400 nm decreased from 16,200 $M^{-1}cm^{-1}$ to 12,200 $M^{-1}cm^{-1}$, whilst at ~490 nm the molar absorption extinction coefficient increased from 12,800 $M^{-1}cm^{-1}$ to 15,800 $M^{-1}cm^{-1}$, equating to a 1.6-fold change in excitation ratio.

$CG15^{Q204C}$ switched 403 nm and 490 nm ratio in the opposite direction to those of the other variants (Fig 7.4). Under oxidizing conditions the molar absorption extinction coefficient at ~400 nm exhibits an increase to 19,400 $M^{-1}cm^{-1}$ from 16,700 $M^{-1}cm^{-1}$ under reducing conditions. Simultaneously the extinction coefficient at ~490 nm decreases from 13,600 $M^{-1}cm^{-1}$ to 6,500 $M^{-1}cm^{-1}$ (Table 7.1). In the presence of DTT (10 mM) the less intense of the two absorption maxima has a λ_{max} of 486 nm. In the presence of H_2O_2 (0.02% (v/v)) this λ_{max} blue shifts by 8 nm to 478 nm, whilst the λ_{max} for the major maxima essentially remains the same under reducing and oxidizing conditions. (Table 7.1).

In comparison to all the double cysteine mutants, CG15 exhibited only minor changes in molar absorption extinction coefficients; 18,400 $M^{-1}cm^{-1}$ at 408 nm to 17,300 $M^{-1}cm^{-1}$ and 14,800 $M^{-1}cm^{-1}$ at 487 nm to 15,200 $M^{-1}cm^{-1}$ under oxidizing and reducing conditions respectively (Table 7.1). The observation that CG15 exhibited only very small changes in molar absorption extinction coefficients under oxidizing and reducing conditions implies that a second cysteine residue is required in close proximity to C147 in order to elicit a significant change in the spectral properties. CG15 did however exhibit an 8 nm blue shift in its λ_{max} from 416 nm under reducing conditions to 408 nm under oxidizing conditions, whilst its λ_{max} value for its low intensity absorption maxima at 487 nm remained the same under both reducing and oxidizing conditions (Table 7.1).

7.2.2.2 The haem binding properties of CG15^{CC} variants.

As shown in section 5.2.4 (Fig 5.9), haem quenches CG8 fluorescence. Therefore, it is expected that CG15 and its associated variants should retain their ability to bind haem. To probe the capacity of CG15^{N146C}, CG15^{K166C}, CG15^{R168C} and CG15^{Q204C} variants to bind haem, the UV-visible absorption spectra were determined in the presence of haem (Fig 7.5). Absorption spectra of purified protein samples were measured in the absence or presence of equimolar haem under reducing and oxidizing conditions (Fig 7.5). All the variants exhibited absorption spectra similar to that of wild type holo-cyt b_{562}, with λ_{max} values of 426 nm and ~417 nm under reducing and oxidizing conditions respectively (Table 7.1), implying the haem binding environment has not been altered by the cysteine substitution mutations.

Given that the fluorescence of the cyt b_{562}-EGFP chimeras is quenched when haem binds to the cyt b_{562} domain and ratiometric redox sensing requires a fluorescent signal, no further spectroscopic analysis on the effect of haem on the spectral properties of each CG15^{CC} variant was performed.

7.2.2.3 Effect of redox state on the fluorescence properties of CG15 and the CG15^{CC} variants.

To determine the influence of redox conditions on the fluorescence spectra of CG15 and the CG15^{CC} variants, purified protein samples were incubated for 24 hr at 25 °C with either DTT or H_2O_2, to induce reducing or oxidizing conditions respectively. The buffer was degassed prior to use and the samples incubated in an airtight container under anaerobic conditions (Section 2.6.1.10), to minimize any effects of air oxidation.

The changes to the excitation spectra for each variant mirrored those observed for the absorbance spectra (Fig 7.6). Under oxidizing conditions variants CG15^{N146C}, CG15^{K166C} and CG15^{R168C} all showed an increase in intensity for the ~490 nm excitation maxima and a decrease in the ~400 nm excitation maxima intensity (Fig 7.6). Again in contrast, CG15^{Q204C} exhibited the opposite change in excitation spectra with the ~400 nm excitation maxima increasing and the ~490 nm excitation maxima decreasing in intensity in the presence of oxidant. As with the absorption spectrum for CG15, the excitation spectrum also shows a very small change (~10%) in the excitation maxima intensity ratio (Fig 7.6, Table 7.2).

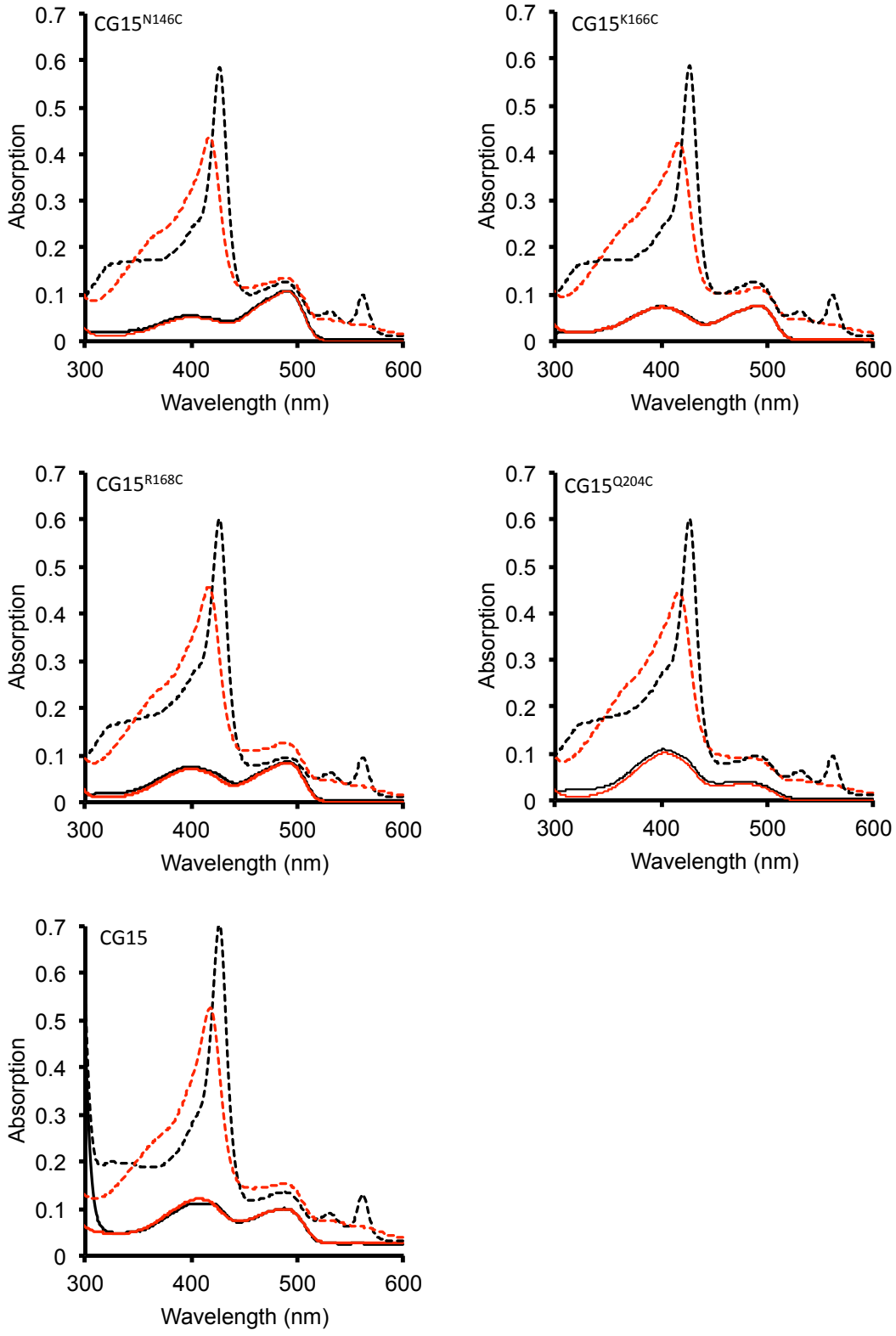


Fig 7.5 UV-visible absorption spectra of holo-CG15 and holo-CG15^{CC} variants. UV-visible absorption spectra of 5 μM holo-CG15 or holo-CG15^{CC} variants (as indicated in the figure were measured in the presence of 1 mM ascorbic acid (black line) or 1 mM KNO_3 (red line) for reducing and oxidizing conditions respectively, in the absence (solid lines) and presence (dashed lines) of equimolar haem.

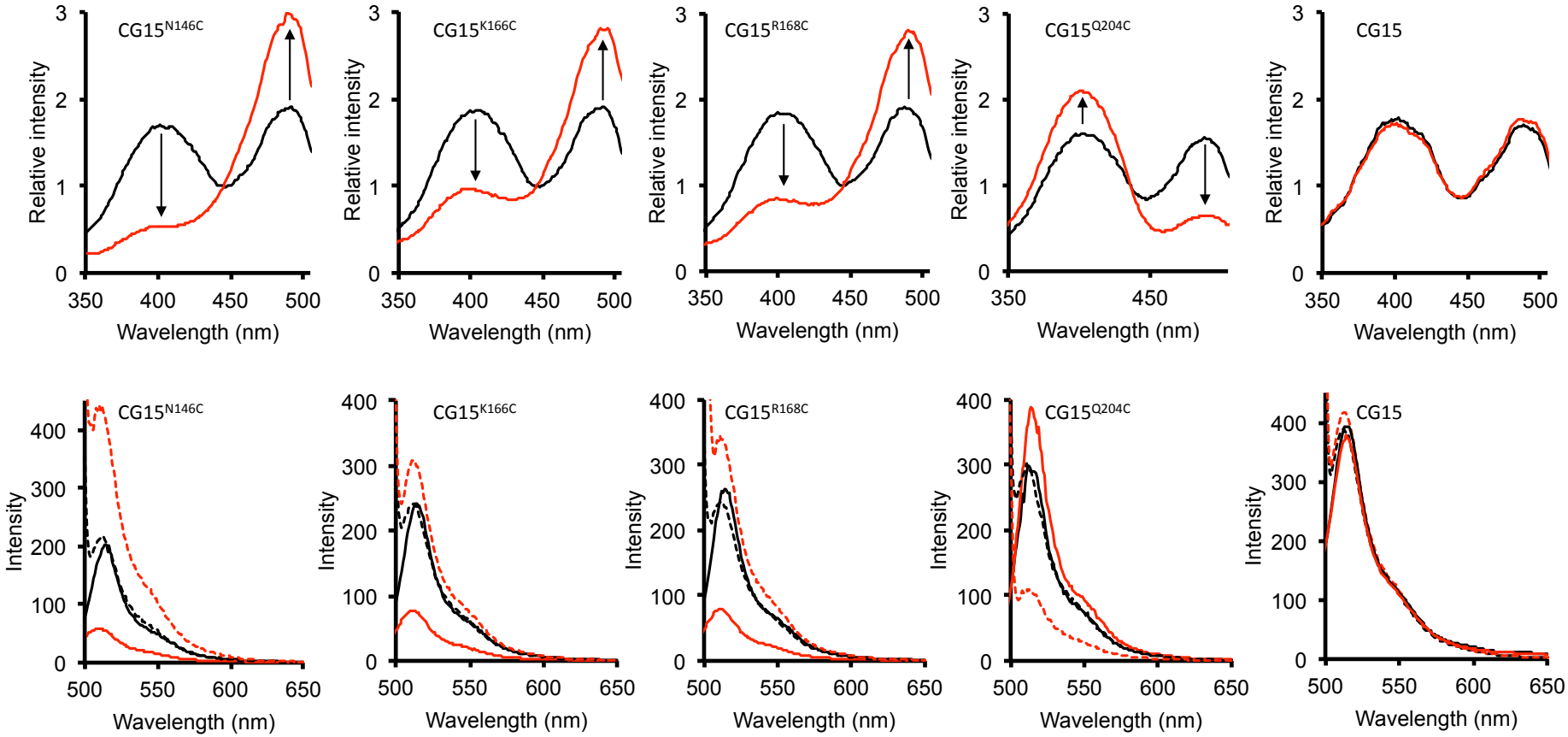


Fig 7.6 Excitation and emission spectra for CG15 and CG15^{CC} variants. All fluorescence spectra were measured with 50 nM protein under reducing (black lines) and oxidizing (0.003% v/v H_2O_2; red lines) conditions. Excitation spectra (top panels) were measured whilst monitoring fluorescence emission at 530 nm. Emission spectra (bottom panels) were measured after excitation at either 400 nm (solid lines) or 490 nm (dashed lines). Excitation spectra were normalized to a value of 1 at their isosbestic point.

Table 7.2 Fluorescence spectral properties

Variant	Reduced[a]		Oxidised[a]		λ_{iso} (nm) 530nm em	δ_{ex} [b]
	λ_{ex} (nm)	λ_{em} (nm)	λ_{ex} (nm)	λ_{em} (nm)		
CG15	402 (489)	514 (513)	488 (401)	514 (513)	435	1.1
$CG15^{N146C}$	492 (401)	512 (515)	489 (399)	509 (511)	445	5.0
$CG15^{K166C}$	492 (399)	512 (514)	490 (399)	513 (511)	446	2.9
$CG15^{R168C}$	488 (399)	511 (514)	490 (400)	511 (512)	445	3.2
$CG15^{Q204C}$	397 (488)	512 (511)	401 (490)	514 (513)	436	3.1

[a] λ_{ex} and λ_{em} values in brackets and red text are for excitation/emission maxima with the lower intensity

[b] λ_{iso} is the isosbestic point wavelength determined from excitation spectral scans whilst monitoring emission at 530 nm

[c] Dynamic range value is the maximum δ-fold change in excitation peak ratio

Under reducing conditions, the λ_{em} for all the variants red shifted by ~2-4 nm on excitation at ~400 nm compared to the λ_{em} on excitation at ~490 nm (~512 nm) (Table 7.2). Although there are also minor shifts in the λ_{ex} for all variants between oxidising and reducing conditions, all the excitation peak ratios for the $CG15^{CC}$ variants were calculated from intensity values at either 400 nm or 490 nm. For $CG15^{N146C}$, $CG15^{K166C}$ and $CG15^{R168C}$ a 490:400 nm ratio was used where as for $CG15^{Q204C}$ a 400:490 nm ratio is used, given that its excitation maxima switch in the opposite direction to the other variants. The ratio of the maximum (in the presence of H_2O_2) to minimum (in the presence of DTT) values of the excitation ratios is defined as the dynamic range, and is a useful measure of the maximum signal to noise ratio [165].

The dynamic range for the roGFPs vary from 2.4 – 7.8 dependent on the mutations incorporated to the parent protein (usually roGFP1 or roGFP2). The dynamic ranges for $CG15^{K166C}$, $CG15^{R168C}$ and $CG15^{Q204C}$ are all towards the lower range with δ_{ex} values of 2.9, 3.2 and 3.1 respectively (Table 7.2). $CG15^{N146C}$ in contrast has a higher dynamic range (δ_{ex} = 5) than the other variants (Table 7.2), which is more comparable to the value for roGFP2 (δ_{ex} = 5.8).

7.2.3 Effect of cysteine substitution on CG15 oligomeric state.

The introduction of additional surface exposed cysteine residues into CG15 runs the potential risk of forming intermolecular rather than intramolecular disulphide bonds. Formation of intermolecular disulphide bonds could alter the excitation maxima ratio or spectral properties of the $CG15^{CC}$ variants hindering their use as efficient ratiometric redox sensors. It was therefore important to determine the oligomeric state of $CG15^{N146C}$, $CG15^{K166C}$, $CG15^{R168C}$ and $CG15^{Q204C}$ by analytical size exclusion chromatography.

A standard curve for the relationship between molecular weight and elution volume from a Superdex™ 200 gel filtration column was determined (Appendix C) using the BioRad gel filtration standards, as described previously (Chapter 6, section 6.2.1.6). From this the elution volume for the $CG15^{CC}$ variants at different concentrations under oxidizing and reducing conditions, could be related to the molecular weight of that species.

The $CG15^{CC}$ samples were applied to the column at either 10, 25, 50 μM concentrations under reducing or oxidizing conditions. The elution volumes (Table 7.3) were determined from the absorbance peaks of the elution profiles (Fig 7.7). The elution volumes for all the major elution peaks for all variants only differed by up to ~0.1 ml, equivalent to ~1200 Da. The average calculated molecular weights for all the variants (~40 – ~41 kDa) agreed closely to their theoretical molecular weights (~39 kDa, calculated from their primary sequence) under reducing and oxidizing conditions, except for $CG15^{K166C}$ (Table 7.3).

At the highest tested protein concentration (50 μM) under oxidizing conditions $CG15^{K166C}$, a very low intensity peak was observed at 13.73 ml, which is equivalent to an estimated molecular weight of ~76 kDa (Table 7.3). This agrees closely to the theoretical molecular weight for a $CG15^{K166C}$ dimer (~78 kDa). The difference in absorption between the elution peak at 13.731 ml and ~15.06 ml however would indicate an estimated monomer:dimer ratio of ~1000:36 confirming that the dimer is a minority form. Given that a protein species elutes at a smaller elution volume under oxidizing conditions but not reducing conditions it is probable that the dimer is formed by an intermolecular disulphide bond rather than a tendency for $CG15^{K166C}$ to dimerize naturally at higher concentrations (>50 μM).

Although $CG15^{K166C}$ exhibits the tendency to form intermolecular disulphide bonds at high concentrations (>50 μM) the redox sensing properties could still be assessed as 50-fold less protein (1 μM) is used when determining the standard redox midpoint potentials (see 7.2.4). It is highly unlikely that intermolecular disulphide bonds will be formed under oxidizing conditions at such a low concentration.

Despite the potential dimerization of $CG15^{K166C}$ at higher concentrations the elution profiles for $CG15^{N146C}$, $CG15^{K166C}$ and $CG15^{R168C}$ are essentially symmetrical and do not exhibit the elution volume shift observed for wild type EGFP with increasing concentration (Chapter 4, Fig 4.11). This implies that the inherent weak tendency of EGFP to dimerize at high concentrations has potentially been eliminated by insertion of the cyt b_{562} domain between residues N146C and S147C of EGFP. In contrast the elution profile for $CG15^{Q204C}$ is asymmetric with a hip in the profile with an estimated elution volume ~15.7 ml, equivalent to a calculated molecular weight of ~29.5 kDa. The protein species eluting at this volume

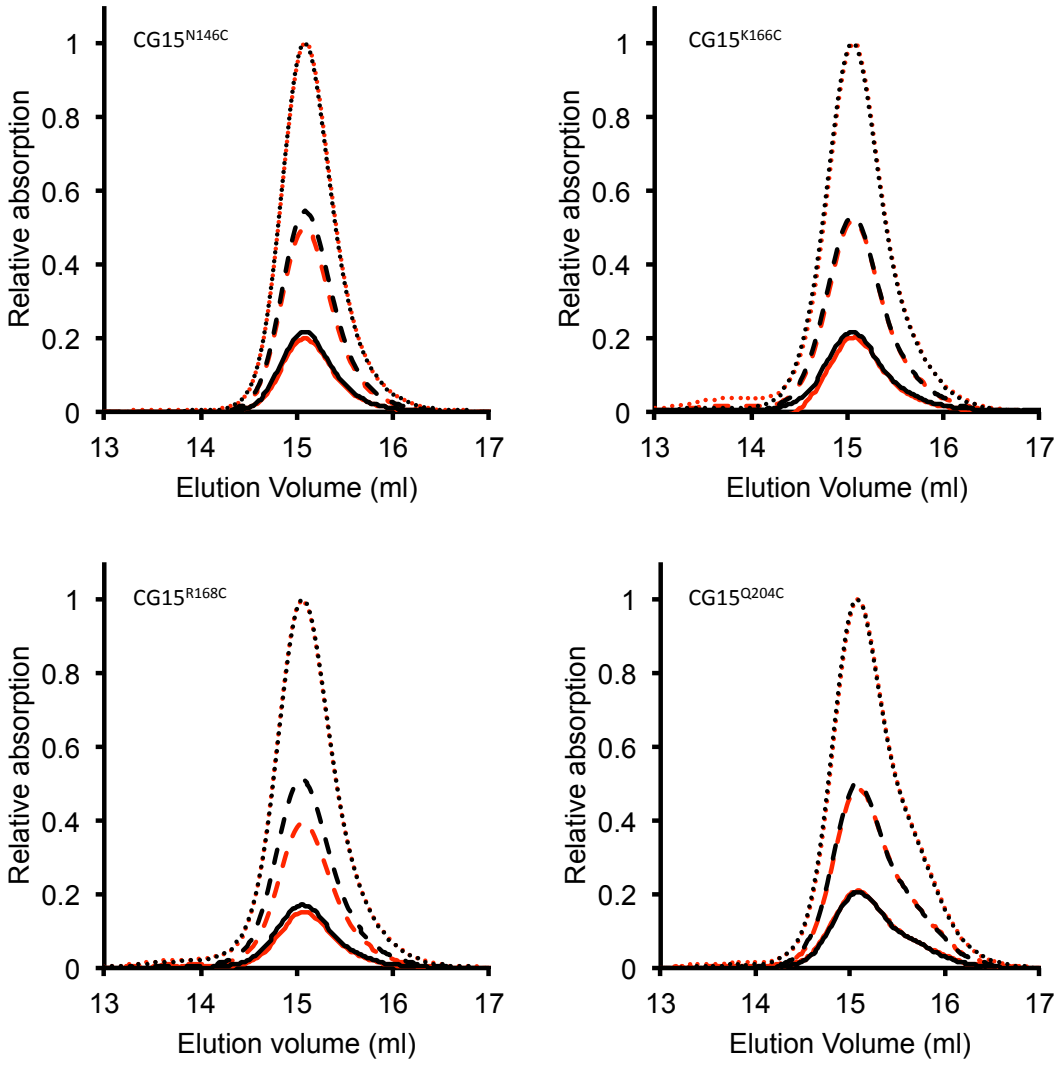


Fig 7.7 Size exclusion chromatography of CG15^{CC} variants. Samples of variants (as indicated in figure) were applied to a Superdex^{TM} 200 gel filtration column and the elution of the protein samples monitored by absorbance at 488 nm for CG15^{N146C}, CG15^{K166C} and CG15^{R168C} or 400 nm for CG15^{Q204C}. Protein concentrations of 10 µM (solid black line), 25 µM (dashed line) or 50 µM (dotted line) were applied to the column under reducing (1 mM DTT; black) or oxidizing (0.0034% (v/v) H_2O_2; red) conditions. The elution of CG15^{N146C}, CG15^{K166C} and CG15^{R168C} from the column was monitored, by absorbance at 488 nm, and CG15^{Q204C} monitored at 400 nm. At these wavelengths all variants have an appreciable absorbance under oxidizing and reducing conditions.

Table 7.3 Size exclusion chromatography analysis

Variant	Redox state[a]	Concentration (μM)	Elution Volume (ml)[b]	Calculated Mw (Da)[c]	Average Mw (Da)[d]	Estimated Mw (Da)[e]
$CG15^{N146C}$	Ox	10 uM	15.1	40580	40667	39109
		25 uM	15.1	40730		
		50 uM	15.1	40690		
	Red	10 uM	15.1	40710	40623	39111
		25 uM	15.1	40580		
		50 uM	15.1	40580		
$CG15^{K166C}$	Ox	10 uM	15.1	40660	40927	39095
		25 uM	15.1	41040		
		50 uM	15.1	41080		
		50 uM	13.7	76680	76680	78190
	Red	10 uM	15.0	41450	41267	39097
		25 uM	15.1	41040		
		50 uM	15.1	41310		
$CG15^{R168C}$	Ox	10 uM	15.1	40660	41047	39067
		25 uM	15.1	41290		
		50 uM	15.1	41190		
	Red	10 uM	15.0	41410	41203	39069
		25 uM	15.1	41160		
		50 uM	15.1	41040		
$CG15^{Q204C}$	Ox	10 uM	15.1	40260	40523	39095
		25 uM	15.1	40620		
		50 uM	15.1	40690		
	Red	10 uM	15.1	40620	40707	39097
		25 uM	15.1	40750		
		50 uM	15.1	40750		

[a]Ox and Red stand for oxidized and reduced respectively
[b]Elution volumes determined from peak absorbance at either 488 or 400 nm (Fig 7.7) dependent on variant (see main text)
[c]Molecular weights calculated using standard curve (Appendix C)
[d]Average calculated molecular weights over the different concentrations
[e]Estimated molecular weight from primary sequence

is present at all concentrations under reducing and oxidizing conditions and is potentially a contaminating protein remaining after purification or more likely a degradation product.

7.2.4 Determination of $CG15^{CC}$ redox midpoint potentials

One of the key parameters of any cellular redox biosensor is redox midpoint potential and its relationship with cellular values. To determine the redox midpoint potentials of $CG15^{CC}$ variants, titration of redox buffers was performed. Given that the relative intensities of the two distinct λ_{ex} values (Fig 7.6) are dependent on the solution oxidation state, titration with DTT redox buffers allows the redox midpoint potential to be determined (Fig 7.8). Using the excitation maxima ratio between 490 nm and the isosbestic point (Table 7.2) for $CG15^{N146C}$, $CG15^{K166C}$ and $CG15^{R168C}$ or $CG15^{Q204}$ between 400 nm and the isosbestic point, the proportion of $CG15^{CC}$ protein in the reduced form can be calculated in different buffers. The buffers themselves comprise of different ratios of oxidized DTT (DTT_{ox}) and reduced DTT (DTT_{red}) (total concentration of DTT_{red} + DTT_{ox} = 1 mM in ratios of 10:0 to 0:10) (Fig 7.7).

The redox midpoint potentials for the $CG15^{CC}$ variants were determined from the equilibrium constant obtained by fitting a theoretical curve to a plot of the fraction of reduced protein *versus* the concentration ratio of reduced DTT to oxidized DTT (Fig 7.9) (see section 2.6.1.10 for methods and equations used).

Given that size exclusion chromatography confirmed the formation of intramolecular disulphide bonds only (see 7.2.3), at a protein concentration of 1 μM, the redox equilibria data could be fit to a two-state model (Section 2.6.1.10, Equation 8) (Fig 7.9). The redox potential of the $CG15^{CC}$ variants at pH 7.0 and 30 °C could then be calculated using the Nernst equation (Section 2.6.1.10, Equation 9) with the standard reduction potential value for DTT ($E'_{0(DTT)}$) of -323 mV [169]. The calculated redox midpoint potentials for all the $CG15^{CC}$ variants were more reducing than that for roGFP2 (-272 mV)[165], and ranged from -301 to -319 mV (Table 7.4). In particular variant $CG15^{N146C}$ has a redox midpoint potential of -319 mV, which is the most reducing midpoint potential described to date for a ratiometric fluorescent redox sensing protein.

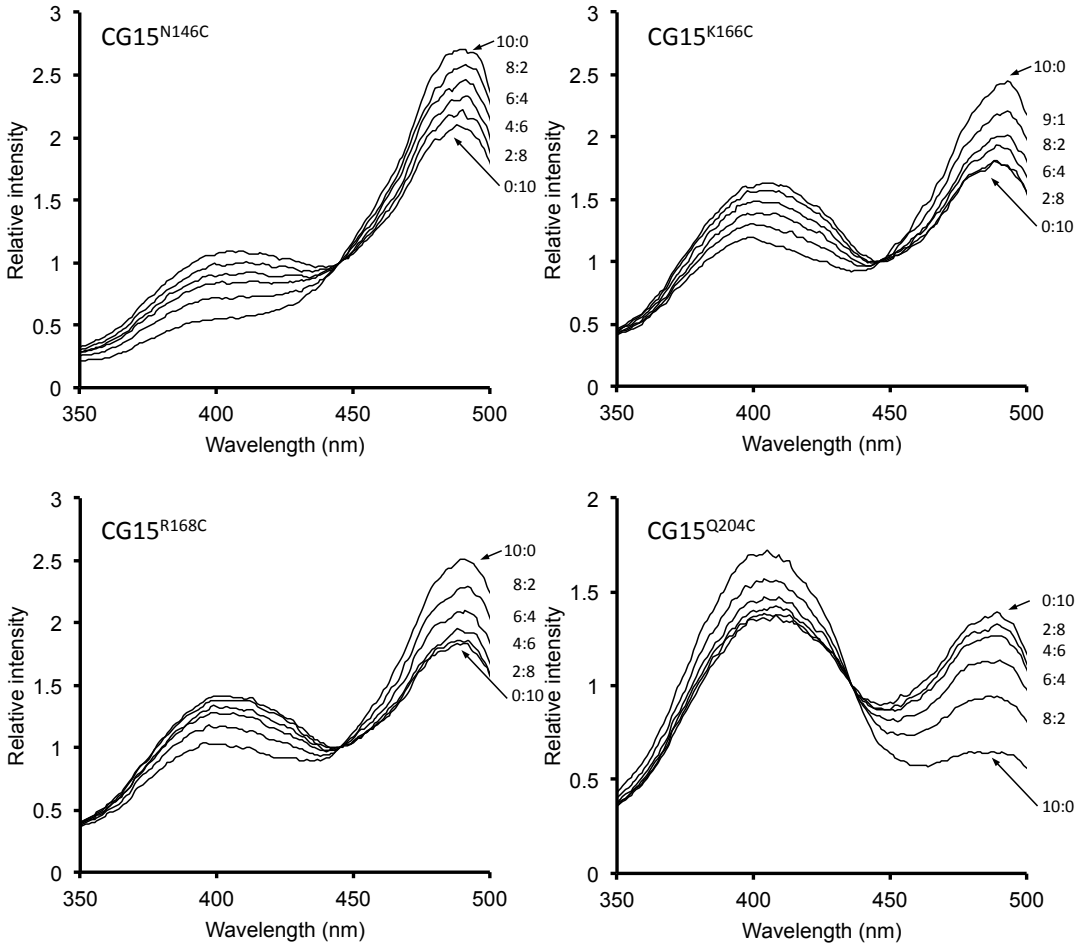


Fig 7.8 Excitaton spectra of CG15^{CC} in various DTT redox buffers. Excitation spectra of CG15^{CC} variants (as indicated in the figure) whilst monitoring emission at 530 nm in various DTT redox buffers (total concentration of DTT_{red} + DTT_{ox} = 1 mM with DTT_{ox}:DTT_{red} ratios of 10:0 to 0:10)

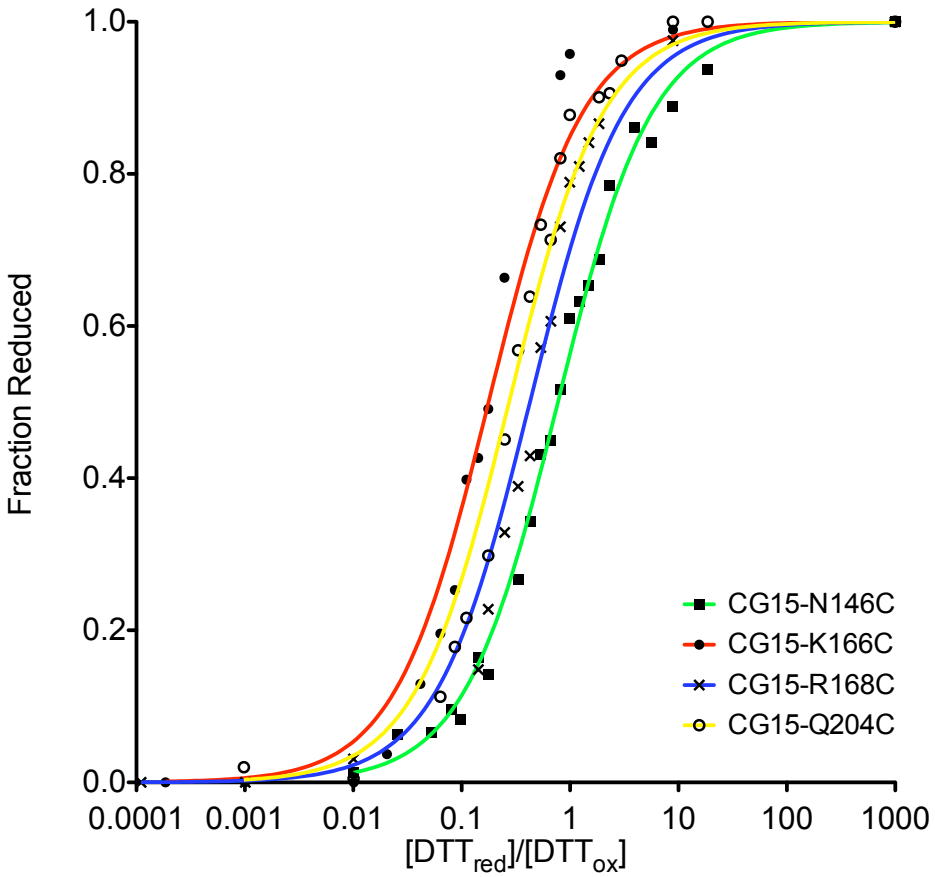


Fig 7.9 Redox midpoint determination of CG15^{CC} variants. Apparent redox midpoint potentials were determined by plotting the fraction of reduced protein against the $[DTT_{red}]$:$[DTT_{ox}]$ ratios and fitting the data to titration curves.

Table 7.4 Redox properties of CG15^{CC} variants

Variant	E_o' (mV)[a]	k_{DTT} (min^{-1})[b]	$k_{H_2O_2}$ (min^{-1})[c]
$CG15^{N146C}$	-319±4	0.59±0.02	2.48±0.11
$CG15^{K166C}$	-301±5	0.47±0.01	0.65±0.01
$CG15^{R168C}$	-311±5	0.12±0.00	0.69±0.01
$CG15^{Q204C}$	-306±5	0.28±0.00	1.05±0.02

[a] Redox midpoint potential calculated from titration against 1mM DTT (Fig 7.10)
[b] Reduction rate constant
[c] Oxidation rate constant

7.2.5 CG15^{CC} variant redox kinetics

In order for high temporal sensing of changes in redox state the kinetics of oxidation and reduction are important factors to consider. *In vitro* rates for the reaction between CG15^{CC} variants with either DTT or H_2O_2 were determined by the change in fluorescence excitation ratio over time after the addition of excess oxidizing (0.0034% (v/v) H_2O_2) or reducing (1 mM DTT) agent. To assure the starting protein samples were either fully oxidized or fully reduced, 10 μM protein samples were incubated in either 1 mM DTT or H_2O_2 for 2 hours at 30 °C. The protein samples were then diluted to a final concentration of 0.5 μM in fresh redox agent free buffer.

The fraction of reduced protein was plot against time and the data fit to single exponential functions (Section 2.6.1.11, Equation 10) to determine pseudo first-order rate constants (Fig 7.10). The reduction rate constants (k_{DTT}) for the CG15^{CC} variants ranged from ~0.12 - ~0.59 min^{-1} (Table 7.4) and were comparable to published values for different roGFP variants (0.10-0.66 min^{-1})[167]. CG15^{R168C} has the slowest k_{DTT} value of 0.12 min^{-1} (Table 7.4), similar to the reduction rates for the majority of the roGFP variants (0.10-0.31 min^{-1}). CG15^{Q204C} has a k_{DTT} rate constant 2.3-fold faster than that of CG15^{R168C} at 0.28 min^{-1} (Table 7.4). CG15^{N146C} and CG15^{K166C} have the fastest reduction rate constants of 0.59 and 0.47 min^{-1} respectively (Table 7.4), similar to the fastest reduction rates measured for the roGFP variants (0.32-0.68 min^{-1}).

The oxidation rate constants ($k_{H_2O_2}$) for CG15^{K166C} and CG15^{R168C} are very similar at 0.65 and 0.69 min^{-1}, whilst CG15^{Q204C} was oxidized slightly quicker with a $k_{H_2O_2}$ value of 1.04 min^{-1} (Table 7.4). Variant CG15^{N146C} was oxidized up to 3.8-fold faster than the other variants as was evident from its $k_{H_2O_2}$ value of 2.48 min^{-1} (Table 7.4). This is also a faster observed response to H_2O_2 than any other roGFP studied to date, with the fastest until now being 2.05 min^{-1} [167].

7.2.6 CG15^{CC} variant pH sensitivity

Although the S65T mutation of EGFP confers it with ~six-fold increase in fluorescence intensity at ~488 nm [72] it has also been shown to make several fluorescent proteins pH sensitive [170]. Changes in pH alter the spectral properties of the S65T GFPs and at pH < 5.0 rapid loss of fluorescence is observed [170]. Cyt b_{562} domain insertion into EGFP between residues N146 and S147 could also have

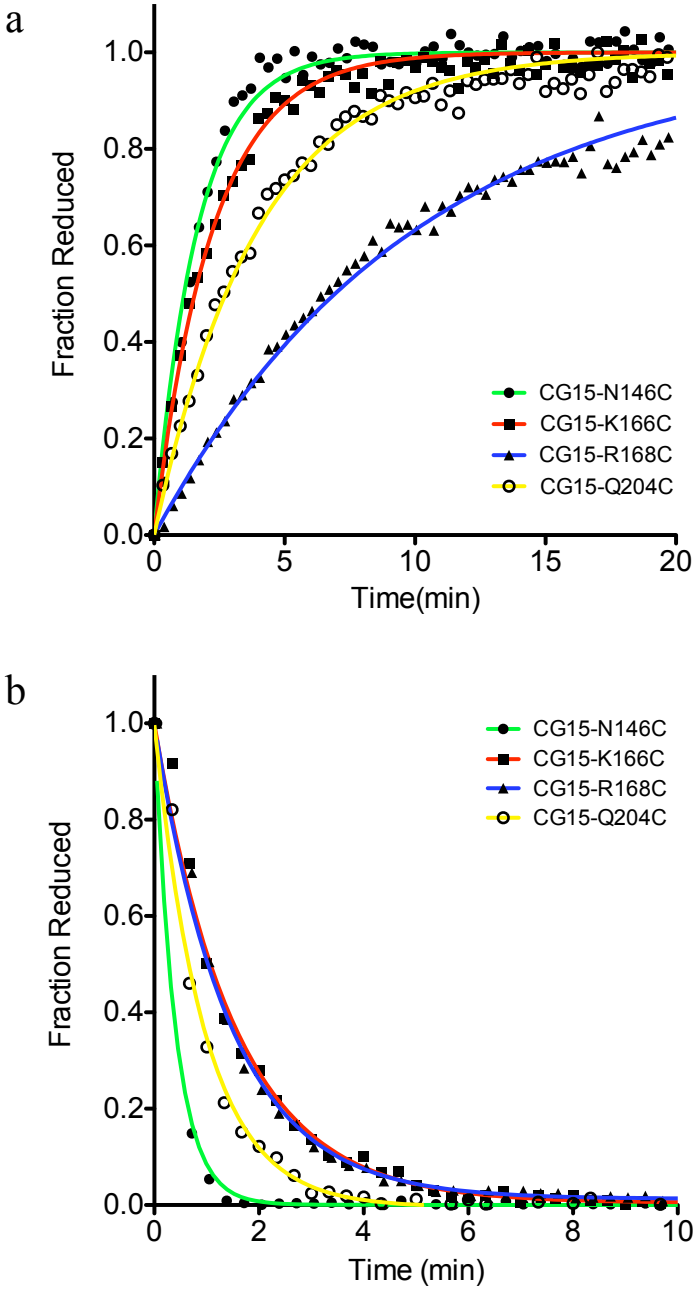


Fig 7.10 Oxidation and reduction kinetics of CG15^{CC} variants. The fraction of reduced protein determined from excitation maxima ratios was monitored over time after the addition of an excess of **a,** 1 mM DTT or **b,** 1mM H_2O_2. Fluorescence excitation intensities at 400 and 490 nm were measured over 20 mins while monitoring emission at 530 nm, after the addition of DTT or H_2O_2 to initiate reduction or oxidation respectively. Pseudo first-order rate constants were determined by fitting the data to single exponential functions.

increased the chromophores sensitivity to changes in pH. Given that the $CG15^{CC}$ mutants have the S65T mutation in the EGFP domain and any redox state measurements rely on the spectral properties of the proteins the pH sensitivity was investigated.

Samples of each variant were incubated for 24 hrs in phosphate buffers with a range of pHs (4 – 9), with either 1 mM DTT or 1 mM H_2O_2 to induce reducing or oxidizing conditions respectively. All buffers were degassed and the samples were incubated in an airtight container under anaerobic conditions to reduce any effects by air oxidation. Excitation spectra were measured whilst monitoring fluorescence emission at 530 nm (Fig 7.11). The excitation intensity decreased at both excitation maxima for all variants with decreasing pH, under reducing and oxidizing conditions (Fig 7.11).

The 490:400 nm excitation ratio for $CG15^{N146C}$, $CG15^{K166C}$ and $CG15^{R168C}$ or 400:490 nm ratio for $CG15^{Q204C}$ were also plot against pH and showed that under reducing conditions the excitation ratios did not significantly alter across the pH range (Fig 7.12). However, under oxidizing conditions all variants exhibited a dependence between the excitation ratios and decreasing pH (Fig 7.12). $CG15^{K166C}$ and $CG15^{R168C}$ both showed the biggest dependence between pH and their excitation ratios under oxidizing conditions, with the excitation ratio decreasing steadily below a pH = 9.0. $CG15^{Q204C}$ also showed a decrease in its excitation ratio below a pH = 8.0, although not to the same degree as $CG15^{K166C}$ or $CG15^{R168C}$. $CG15^{N146C}$ was the most pH stable of all the variants only showing a decrease in the excitation ratio under oxidizing conditions below a pH = 7.0.

7.3 Discussion

In this Chapter, a series of ratiometric redox biosensors that can be genetically encoded have been produced and have the potential to be used within the cell. The ability to construct custom genetically-encoded biomolecular switches whose properties change in response to a desired input opens new possibilities for generating sensitive, non-invasive cell-based biosensors to report on important biological processes. Such biosensors have a range of potential applications, including monitoring cellular events to the screening of therapeutic drugs.

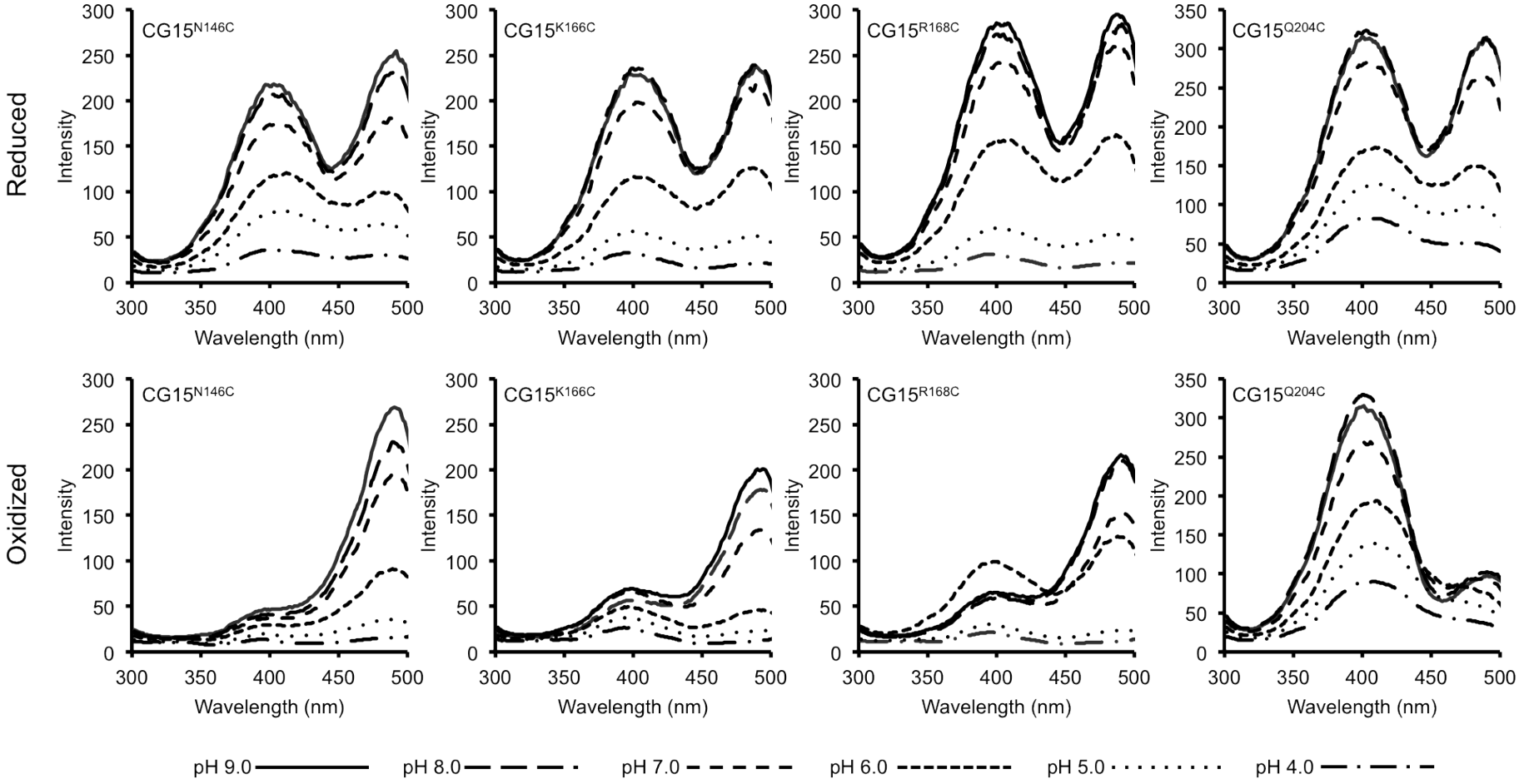


Fig 7.11. Effect of pH on CG15^{CC} excitation spectra. Excitation spectra for the CG15^{CC} variants (as indicated in figure) at different pHs (as indicated in figure) under reducing (1 mM DTT, top panels) or oxidising (0.0034% (v/v) H_2O_2, bottom panels).

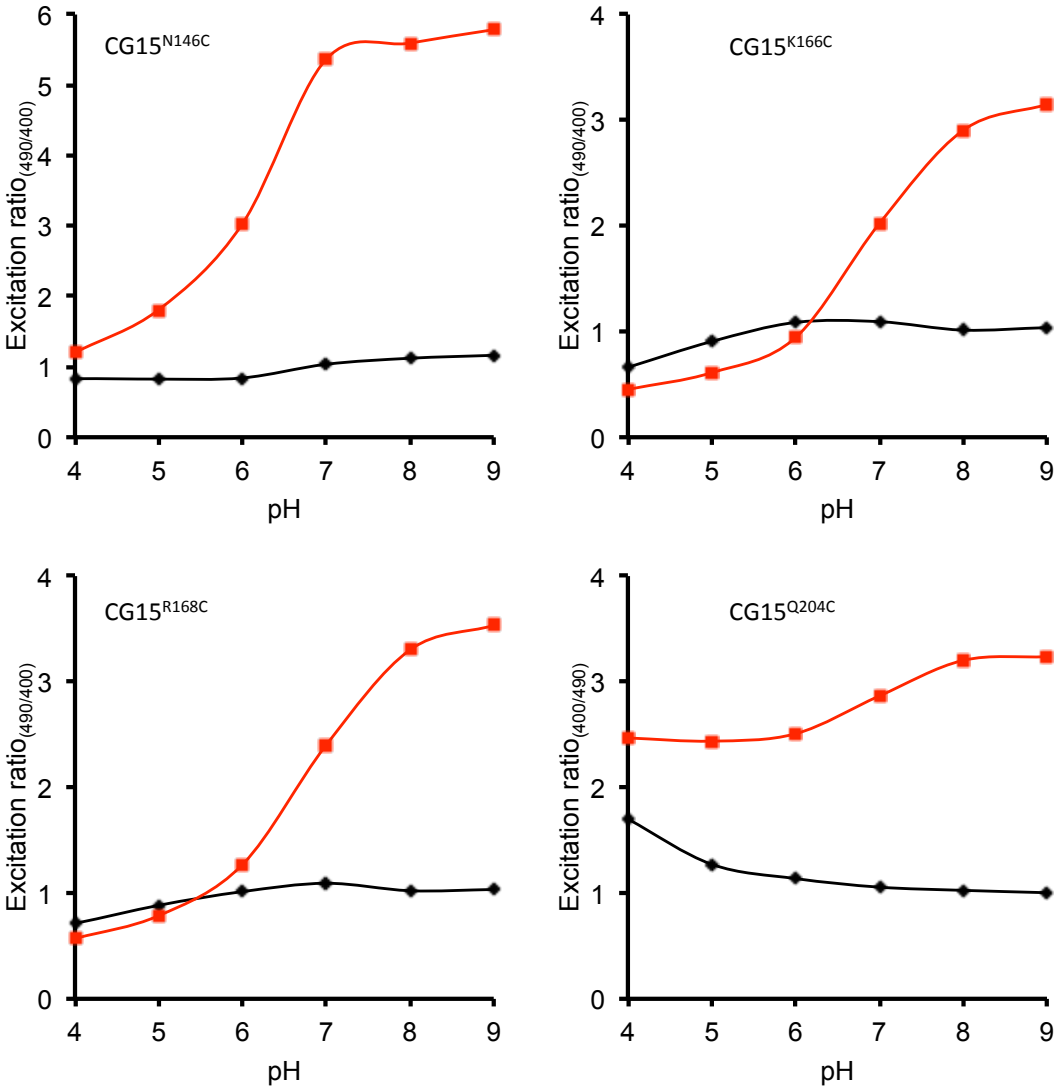


Fig 7.12 Effect of pH on excitation peak ratio. Excitation peak ratios calculated for the $CG15^{CC}$ variants (as indicated in the figure) under reducing (1 mM DTT, black line) or oxidising (0.0034% (v/v) H_2O_2, red lines) conditions at different pHs.

Specifically, our aim is to construct GFP-based sensors for the biologically important area of redox sensing. Oxidative stress and the production of reactive oxygen species (ROS) alter cellular balance and damage cell components, leading to disease, including cancer [171]. However, ROS also plays essential roles in cell processes (e.g. cellular defence, apoptosis and as a secondary messenger in signal transduction). Therefore, maintaining the correct balance between ROS production and consumption is critical to the cell. Drug screening also routinely uses changes in cellular redox state to assess the non-specific effects of early drug candidates.

Traditional approaches for detecting changes in cellular redox conditions are hampered by the lack of reliable non-invasive methods. Redox-sensitive chemical dyes are one of the most widely used products for sensing cellular oxidation state [172]. However, these have several drawbacks: oxidation of the probe is irreversible, the probe's optical response may also be due to other cellular radicals (e.g. NO radical) [172], the dyes may produce further toxic radicals, and they can be sensitive to photo-oxidation [172]. The dyes also readily cross membranes so targeting to specific cellular locations is not possible.

Other procedures for monitoring cell redox state are generally invasive, difficult to implement and have limitations in detection eg. measuring the amounts of the reduced and oxidised forms of natural cell redox mediators such as glutathione [165]. However, cells need to be harvested meaning that single cells cannot be analysed. Another approach is to link a reporter protein such as GFP to a redox regulated promoter system. However there will be transcriptional and translational delay, potential for "noise" due to a leaky promoter or changes to other cellular events and the output is non-reversible as signal is only removed by reporter breakdown. Monitoring the transcriptional activity of genes associated with redox regulation *via* changes in mRNA levels is also used but this again has many of the above problems. Therefore, genetically encoded biosensors based around GFP that can monitor fluctuations were sort through the use of protein engineering to overcome the problems associated with these traditional approaches.

Introduction of two cysteine residues in close proximity to the GFP chromophore attributed redox sensing properties to the proteins, termed roGFPs. Changes in redox state mediated the oxidation or reduction of the cysteine residues resulting in altered spectral properties (Section 7.1). Cysteine residues have been

shown to play an important role in allowing proteins to respond to ROS [159]. The cellular redox state can be attributed in many cases to the ratio of disulphide to free thiol groups of proteins and small molecules such as glutathione. It therefore stands to reason that introduction of cysteine residues into protein scaffolds will elicit redox sensing properties and are a suitable sensing system for monitoring changes in redox state *in vivo*.

The roGFPs have redox midpoint potentials that range from -229 mV to -299 mV with various oxidation and reduction rates [165, 167]. However, the development of redox sensors with redox midpoint potentials outside of this range with quicker response rates to oxidants would be most advantageous for high time resolution in different cellular compartments.

Here, we have described how a basic domain insert scaffold termed CG15 that is redox agent inert (Fig 7.4 and Fig 7.6) can be adapted through the use of molecular modeling and rational protein engineering that responds to changes in redox conditions. The domain insert between EGFP residues N146 and S147 provide the initial important fluorescence properties: two excitation maxima at ~400 nm and ~488 nm, potentially due to mixed populations of CG15 with either a protonated or deprotonated fluorophore respectively. The ratios of the two excitation maxima have been made to be dependent on redox state by the incorporation of cysteine residues at various positions within CG15 (Fig 7.3), capable of forming intramolecular disulphide bonds with C147 native to CG15.

7.3.1 Redox sensing properties of CG15^{CC} variants

The CG15^{Q204C} contains a double cysteine mutation at the same residue in EGFP (147 and 204) as that of a recently developed GFP-based redox sensor roGFP2 [165]. The major difference between the two is the domain insertion between residue N146 and C147. Insertion of the cyt b_{562} domain at this position could potentially alter the pK_a values of the cysteine thiol groups therefore altering the redox sensing properties of CG15^{Q204C} with respect to roGFP2. The absorbance and fluorescence intensity at 490 nm for both CG15^{Q204C} (Fig 7.4 and Fig 7.6) and roGFP2 [165] decreases in the presence of oxidant whist the excitation maxima at ~400 nm increases. This would suggest that in both these cases, the formation of a disulphide bond promotes the formation of the protonated form of the chromophore.

Crystal structure analysis of roGFP2 identified the molecular basis for the oxidant induced altered spectral properties. Under reducing conditions no disulphide bond is formed between C147 and C204 allowing a hydrogen-bonding network to exist between the fluorophore and its surrounding residues (Fig 7.1 a) similar to that of S65T-GFP (Chapter 1, Fig 1.12 b). This hydrogen-bonding network is capable of stabilizing a negative charge on the phenolate group of the fluorophore (Y66), responsible for the excitation maxima at 490 nm [62, 69, 173]. Under oxidizing conditions the disulphide bond formed causes a very minor shift in residue S205 resulting in an extended hydrogen-bonding network (Fig 7.1 b) similar to that of wt GFP (Chapter 1, Fig 1.12 a). The hydroxyl of T65 is linked to the phenolate of Y66 by hydrogen bonds through the side chains of E222, S205 and a conserved water molecule [69]. In this conformation, a proton can be shuffled from the carboxylate of E222 to the phenolate group of Y66 [69]. Protonation of the fluorophore results in an excitation maxima at 400 nm with a decrease in intensity of the 495 nm excitation maxima.

In contrast, $CG15^{N146C}$, $CG15^{K166C}$, and $CG15^{R168C}$ appear to promote the ionized chromophore when forming the local disulphide bond, as suggested by the increase in intensity at 488 nm and decrease at 400 nm in the presence of oxidant (Fig 7.6). This is potentially due to the relative position of the second introduced cysteine residue to C147, native to CG15. In $CG15^{K166C}$ and $CG15^{R168C}$ the mutated residues are in β-strand 8 adjacent to β-strand 7 that contains C147 (Fig 7.13). Formation of a disulphide bond between C147 and either K166C or R166C (in $CG15^{K166C}$or $CG15^{R168C}$) is less likely to cause a shift in the position of residue S205 but could potentially cause a shift in the position of residue H148, preferentially stabilizing a negative charge on the chromophore therefore resulting in an increase in the 488 nm excitation maxima (Fig 7.13). Similar changes in the local environment around the chromophore upon the formation of a disulphide bond between C146 and C147 in $CG15^{N146C}$ could also potentially give a preference to charge stabilization on the fluorophore, given that C146, S147 and H148 are all adjacent to one another on the same β-strand (S7) (Fig 7.13). Therefore, depending on which residue is mutated to partner C147 will influence the protonated state of the chromophore thus the fluorescence properties. These fluorescence properties are in turn influenced by the chemical state of the two cysteine pairs.

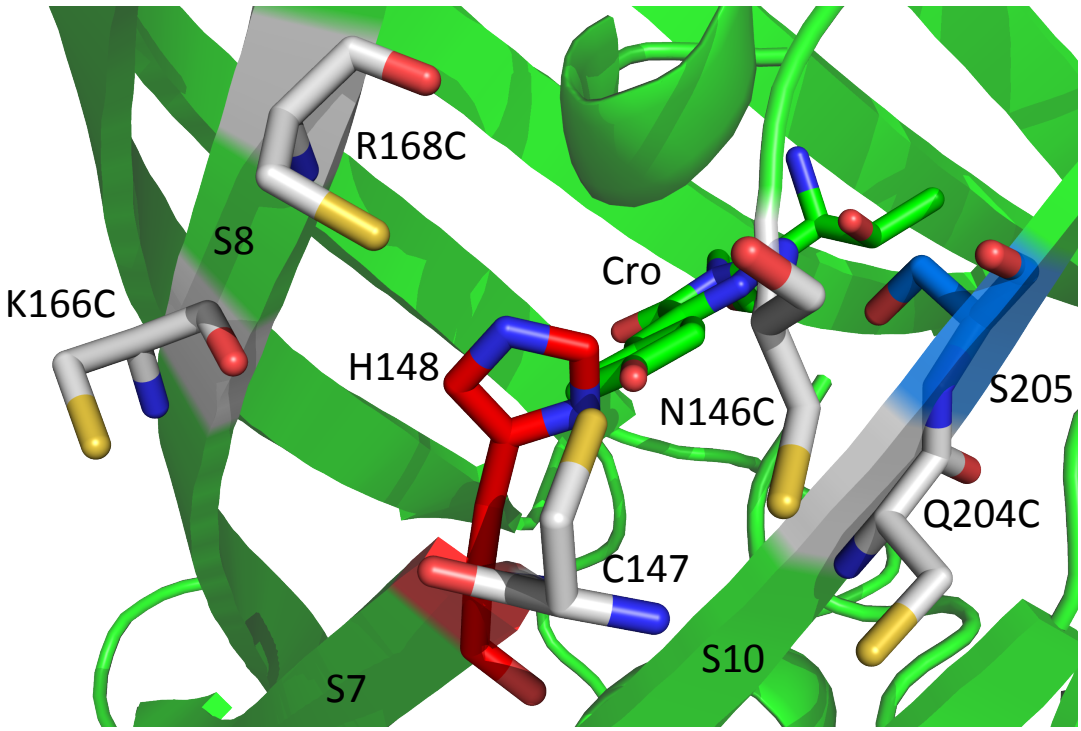


Fig 7.13 Cysteine mutation positions within a model of CG15 tertiary structure. Relative cysteine mutations (CPK sticks) have been introduced into the model structure of CG15 highlighting their positions with respect to C147, native to CG15. Residues H148 (red sticks) and S205 (blue sticks) have been shown as they play an important role in modulating the protonation state of the fluorophore (Cro: green sticks). β-strands 7, 8 and 10 are labeled S7, S8 and S9 respectively. For clarity the cyt b_{562} domain and GlyGlySer linkers have been removed. A full model of CG15 can be seen in Fig 7.3. Formation of disulphides between C147 and either C146, C166 or R168 are likely to have an effect on the position of H148 potentially promoting charge stabilization on the chromophore. Formation of a disulphide between C147 and C204 has an effect on the position of the S205 side chain promoting the protonated form of the chromophore.

Comparison of the UV-visible absorption spectra (Fig 7.4) and fluorescence excitation spectra (Fig 7.6) showed distinct differences in the intensity of the absorption maxima to those of the excitation maxima under reducing conditions for $CG15^{N146C}$, $CG15^{K166C}$ and $CG15^{R168C}$. In the UV-visible absorption spectra the ~400 nm absorption maxima is predominant where as in the excitation spectra the ~490 nm excitation maxima is predominant. The discrepancy between the different spectra indicates that under reducing conditions either the quantum yield (QY) at ~400 nm is decreased, the QY at ~490 nm is increased or both.

Quantum yield determination against a 9-aminoacridine standard (for QY determination at 400 nm) and a fluorescein standard (for QY determination at 490 nm) would be required to confirm this. Modulation of the quantum yield by disulphide formation may be due to insertion of the cyt b_{562} domain affecting the local structure adjacent to the chromophore. External solvent could potentially access the internal environment surrounding the chromophore and quench the chromophore in the excited state thereby reducing the quantum yield.

7.3.2 $CG15^{CC}$ variant redox midpoint potential and redox kinetics.

The redox midpoints determined for the $CG15^{CC}$ variants are the most reducing identified to date for ratiometric fluorescent redox sensors (>300 mV) (Fig 7.9 and Table 7.4). Studies using roGFPs have identified the basal redox potential of the cytoplasm in HeLa cells to be more reducing than previously thought with values between -315 - -325 mV [154, 166], dependent on which roGFP sensor was used. These sensors were also used to identify the redox potential inside mitochondria to be more reducing than the cytoplasm at -360 mV [165]. Given that the redox midpoint potentials for the roGFPs are <300 mV, accurately measuring redox potentials above these values (especially in mitochondria) would be hard to achieve. Here, we have shown that all of the $CG15^{CC}$ redox sensors have redox midpoint potentials >300 mV (Fig 7.9 and Table 7.4) and therefore makes them more suitably placed for measurements of redox potential around these values. This implies that they could potentially be the most sensitive sensors to increases in cellular oxidants providing a much higher resolution for measurements of cellular redox state in more reducing cellular compartments (cytoplasm and mitochondria).

The observation that CG15^{N14C} has fast oxidation and reduction kinetics of 2.48 and 0.59 min^{-1} respectively is probably due to the two cysteine residues (EGFP^{C146} and EGFP^{C147}) being in very close proximity to one another (Fig 7.13). In wt GFP the residues at positions 146 and 147 are adjacent to one another and connected by a peptide bond, forming part of β-strand 7. If these residues were mutated to cysteines, formation of disulphides under oxidising conditions may not elicit the structural change required for modulation of the chromophore protonation state. In the case of CG15^{N146C} β-strand 7 is separated between residues C146 and C147 by the cyt b_{562} domain insertion (Fig 7.14), therefore formation or breaking of a disulphide bond potentially reconnects or separates the β-strand respectively (Fig 7.14). Therefore, insertion of the cyt b_{562} domain at this position is potentially critical for the function of this particular redox sensing GFP, allowing for a new residue to be sampled (N146) for mutation to a cysteine. To assess if the insertion of cyt b_{562} at this position is critical for the results observed here the N146C and S147C mutations could be introduced into EGFP alone and used as a comparison to the CG15^{N146C} variant. The redox kinetics for CG15^{N146C} determined here are the fastest described to date in comparison to the roGFP variants [167].

The redox midpoint potential and redox kinetics for roGFP variants have been shown to be highly dependent on the number of positively charged residues in the immediate vicinity of the disulphide forming cysteines, by modulating their pK_a values [159, 166, 167]. The more positively charged residues in the cysteine local environment confers increased rates of oxidation and reduction [167], and more oxidizing midpoint potentials [166]. It is therefore not surprising that CG15^{K166C} and CG15^{R168C} have decreased oxidation rate constants (0.65 and 0.69 min^{-1} respectively) with respect to CG15^{N146C} (2.48 min^{-1}), as they both have a positively charged residue substituted for a cysteine. Both residues K166 and R168 are in close proximity to C146 in CG15^{N146C} and could be having a positive effect on its redox rates, although these residues do not appear to have conferred a more oxidizing midpoint potential. CG15^{Q204C} has no positively charged residues in the EGFP domain in close proximity to it and could therefore account for the ~two-fold decrease in oxidation and reduction rate constants with respect to CG15^{N146C}.

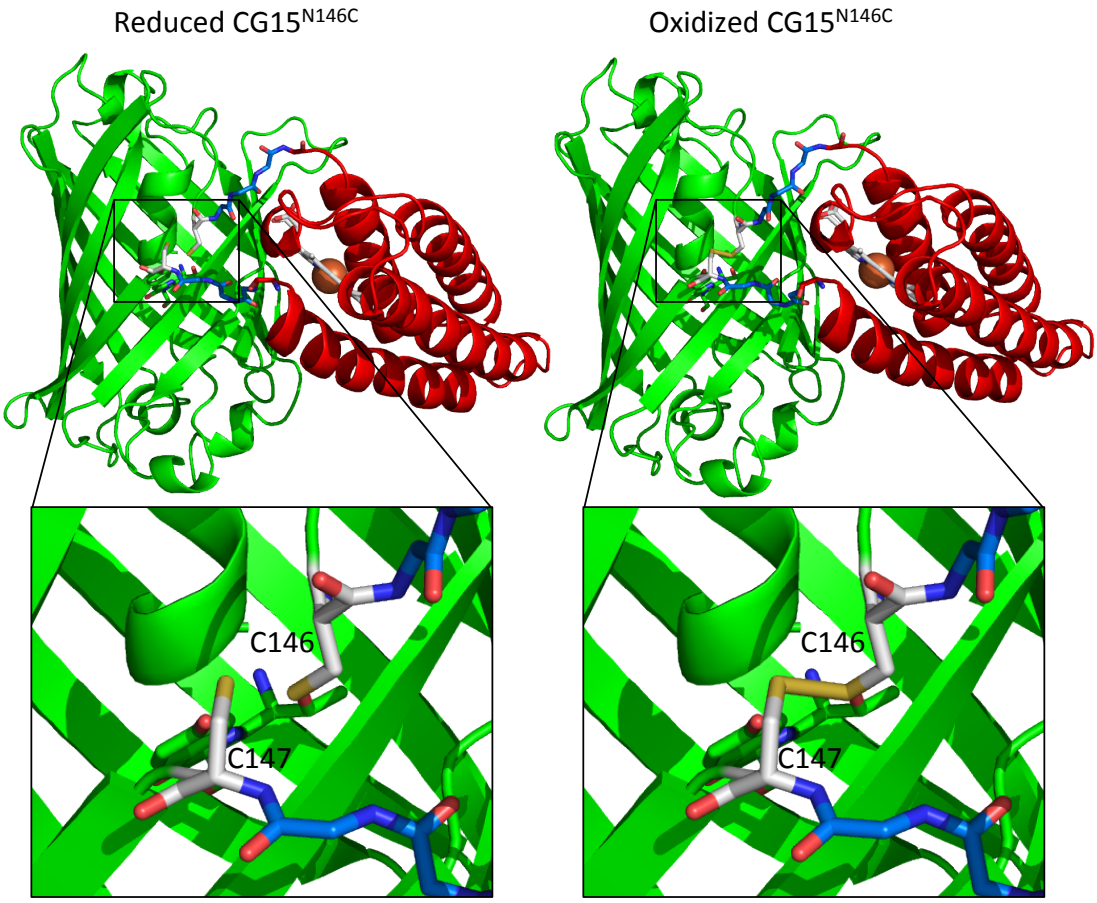


Fig 7.14 Model of reduced and oxidized CG15^{N146C}. Models of reduced CG15^{N146C} (left) and oxidized CG15^{N146C} (right). Formation of a disulphide between C146 and C147 (EGFP numbering) essentially reconstitutes β-strand 7 where as under reducing conditions the β-strand is separated.

It is possible that with further substitution of residues surrounding the cysteines of the CG15^{CC} variants for more positively charged residues the redox rates could be enhanced more than they already have been. It is also possible that the redox midpoint potential for the CG15^{CC} variants could be altered to provide a toolbox of fluorescent redox sensors that sample a wide range of midpoint potentials.

7.3.3 CG15^{CC} variant pH sensitivity

Given that all of the CG15^{CC} variants are sensitive to changes in pH it would be advantageous to mutate these variants further to eradicate their pH sensitivity. As mentioned previously (Section 7.2.6) it is known that the S65T mutation of EGFP, and other green fluorescent proteins, confers pH sensitivity [170]. Therefore to potentially reduce the pH sensitive nature of the CG15^{CC} variants the T65 residue of the fluorophore could be mutated back to a serine. However, in doing so the excitation ratios, dynamic range and redox midpoint potential could be altered.

The cyt b_{562} domain insertion may also be playing a role in the pH sensitivity of the chromophore by disrupting the local structure adjacent to the chromophore. Disrupting the structure of the β-strands in this region of EGFP may allow solvent to access the interior environment surrounding the chromophore influencing its protonation state. To determine if it is the S65T mutation, the cyt b_{562} domain insertion or both that is responsible for the pH sensitivity would require mutating the T65 residue back to a serine or by inserting the cyt b_{562} domain into wt GFP and assessing the pH sensitivity.

Given that there are currently no crystal structures for the CG15^{CC} variants makes it difficult to identify the exact molecular mechanism behind their redox sensing properties and pH sensitivity. Once crystal structures have been determined and the precise EGFP and cyt b_{562} domain arrangement is known it will be easier to further rationally design the sensors to potentially improve their function and eradicate their pH sensitivity.

Although all the variants exhibited pH sensitivity, the excitation peak ratios of CG15^{N146C} were stable above a pH=7.0 therefore could still be an adequate *in vivo* redox sensor, given that generally physiological pH is >7. The S65T mutation is also in roGFP2, which has been shown to be successfully used *in vivo* to measure redox state in different cellular compartments for both mammalian [154, 165] and plant cells

[170].Therefore the fact that the $CG15^{CC}$ variant spectral properties are sensitive below a pH of ~7 may not eliminate them as suitable *in vivo* redox sensors.

7.3.4 Conclusion

In conclusion, we have demonstrated here that using a computer aided rational design approach a directly evolved protein scaffold, CG15, has been further engineered for ratiometric fluorescent redox sensing. In particular variant $CG15^{N146C}$ exhibited the most reducing redox midpoint potential and the fastest response rates to H_2O_2 developed to date, with the potential for further enhancement with additional mutations. In particular, one benefit would be to mutate the residues that coordinate haem (M7 and H105 of the cyt b_{562} domain) so as to eradicate any haem mediated fluorescence quenching. Despite the pH sensitivity of $CG15^{N146C}$ below a pH=7.0 it would still potentially make a useable *in vivo* cellular sensor for redox state.

Chapter 8: General discussion and conclusion

8.1 The strength and weakness of the MuDel transposon directed evolution approach

Directed evolution has many benefits over rational design approaches towards protein engineering. A detailed knowledge of protein structure and function is not required as many different mutational events can be sampled in a single reaction and variants with desired properties selected for [38]. Also mutations can be sampled that may not be intuitively selected during rational design.

The transposon based directed evolution approach described in this Thesis (Chapter 3) has many strengths over other available directed evolution approaches as well as rational design approaches. Firstly, the MuDel system is capable of sampling a diverse range of insertion sites throughout a target gene, pivotal to directed evolution methods (Fig 3.6). Other directed evolution approaches such as introducing random breaks in target DNA by either enzymatic or chemical cleavage have been shown to result in libraries with site-specific bias [38, 48]. It has been demonstrated here that other transposon base techniques such as the Tn5 system (Fig 3.6) and in other studies with the Tn7 tansposon system [174] that libraries are produced with target site preference, reducing their usefulness for producing libraries with extensive diversity.

After creation of a transposon insertion library most transposons can be removed by restriction digestion to introduce single random breaks into the target gene. Whilst with the majority of transposon systems the bulk of the transposon is removed from the library, the transposase recognition elements (TREs) and target site duplications (introduced due to the mechanism of transposition) are left behind (Fig 1.8) [47]. This can be useful for probing sites within proteins that are tolerant to peptide insertions but there is no control over the coding sequence inserted [47].

As demonstrated here the MuDel transposon system is not limited by this problem. Removal of MuDel from a DNA library after transposition, by *Mly*I restriction digestion, removes the entire transposon including its TREs, the target site duplication and a triplet nucleotide from the target gene (Fig 3.1 a). This alone can be used to sample single amino acid deletion mutations throughout a target protein (Section 3.2.6). The random break introduced by MuDel removal is also well defined unlike with other directed evolution approaches for introducing random breaks into

target genes (DNAse I or Ce(IV)-EDTA), which can result in multiple breaks per molecule of DNA, nested deletions and tandem duplications [32].

Given that controlled blunt ended breaks can be introduced randomly throughout a target gene provides the strength of the MuDel system, allowing full control of downstream mutagenesis techniques for sampling single amino acid deletions (Chapter 3 and Chapter 4), domain insertions (Chapter 3 and Chapter 5) or for the incorporation of non-natural amino acids to increase the chemical diversity of proteins (Chapter 1, Fig 1.9) [50].

Owing to the nature of transposition and subsequent removal by *Mly*I restriction digestion the blunt ended breaks produced randomly throughout the target gene are not always in frame (Table 3.1 and Section 3.2.7). However, DNA cassettes encoding whole protein domains for insertion into the random breaks can be designed with additional random nucleotides at their 5' and 3' ends allowing all three reading frames to be sampled within the library, as demonstrated here (Table 3.2). This also gives the freedom to design and encode the linking peptides to separate the two domains in the resulting integral fusion scaffold. Although in the past limited importance has been placed on the linking peptides separating two domains of an integral fusion scaffold [34] it has been demonstrated here (Chapter 6) that they play a critical role in the magnitude of functional coupling achieved (discussed in more detail in Section 8.3).

Taking all these benefits into account validates the strength of the MuDel system for creating diverse DNA libraries that provide the freedom and control to create sub-libraries sampling different mutational events that target the polypeptide backbone and residue side chains. The only weakness of the MuDel transposon method is that the initial library generation can be time consuming, however, this is far outweighed by the benefits of the technique.

Creation of a MuDel insertion library in the target gene, *egfp*, has been described in detail in Chapter 3 and was used to sample single amino acid deletions (Section 3.2.6) and cyt b_{562} domain insertions throughout EGFP (Section 3.2.7). Between the two libraries 139 unique sites within EGFP were sampled for mutagenesis. However, with improved high throughput screening techniques it is possible that even more sites could have been identified for mutational sampling, allowing the full potential of the MuDel insertion library to be exploited.

8.2 Protein engineering through single amino acid deletion mutagenesis

Although dogma suggests that amino acid deletions are detrimental to protein structure and therefore function (Chapter 1, section 1.2.2), causing registry shifts in organized secondary structure (Chapter 1, Fig 1.1), it has been demonstrated here (Chapter 4) that they can also be beneficial. Until now there has been a lack of directed evolution techniques capable of introducing single amino acid deletion mutations randomly throughout a target protein [39]. For this reason there has been relatively little investigation into the value of deletion mutations and their effects on protein structure and function.

Whilst deletion mutations have been studied in EGFP prior to this investigation they were rationally localized to particular loops or towards the N- or C-termini of GFP [122, 123, 125]. However, for the majority of cases more than a single amino acid was deleted at a time, which was shown to be detrimental to GFP fluorescence [122, 123, 125]. Using the MuDel transposon based directed evolution approach it has been possible to sample single amino acid deletion mutations throughout EGFP (Fig 4.4 and Fig 4.5). Screening of the EGFPΔ library identified novel EGFP variants with increased protein fluorescence conferred to cell cultures with respect to EGFP (Fig 4.6). In particular deletion of residue G4, D190 or A227 from EGFP resulted in a marked increase in fluorescence observed in cell cultures (Fig 4.8). Further analysis identified the effects of the mutations were not due to altered fluorescence properties of EGFP (Table 4.5) but were elicited through increased stability (Fig 4.13 and Table 4.8) or improved folding (Fig 4.14 and Table 4.8).

Of the three variants studied in detail the G4Δ mutation in particular was identified as being the most beneficial of the deletion mutations, dramatically increasing protein production levels (Fig 4.9 a) increasing protein stability (Fig 4.6 and Table 4.8), improving folding (Fig 4.7) and increasing the fluorescence observed in cell cultures more than any of the other mutations (Fig 4.8). The effects of the G4Δ mutation was also demonstrated to be transferrable to other variants identified from the library increasing their cellular fluorescence (Fig 4.9 b). The effects of the G4Δ mutation were also not just specific to EGFP with the incorporation of this mutation into EYFP also resulting in increased cellular fluorescence (Fig 4.9 b). These observations indicate that the G4Δ mutation is having a generic effect on fluorescent

proteins and should potentially be incorporated into all fluorescent proteins to improve their folding and stability.

The most interesting observation was that the positions of these deletion mutations were unintuitive. In particular deletion of G4 or A227 which are positioned in the first α-helix or at the C-terminal side of the last β-strand in EGFP respectively (Fig 4.7). If using a rational design approach these residues would not necessarily have been selected with the intention of improving stability and folding of EGFP. This highlights the importance and power of the MuDel directed evolution approach for identifying mutational events that would not otherwise be obvious for influencing the function of the target protein.

As well as the MuDel directed evolution approach being important for identifying novel variants it has also allowed a survey of tolerated and non-tolerated single amino acid deletions throughout EGFP (Table 4.3,Table 4.4, Fig 4.4 and Fig 4.5) to help formulate very preliminary rules (Section 4.3.2) to help guide rational design attempts utilizing this mutational technique. With further production and analysis of single amino acid deletion libraries within other target proteins it will potentially be possible to build upon these rules, provide more information on the structural impacts of deletion mutations and increase the usefulness of this mutational technique as a protein engineering tool.

8.3 Creation of artificial biomolecular switches through domain insertion.

In this Thesis we have shown how domain insertion provides a general approach to functionally link normally disparate proteins to act as aritificial biomolecular switches (Chapter 5 and Chapter 6). Whilst there has been limited success in constructing artificial biomolecular switches using rational design strategies for domain insertion [34, 35] the MuDel directed evolution approach described here and discussed previously (Section 8.1) provides many advantages over rational design approaches.

Using the MuDel directed evolution approach it has been possible to identify sites within EGFP that are not only tolerant to cyt b_{562} domain insertion (Fig 5.3 and Fig 5.4) but also display functional coupling of the two domains (Fig 5.9). Previous work constructing simple head-to-tail fusions between cyt b_{562} and

EGFP exhibited only moderate functional coupling (~65%) [82], agreeing with observations for the N-and C-terminal fusions, CG1 and CG12, generated here (Fig 5.9). It was observed that cyt b_{562} domain insertion position was crucial to the extent of functional coupling observed between the different variants (Fig 5.9) and that in one particular variant, CG6, the differential linker length separating the two domains was key to the maximal functional coupling observed (up to 100%) (Fig 5.9 and section 6.2.2.2).

Crystal structure determination of the cyt b_{562}-EGFP integral domain fusion scaffold, CG6, has helped to explain the molecular basis for the functional coupling of the two domains and some of the observed redox-dependent switching properties (Fig 5.11 a and section 6.2.2). The two domains in the integral domain fusion scaffold have a side-by-side domain arrangement (Fig 6.7), which is facilitated by the differential linker length connecting the two domains providing a constrained molecular pivot point (Fig 6.10). This resulted in a domain-domain interface being formed, further stabilizing the side-by-side domain arrangement required for the integral domain fusion scaffold function (Fig 6.12).

Retrospective structural analysis has provided information on how the extent of functional coupling can be maximized through domain arrangement and has highlighted the importance of the linking regions connecting two domains in an integral fusion scaffold. This will help towards future efforts in designing artificial biomolecular switches that exhibit intimate functional coupling of two domains.

8.4 Rational design of a directly evolved integral domain fusion scaffold, CG15.

Whilst it has been highlighted throughout this Thesis the importance of a directed evolution approach for identifying sites within a target protein that will tolerate the insertion of another protein domain, that retains and couples the functions of the two domains (Chapter 3 and Chapter 5), this does not negate the importance of a rational design approach for the further development of these scaffolds. As has been mentioned previously rational design generally relies on the knowledge of the structure and function of the target protein in order to

make educated decisions on which residues if mutated may elicit a desired outcome [175]. Given the lack of a crystal structure for the cyt b_{562}-EGFP integral domain fusion protein, CG15, a computer-aided approach was used to guide the rational design process.

By creating a model structure for CG15 (Section 7.2.1) it was possible to select sites for the introduction of cysteine residues to create ratiometric fluorescent redox sensors (Fig 7.3). The use of cysteine mutations has been demonstrated before to attribute redox sensing properties to GFP [154, 165-167]. However, through cyt b_{562} domain insertion between residues N146 and S147 in EGFP it has been possible to sample the mutation N146C that would potentially not be effective in the standard roGFP constructs (discussed in more detail in section 7.3.2).

The $CG15^{CC}$ variants exhibited the most reducing redox midpoint potentials (Fig 7.9 and Table 7.4) reported for any protein based redox sensor to date [154, 165-167] with one variant ($CG15^{N146C}$) exhibiting the fastest redox kinetics observed to date [167]. Given that in previous studies the redox midpoint potentials and redox kinetics could be altered through the introduction of positively charged residues about the site of the introduced cysteine residues or by introducing geometric constraint around the disulphide forming residues [166, 167] it is possible that the $CG15^{CC}$ variants could be further optimised.

8.5 Protein engineering through polypeptide backbone mutagenesis

In this Thesis it has been shown that mutational events that target the polypeptide backbone can be beneficial and not just detrimental as dogma suggests. Here we have demonstrated that shortening the polypeptide backbone by single amino acid deletion mutagenesis can provide novel protein variants with improved stability and folding, sampling conformational space not accessible through substitution mutations alone. By performing a thorough survey of tolerated and non-tolerated single amino acid deletion positions throughout EGFP it has been possible to start formulating a set of rules to help guide future rational design utilizing this mutagenesis technique.

It has also be demonstrated how disrupting the continuity of the polypeptide backbone by whole domain insertion can lead to the identification of novel integral domain fusion scaffolds for the use as artificial biomolecular sensors. Structural analysis has increased our understanding of the molecular mechanism behind the functional coupling of two domains in an integral fusion scaffold and again will aid in the future design of protein scaffolds with more specific tailored properties.

Whilst both of these mutational techniques were accomplished using a directed evolution approach, it has been demonstrated how a computer aided rational design process can be used in conjunction with directed evolution to further develop novel protein variants with desired properties.

8.6 Future work

8.6.1 Determination of EGFPΔ structures by X-Ray crystallography

Whilst the X-ray crystal structure for EGFP has been determined in this study and helps towards understanding the potential impacts a single amino acid deletion mutation could potentially have on its structure and there function, the exact molecular basis for the effects of the mutations can not be identified. Crystal structure determination for the EGFP$^{D190\Delta}$ and EGFP$^{A227\Delta}$ is currently being performed and will help to explain the molecular effects of these particular single amino acid deletions on EGFP structure and function. Of most interest however is how the G4Δ mutation increases EGFP stability and improves folding, therefore understanding the molecular mechanisms by which it elicits its effects would be most advantageous. Whilst attempts to crystalize EGFP$^{G4\Delta}$ have to date failed, continued attempts will hopefully yield protein crystals with which the structure can be determined. Apart from the variants mentioned here it would also be interesting to investigate the molecular effects of single amino acid deletions that reduce EGFP fluorescence. This is important as it will help identify the molecular mechanisms of single amino acid deletions that are detrimental to protein structure and function furthering our understanding of polypeptide backbone mutations and how they can be used usefully in protein engineering by rational design.

8.6.2 Further rational design of the CG6 integral domain fusion scaffold.

The CG6 scaffold has already been shown to act as a haem sensor (Fig 5.9), potential sensor for redox state (Fig 5.9 and Fig 5.10) and as a H_2O_2 specific oxidant sensor (Fig 5.11). However, it also has the potential to be used in bionanoelectronics as a photovoltaic component. Cyt b_{562} is already being considered for its potential use in molecular electronics exploiting its redox properties for electrochemical gating providing a route to the modulation of current flow through the protein [85, 86]. This has been achieved by exploiting thiol gold chemistry allowing cyt b_{562} to be chemically attached to gold surfaces for current transfer [86]. Work has also been performed on the light induced electron transfer from the terminal fusion scaffold between cytb_{562}-EGFP and a gold surface, showing that wavelength specific light (488 nm) induced electron transfer from the EGFP domain through the cyt b_{562} domain to the gold surface [149, 176]. However, as has been shown in this thesis (Fig 5.9) and previous work [82] the energy transfer efficiency between the EGFP domain and cyt b_{562} domain of terminal fusion scaffolds is only 65%, whilst that of CG6 is up to 100% (Fg 5.9), akin to the energy transfer efficiencies of natural light harvesting systems. Introduction of cysteine residues into the cyt b_{562} domain of CG6 will allow its interaction with gold surfaces and for its potential use as a high efficiency photovoltaic component. CG6 may therefore play an important role in the future development of novel protein based molecular components for use in solar panels.

It was observed in the crystal structure for CG6 that the three molecules in the asymmetric unit had the cyt b_{562} domain at slightly different angles to the EGFP domain indicating possible flexibility at the hinge region. Whilst SAXS analysis of CG6 in solution supported the side-by-side domain arrangement observed in the crystal structure it would be interesting to investigate the potential flexibility between the two domains. NMR solution structure determination could highlight potential regions with increased dynamics in the CG6 scaffold and could potentially direct rational design of the linkers forming the hinge region to reduce the dynamics between the two domains. With the determined X-ray crystal structure and the potential insights that could be

provided by NMR studies could also help direct the further mutation of the domain-domain interface to further stabilize the side-by-side domain arrangement and potentially increase the stability of the entire scaffold.

The CG6 scaffold developed here utilized the widely used EGFP. However, EGFP is neither the brightest fluorescent protein nor the most stable. It would therefore be interesting to create a new integral domain fusion scaffold, using CG6 as a template, between cyt b_{562} and superfolder GFP (sfGFP) [71]. Both EGFP and sfGFP have very similar spectral properties and structures however sfGFP has superior stability and brightness. An integral fusion scaffold between cyt b_{562} and sfGFP could further improve the ability of the scaffold to act as a photovoltaic component.

8.6.3 Determination of CG15^{CC} variants structure by X-ray crystallography.

Whilst a molecular model for CG15 was developed to aid the design process of the double cysteine mutants of CG15, it does not shed any light on the mechanism of ratiometric fluorescent redox sensing. Although changes in the excitation maxima in the presence of an oxidant or reductant is indirect evidence for the formation and breaking of disulphide bonds in the vicinity of the chromophore there is currently no structural evidence confirming this. Therefore structure determination of the CG15^{CC} variants under reducing and oxidizing conditions would confirm the formation of disulphide bonds between the two introduced cysteine residues and would also provide an explanation for the changes in the spectral characteristics observed at the molecular level.

The crystal structures would also help towards understanding what affect the cyt b_{562} domain insertion plays in the redox sensing mechanism. Having crystal structures for the CG15^{CC} variants would also help towards identifying potential sites for rational design to further improve the redox kinetics or to alter their redox midpoint potentials. This can be achieved by increasing the reactivity of the cysteine residues by modulating the thiol group pKa values with the introduction of basic substitution mutations in their close vicinity.

References

1 Fersht, A. (1999) Structure and mechanism in protein science : a guide to enzyme catalysis and protein folding. W.H. Freeman, New York

2 Gaston, M. A., Zhang, L., Green-Church, K. B. and Krzycki, J. A. (2011) The complete biosynthesis of the genetically encoded amino acid pyrrolysine from lysine. Nature. **471**, 647-650

3 Stadtman, T. C. (1996) Selenocysteine. Annu Rev Biochem. **65**, 83-100

4 Hutchison, C. A., 3rd, Phillips, S., Edgell, M. H., Gillam, S., Jahnke, P. and Smith, M. (1978) Mutagenesis at a specific position in a DNA sequence. J Biol Chem. **253**, 6551-6560

5 Brannigan, J. A. and Wilkinson, A. J. (2002) Protein engineering 20 years on. Nat Rev Mol Cell Biol. **3**, 964-970

6 Sigal, I. S., Harwood, B. G. and Arentzen, R. (1982) Thiol-beta-lactamase: replacement of the active-site serine of RTEM beta-lactamase by a cysteine residue. Proc Natl Acad Sci U S A. **79**, 7157-7160

7 Oliphant, A. R. and Struhl, K. (1989) An efficient method for generating proteins with altered enzymatic properties: application to beta-lactamase. Proc Natl Acad Sci U S A. **86**, 9094-9098

8 Schneider, K. D., Bethel, C. R., Distler, A. M., Hujer, A. M., Bonomo, R. A. and Leonard, D. A. (2009) Mutation of the active site carboxy-lysine (K70) of OXA-1 beta-lactamase results in a deacylation-deficient enzyme. Biochemistry. **48**, 6136-6145

9 Petrosino, J. F. and Palzkill, T. (1996) Systematic mutagenesis of the active site omega loop of TEM-1 beta-lactamase. J Bacteriol. **178**, 1821-1828

10 Knox, J. R. (1995) Extended-spectrum and inhibitor-resistant TEM-type beta-lactamases: mutations, specificity, and three-dimensional structure. Antimicrob Agents Chemother. **39**, 2593-2601

11 Leung, Y. C., Robinson, C. V., Aplin, R. T. and Waley, S. G. (1994) Site-directed mutagenesis of beta-lactamase I: role of Glu-166. Biochem J. **299 (Pt 3)**, 671-678

12 Brown, N. G., Shanker, S., Prasad, B. V. and Palzkill, T. (2009) Structural and biochemical evidence that a TEM-1 beta-lactamase N170G active site mutant acts via substrate-assisted catalysis. J Biol Chem. **284**, 33703-33712

13 Giron, M. D. and Salto, R. (2011) From green to blue: site-directed mutagenesis of the green fluorescent protein to teach protein structure-function relationships. Biochem Mol Biol Educ. **39**, 309-315

14 Sawano, A. and Miyawaki, A. (2000) Directed evolution of green fluorescent protein by a new versatile PCR strategy for site-directed and semi-random mutagenesis. Nucleic Acids Res. **28**, E78

15 Pascarella, S. and Argos, P. (1992) Analysis of insertions/deletions in protein structures. J Mol Biol. **224**, 461-471

16 Taylor, M. S., Ponting, C. P. and Copley, R. R. (2004) Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. Genome Res. **14**, 555-566

17 de Jong, W. W. and Ryden, L. (1981) Causes of more frequent deletions than insertions in mutations and protein evolution. Nature. **290**, 157-159

18 de Wildt, R. M., van Venrooij, W. J., Winter, G., Hoet, R. M. and Tomlinson, I. M. (1999) Somatic insertions and deletions shape the human antibody repertoire. J Mol Biol. **294**, 701-710
19 Simm, A. M., Baldwin, A. J., Busse, K. and Jones, D. D. (2007) Investigating protein structural plasticity by surveying the consequence of an amino acid deletion from TEM-1 beta-lactamase. FEBS Lett. **581**, 3904-3908
20 Wood, N., Bhattacharya, T., Keele, B. F., Giorgi, E., Liu, M., Gaschen, B., Daniels, M., Ferrari, G., Haynes, B. F., McMichael, A., Shaw, G. M., Hahn, B. H., Korber, B. and Seoighe, C. (2009) HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. PLoS Pathog. **5**, e1000414
21 Jones, D. D. (2005) Triplet nucleotide removal at random positions in a target gene: the tolerance of TEM-1 beta-lactamase to an amino acid deletion. Nucleic Acids Res. **33**, e80
22 Russell, R. B. (1994) Domain insertion. Protein Eng. **7**, 1407-1410
23 Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S. K., Teichmann, S. A. and Weiner, J., 3rd. (2005) The evolution of domain arrangements in proteins and interaction networks. Cell Mol Life Sci. **62**, 435-445
24 Aroul-Selvam, R., Hubbard, T. and Sasidharan, R. (2004) Domain insertions in protein structures. J Mol Biol. **338**, 633-641
25 Selvam, R. A. and Sasidharan, R. (2004) DomIns: a web resource for domain insertions in known protein structures. Nucleic Acids Res. **32**, D193-195
26 Xu, Z. and Sigler, P. B. (1998) GroEL/GroES: structure and function of a two-stroke folding machine. J Struct Biol. **124**, 129-141
27 Sigler, P. B., Xu, Z., Rye, H. S., Burston, S. G., Fenton, W. A. and Horwich, A. L. (1998) Structure and function in GroEL-mediated protein folding. Annu Rev Biochem. **67**, 581-608
28 Ostermeier, M. (2005) Engineering allosteric protein switches by domain insertion. Protein Eng Des Sel. **18**, 359-364
29 Doi, N. and Yanagawa, H. (1999) Design of generic biosensors based on green fluorescent proteins with allosteric sites by directed evolution. FEBS Lett. **453**, 305-307
30 VanEngelenburg, S. B. and Palmer, A. E. (2008) Fluorescent biosensors of protein function. Curr Opin Chem Biol. **12**, 60-65
31 Edwards, W. R., Busse, K., Allemann, R. K. and Jones, D. D. (2008) Linking the functions of unrelated proteins using a novel directed evolution domain insertion method. Nucleic Acids Res. **36**, e78
32 Guntas, G. and Ostermeier, M. (2004) Creation of an allosteric enzyme by domain insertion. J Mol Biol. **336**, 263-273
33 Takeda, S., Kamiya, N., Arai, R. and Nagamune, T. (2001) Design of an artificial light-harvesting unit by protein engineering: cytochrome b_{562}-green fluorescent Protein chimera. Biochem Biophys Res Commun. **289**, 299-304
34 Baird, G. S., Zacharias, D. A. and Tsien, R. Y. (1999) Circular permutation and receptor insertion within green fluorescent proteins. Proc Natl Acad Sci U S A. **96**, 11241-11246
35 Collinet, B., Herve, M., Pecorari, F., Minard, P., Eder, O. and Desmadril, M. (2000) Functionally accepted insertions of proteins within protein domains. J Biol Chem. **275**, 17428-17433

36 Tucker, C. L. and Fields, S. (2001) A yeast sensor of ligand binding. Nat Biotechnol. **19**, 1042-1046
37 Edwards, W. R., Williams, A. J., Morris, J. L., Baldwin, A. J., Allemann, R. K. and Jones, D. D. (2010) Regulation of beta-lactamase activity by remote binding of heme: functional coupling of unrelated proteins through domain insertion. Biochemistry. **49**, 6541-6549
38 Lutz, S. and Patrick, W. M. (2004) Novel methods for directed evolution of enzymes: quality, not quantity. Curr Opin Biotechnol. **15**, 291-297
39 Neylon, C. (2004) Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. Nucleic Acids Res. **32**, 1448-1459
40 Mizuuchi, K. (1992) Transpositional recombination: mechanistic insights from studies of mu and other elements. Annu Rev Biochem. **61**, 1011-1051
41 Merkulov, G. V. and Boeke, J. D. (1998) Libraries of green fluorescent protein fusions generated by transposition *in vitro*. Gene. **222**, 213-222
42 Sheridan, D. L. and Hughes, T. E. (2004) A faster way to make GFP-based biosensors: two new transposons for creating multicolored libraries of fluorescent fusion proteins. BMC Biotechnol. **4**, 17
43 Savilahti, H., Rice, P. A. and Mizuuchi, K. (1995) The phage Mu transpososome core: DNA requirements for assembly and function. Embo J. **14**, 4893-4903
44 Mizuuchi, M., Rice, P. A., Wardle, S. J., Haniford, D. B. and Mizuuchi, K. (2007) Control of transposase activity within a transpososome by the configuration of the flanking DNA segment of the transposon. Proc Natl Acad Sci U S A. **104**, 14622-14627
45 Craig, N. L. (1991) Tn7: a target site-specific transposon. Mol Microbiol. **5**, 2569-2573
46 Davies, D. R., Goryshin, I. Y., Reznikoff, W. S. and Rayment, I. (2000) Three-dimensional structure of the Tn5 synaptic complex transposition intermediate. Science. **289**, 77-85
47 Hayes, F. and Hallet, B. (2000) Pentapeptide scanning mutagenesis: encouraging old proteins to execute unusual tricks. Trends Microbiol. **8**, 571-577
48 Murakami, H., Hohsaka, T. and Sisido, M. (2002) Random insertion and deletion of arbitrary number of bases for codon-based random mutation of DNAs. Nat Biotechnol. **20**, 76-81
49 Baldwin, A. J., Busse, K., Simm, A. M. and Jones, D. D. (2008) Expanded molecular diversity generation during directed evolution by trinucleotide exchange (TriNEx). Nucleic Acids Res. **36**, e77
50 Baldwin, A. J., Arpino, J. A., Edwards, W. R., Tippmann, E. M. and Jones, D. D. (2009) Expanded chemical diversity sampling through whole protein evolution. Mol Biosyst. **5**, 764-766
51 Haapa, S., Taira, S., Heikkinen, E. and Savilahti, H. (1999) An efficient and accurate integration of mini-Mu transposons *in vitro*: a general methodology for functional genetic analysis and molecular biology applications. Nucleic Acids Res. **27**, 2777-2784
52 Seitz, T., Berger, B., Nguyen, V. T., Tricot, C., Villeret, V., Schmid, S., Stalon, V. and Haas, D. (2000) Linker insertion mutagenesis based on IS21 transposition:

isolation of an AMP-insensitive variant of catabolic ornithine carbamoyltransferase from *Pseudomonas aeruginosa*. Protein Eng. **13**, 329-337
53 Hallet, B., Sherratt, D. J. and Hayes, F. (1997) Pentapeptide scanning mutagenesis: random insertion of a variable five amino acid cassette in a target protein. Nucleic Acids Res. **25**, 1866-1867
54 Biery, M. C., Stewart, F. J., Stellwagen, A. E., Raleigh, E. A. and Craig, N. L. (2000) A simple *in vitro* Tn7-based transposition system with low target site selectivity for genome and gene analysis. Nucleic Acids Res. **28**, 1067-1077
55 Taira, S., Tuimala, J., Roine, E., Nurmiaho-Lassila, E. L., Savilahti, H. and Romantschuk, M. (1999) Mutational analysis of the *Pseudomonas syringae* pv. tomato hrpA gene encoding Hrp pilus subunit. Mol Microbiol. **34**, 737-744
56 Ehrmann, M., Bolek, P., Mondigler, M., Boyd, D. and Lange, R. (1997) TnTIN and TnTAP: mini-transposons for site-specific proteolysis in vivo. Proc Natl Acad Sci U S A. **94**, 13111-13115
57 Manoil, C. and Bailey, J. (1997) A simple screen for permissive sites in proteins: analysis of *Escherichia coli* lac permease. J Mol Biol. **267**, 250-263
58 Hoekstra, M. F., Burbee, D., Singer, J., Mull, E., Chiao, E. and Heffron, F. (1991) A Tn3 derivative that can be used to make short in-frame insertions within genes. Proc Natl Acad Sci U S A. **88**, 5457-5461
59 Ross-Macdonald, P., Sheehan, A., Roeder, G. S. and Snyder, M. (1997) A multipurpose transposon system for analyzing protein production, localization, and function in *Saccharomyces cerevisiae*. Proc Natl Acad Sci U S A. **94**, 190-195
60 Ehrhardt, D. (2003) GFP technology for live cell imaging. Curr Opin Plant Biol. **6**, 622-628
61 Hoffman, R. M. (2005) The multiple uses of fluorescent proteins to visualize cancer *in vivo*. Nat Rev Cancer. **5**, 796-806
62 Tsien, R. Y. (1998) The green fluorescent protein. Annu Rev Biochem. **67**, 509-544
63 Zhang, J., Campbell, R. E., Ting, A. Y. and Tsien, R. Y. (2002) Creating new fluorescent probes for cell biology. Nat Rev Mol Cell Biol. **3**, 906-918
64 Barondeau, D. P., Tainer, J. A. and Getzoff, E. D. (2006) Structural evidence for an enolate intermediate in GFP fluorophore biosynthesis. J Am Chem Soc. **128**, 3166-3168
65 Zhang, L., Patel, H. N., Lappe, J. W. and Wachter, R. M. (2006) Reaction progress of chromophore biogenesis in green fluorescent protein. J Am Chem Soc. **128**, 4766-4772
66 Sniegowski, J. A., Lappe, J. W., Patel, H. N., Huffman, H. A. and Wachter, R. M. (2005) Base catalysis of chromophore formation in Arg96 and Glu222 variants of green fluorescent protein. J Biol Chem. **280**, 26248-26255
67 Barondeau, D. P., Putnam, C. D., Kassmann, C. J., Tainer, J. A. and Getzoff, E. D. (2003) Mechanism and energetics of green fluorescent protein chromophore synthesis revealed by trapped intermediate structures. Proc Natl Acad Sci U S A. **100**, 12111-12116
68 Jackson, S. E., Craggs, T. D. and Huang, J. R. (2006) Understanding the folding of GFP using biophysical techniques. Expert Rev Proteomics. **3**, 545-559
69 Brejc, K., Sixma, T. K., Kitts, P. A., Kain, S. R., Tsien, R. Y., Ormo, M. and Remington, S. J. (1997) Structural basis for dual excitation and

photoisomerization of the *Aequorea victoria* green fluorescent protein. Proc Natl Acad Sci U S A. **94**, 2306-2311
70 Wachter, R. M., Elsliger, M. A., Kallio, K., Hanson, G. T. and Remington, S. J. (1998) Structural basis of spectral shifts in the yellow-emission variants of green fluorescent protein. Structure. **6**, 1267-1277
71 Pedelacq, J. D., Cabantous, S., Tran, T., Terwilliger, T. C. and Waldo, G. S. (2006) Engineering and characterization of a superfolder green fluorescent protein. Nat Biotechnol. **24**, 79-88
72 Heim, R., Cubitt, A. B. and Tsien, R. Y. (1995) Improved green fluorescence. Nature. **373**, 663-664
73 Arnesano, F., Banci, L., Bertini, I., Faraone-Mennella, J., Rosato, A., Barker, P. D. and Fersht, A. R. (1999) The solution structure of oxidized *Escherichia coli* cytochrome b_{562}. Biochemistry. **38**, 8657-8670
74 Hamada, K., Bethge, P. H. and Mathews, F. S. (1995) Refined structure of cytochrome b_{562} from *Escherichia coli* at 1.4 A resolution. J Mol Biol. **247**, 947-962
75 Feng, Y., Sligar, S. G. and Wand, A. J. (1994) Solution structure of apocytochrome b_{562}. Nat Struct Biol. **1**, 30-35
76 Robinson, C. R., Liu, Y., Thomson, J. A., Sturtevant, J. M. and Sligar, S. G. (1997) Energetics of heme binding to native and denatured states of cytochrome b_{562}. Biochemistry. **36**, 16141-16146
77 Rice, J. K., Fearnley, I. M. and Barker, P. D. (1999) Coupled oxidation of heme covalently attached to cytochrome b_{562} yields a novel biliprotein. Biochemistry. **38**, 16847-16856
78 Park, S. Y., Yokoyama, T., Shibayama, N., Shiro, Y. and Tame, J. R. (2006) 1.25 A resolution crystal structures of human haemoglobin in the oxy, deoxy and carbonmonoxy forms. J Mol Biol. **360**, 690-701
79 Melik-Adamyan, W., Bravo, J., Carpena, X., Switala, J., Mate, M. J., Fita, I. and Loewen, P. C. (2001) Substrate flow in catalases deduced from the crystal structures of active site variants of HPII from *Escherichia coli*. Proteins. **44**, 270-281
80 Simonneaux, G. and Bondon, A. (2005) Mechanism of electron transfer in heme proteins and models: the NMR approach. Chem Rev. **105**, 2627-2646
81 Willis, K. J., Szabo, A. G., Zuker, M., Ridgeway, J. M. and Alpert, B. (1990) Fluorescence decay kinetics of the tryptophyl residues of myoglobin: effect of heme ligation and evidence for discrete lifetime components. Biochemistry. **29**, 5270-5275
82 Takeda, S., Kamiya, N. and Nagamune, T. (2003) A novel protein-based heme sensor consisting of green fluorescent protein and apocytochrome b_{562}. Anal Biochem. **317**, 116-119
83 Kumar, S. and Bandyopadhyay, U. (2005) Free heme toxicity and its detoxification systems in human. Toxicol Lett. **157**, 175-188
84 Jones, D. D. and Barker, P. D. (2004) Design and characterisation of an artificial DNA-binding cytochrome. Chembiochem. **5**, 964-971
85 Pia, E. A., et al (2011) Single molecule electron dynamics of a natively-structured redox protein. Submitted to Science

86 Pia, E. A., Chi, Q., Jones, D. D., Macdonald, J. E., Ulstrup, J. and Elliott, M. (2011) Single-molecule mapping of long-range electron transport for a cytochrome b_{562} variant. Nano Lett. **11**, 176-182
87 Vagin, A. A., Teplyakov, A. . (1997) MOLREP: an automated program for molecular replacement. J. Appl. Cryst. **30**, 1022-1025
88 Emsley, P., Lohkamp, B., Scott, W. G. and Cowtan, K. (2010) Features and development of Coot. Acta Crystallogr D Biol Crystallogr. **66**, 486-501
89 Murshudov, G. N., Vagin, A. A. and Dodson, E. J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr. **53**, 240-255
90 Kissinger, C. R., Gehlhaar, D. K. and Fogel, D. B. (1999) Rapid automated molecular replacement by evolutionary search. Acta Crystallogr D Biol Crystallogr. **55**, 484-491
91 McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. and Read, R. J. (2007) Phaser crystallographic software. J. Appl. Cryst.. **40**, 658-674
92 Svergun, D. I., Barberato C. and Koch M.H.J. (1995) CRYOSOL - a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J. Appl. Cryst. **28**, 768-773
93 Franke, D., Svergun, D.I. (2009) DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. J. Appl. Cryst. **42**, 342-346
94 Nixon, A. E. and Firestine, S. M. (2000) Rational and "irrational" design of proteins and their use in biotechnology. IUBMB Life. **49**, 181-187
95 Floudas, C. A., Fung, H.K., McAllister, S.R., Monnigmann, Rajgaria, R. (2006) Advances in protein structure prediction and *de novo* protein design: A review. Chem Eng Sci. **61**, 966-988
96 Naumann, T. A. and Reznikoff, W. S. (2002) Tn5 transposase with an altered specificity for transposon ends. J Bacteriol. **184**, 233-240
97 Reznikoff, W. S., Bhasin, A., Davies, D. R., Goryshin, I. Y., Mahnke, L. A., Naumann, T., Rayment, I., Steiniger-White, M. and Twining, S. S. (1999) Tn5: A molecular window on transposition. Biochem Biophys Res Commun. **266**, 729-734
98 Steiniger-White, M., Bhasin, A., Lovell, S., Rayment, I. and Reznikoff, W. S. (2002) Evidence for "unseen" transposase--DNA contacts. J Mol Biol. **322**, 971-982
99 Royant, A. and Noirclerc-Savoye, M. (2011) Stabilizing role of glutamic acid 222 in the structure of Enhanced Green Fluorescent Protein. J Struct Biol. **174**, 385-390
100 Shortle, D. and Sondek, J. (1995) The emerging role of insertions and deletions in protein engineering. Curr Opin Biotechnol. **6**, 387-393
101 Lodge, J. K., Weston-Hafer, K. and Berg, D. E. (1988) Transposon Tn5 target specificity: preference for insertion at G/C pairs. Genetics. **120**, 645-650
102 Berg, D. E., Schmandt, M. A. and Lowe, J. B. (1983) Specificity of transposon Tn5 insertion. Genetics. **105**, 813-828
103 Chothia, C., Gough, J., Vogel, C. and Teichmann, S. A. (2003) Evolution of the protein repertoire. Science. **300**, 1701-1703
104 Frishman, D. and Argos, P. (1995) Knowledge-based protein secondary structure assignment. Proteins. **23**, 566-579

105 Gerstein, M. (1992) A resolution-sensitive procedure for comparing protein surfaces and its application to the comparison of antigen-combining sites. Acta crystallographica. **A48**, 271-276
106 Yang, F., Moss, L. G. and Phillips, G. N., Jr. (1996) The molecular structure of green fluorescent protein. Nat Biotechnol. **14**, 1246-1251
107 Yang, T. T., Cheng, L. and Kain, S. R. (1996) Optimized codon usage and chromophore mutations provide enhanced sensitivity with the green fluorescent protein. Nucleic Acids Res. **24**, 4592-4593
108 Phillips, G. N., Jr. (1997) Structure and dynamics of green fluorescent protein. Curr Opin Struct Biol. **7**, 821-827
109 Zacharias, D. A., Violin, J. D., Newton, A. C. and Tsien, R. Y. (2002) Partitioning of lipid-modified monomeric GFPs into membrane microdomains of live cells. Science. **296**, 913-916
110 van der Krogt, G. N., Ogink, J., Ponsioen, B. and Jalink, K. (2008) A comparison of donor-acceptor pairs for genetically encoded FRET sensors: application to the Epac cAMP sensor as an example. PLoS One. **3**, e1916
111 Stepanenko, O. V., Verkhusha, V. V., Kazakov, V. I., Shavlovsky, M. M., Kuznetsova, I. M., Uversky, V. N. and Turoverov, K. K. (2004) Comparative studies on the structure and stability of fluorescent proteins EGFP, zFP506, mRFP1, "dimer2", and DsRed1. Biochemistry. **43**, 14913-14923
112 Verkhusha, V. V., Kuznetsova, I. M., Stepanenko, O. V., Zaraisky, A. G., Shavlovsky, M. M., Turoverov, K. K. and Uversky, V. N. (2003) High stability of *Discosoma* DsRed as compared to *Aequorea* EGFP. Biochemistry. **42**, 7879-7884
113 Huang, J. R., Craggs, T. D., Christodoulou, J. and Jackson, S. E. (2007) Stable intermediate states and high energy barriers in the unfolding of GFP. J Mol Biol. **370**, 356-371
114 Andrews, B. T., Schoenfish, A. R., Roy, M., Waldo, G. and Jennings, P. A. (2007) The rough energy landscape of superfolder GFP is linked to the chromophore. J Mol Biol. **373**, 476-490
115 Steiner, T., Hess, P., Bae, J. H., Wiltschi, B., Moroder, L. and Budisa, N. (2008) Synthetic biology of proteins: tuning GFPs folding and stability with fluoroproline. PLoS One. **3**, e1680
116 Wedemeyer, W. J., Welker, E. and Scheraga, H. A. (2002) Proline cis-trans isomerization and protein folding. Biochemistry. **41**, 14637-14644
117 Siemering, K. R., Golbik, R., Sever, R. and Haseloff, J. (1996) Mutations that suppress the thermosensitivity of green fluorescent protein. Curr Biol. **6**, 1653-1663
118 Zimmer, M. (2002) Green fluorescent protein (GFP): applications, structure, and related photophysical behavior. Chem Rev. **102**, 759-781
119 Saeed, I. A. and Ashraf, S. S. (2009) Denaturation studies reveal significant differences between GFP and blue fluorescent protein. Int J Biol Macromol. **45**, 236-241
120 Attila Nagy, A. M. C., Bela Somogyi, Denes Lorinczy. (2004) Thermal stability of chemically denatured green fluorescent protein (GFP) A preliminary study. Thermochimica Acta. **410**, 161-163
121 Abedi, M. R., Caponigro, G. and Kamb, A. (1998) Green fluorescent protein as a scaffold for intracellular presentation of peptides. Nucleic Acids Res. **26**, 623-630

122 Flores-Ramirez, G., Rivera, M., Morales-Pablos, A., Osuna, J., Soberon, X. and Gaytan, P. (2007) The effect of amino acid deletions and substitutions in the longest loop of GFP. BMC Chem Biol. **7**, 1
123 Li, X., Zhang, G., Ngo, N., Zhao, X., Kain, S. R. and Huang, C. C. (1997) Deletions of the *Aequorea victoria* green fluorescent protein define the minimal domain required for fluorescence. J Biol Chem. **272**, 28545-28549
124 Chiang, C. F., Okou, D. T., Griffin, T. B., Verret, C. R. and Williams, M. N. (2001) Green fluorescent protein rendered susceptible to proteolysis: positions for protease-sensitive insertions. Arch Biochem Biophys. **394**, 229-235
125 Dopf, J. and Horiagon, T. M. (1996) Deletion mapping of the *Aequorea victoria* green fluorescent protein. Gene. **173**, 39-44
126 Meinhold, D. W. and Wright, P. E. (2011) Measurement of protein unfolding/refolding kinetics and structural characterization of hidden intermediates by NMR relaxation dispersion. Proc Natl Acad Sci U S A. **108**, 9078-9083
127 Campanini, B., Bologna, S., Cannone, F., Chirico, G., Mozzarelli, A. and Bettati, S. (2005) Unfolding of Green Fluorescent Protein mut2 in wet nanoporous silica gels. Protein Sci. **14**, 1125-1133
128 Fukuda, H., Arai, M. and Kuwajima, K. (2000) Folding of green fluorescent protein and the cycle3 mutant. Biochemistry. **39**, 12025-12032
129 Channon, K., Bromley, E. H. and Woolfson, D. N. (2008) Synthetic biology through biomolecular design and engineering. Curr Opin Struct Biol. **18**, 491-498
130 Astier, Y., Bayley, H. and Howorka, S. (2005) Protein components for nanodevices. Curr Opin Chem Biol. **9**, 576-584
131 Vallee-Belisle, A. and Plaxco, K. W. (2010) Structure-switching biosensors: inspired by Nature. Curr Opin Struct Biol. **20**, 518-526
132 Ostermeier, M. (2009) Designing switchable enzymes. Curr Opin Struct Biol. **19**, 442-448
133 Ambroggio, X. I. and Kuhlman, B. (2006) Design of protein conformational switches. Curr Opin Struct Biol. **16**, 525-530
134 Koide, S. (2009) Generation of new protein functions by nonhomologous combinations and rearrangements of domains and modules. Curr Opin Biotechnol. **20**, 398-404
135 Gunasekaran, K., Ma, B. and Nussinov, R. (2004) Is allostery an intrinsic property of all dynamic proteins? Proteins. **57**, 433-443
136 Guntas, G., Mansell, T. J., Kim, J. R. and Ostermeier, M. (2005) Directed evolution of protein switches and their application to the creation of ligand-binding proteins. Proc Natl Acad Sci U S A. **102**, 11224-11229
137 Guntas, G., Mitchell, S. F. and Ostermeier, M. (2004) A molecular switch created by *in vitro* recombination of nonhomologous genes. Chem Biol. **11**, 1483-1487
138 Takeda, S., Kamiya, N. and Nagamune, T. (2004) Rational design of a protein-based molecular device consisting of blue fluorescent protein and zinc protoporphyrin IX incorporated into a cytochrome b_{562} scaffold. Biotechnol Lett. **26**, 121-125
139 Akerboom, J., Rivera, J. D., Guilbe, M. M., Malave, E. C., Hernandez, H. H., Tian, L., Hires, S. A., Marvin, J. S., Looger, L. L. and Schreiter, E. R. (2009) Crystal

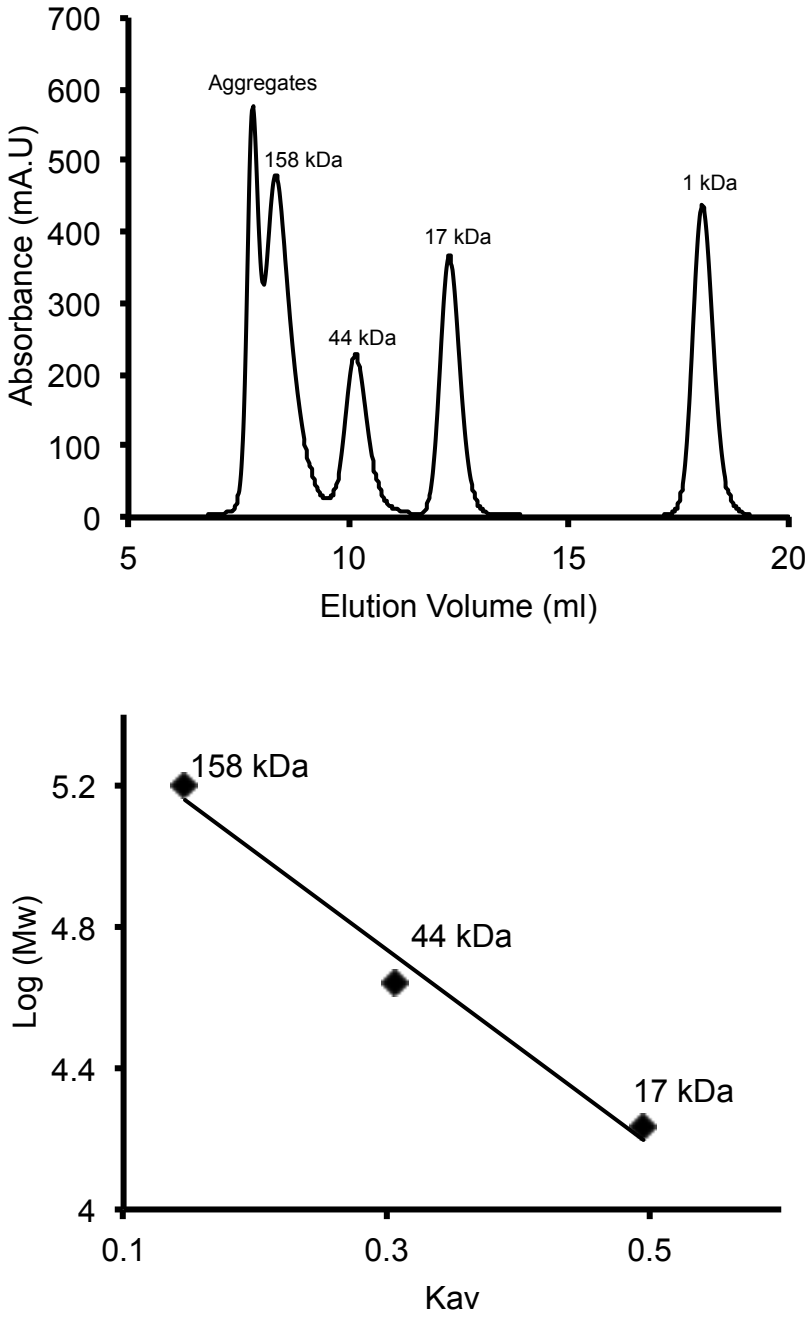
structures of the GCaMP calcium sensor reveal the mechanism of fluorescence signal change and aid rational design. J Biol Chem. **284**, 6455-6464
140 Topell, S., Hennecke, J. and Glockshuber, R. (1999) Circularly permuted variants of the green fluorescent protein. FEBS Lett. **457**, 283-289
141 Itagaki, E. and Hager, L. P. (1966) Studies on cytochrome b_{562} of *Escherichia coli*. I. Purification and crystallization of cytochrome b_{562}. J Biol Chem. **241**, 3687-3695
142 Reid, B. G. and Flynn, G. C. (1997) Chromophore formation in green fluorescent protein. Biochemistry. **36**, 6786-6791
143 Arai, R., Ueda, H., Kitayama, A., Kamiya, N. and Nagamune, T. (2001) Design of the linkers which effectively separate domains of a bifunctional fusion protein. Protein Eng. **14**, 529-532
144 Borst, J. W., Willemse, M., Slijkhuis, R., van der Krogt, G., Laptenok, S. P., Jalink, K., Wieringa, B. and Fransen, J. A. (2010) ATP changes the fluorescence lifetime of cyan fluorescent protein via an interaction with His148. PLoS One. **5**, e13862
145 VanBeek, D. B., Zwier, M. C., Shorb, J. M. and Krueger, B. P. (2007) Fretting about FRET: correlation between kappa and R. Biophys J. **92**, 4168-4178
146 Copty, A., Sakran, F., Popov, O., Ziblat, R., Danieli, T., Golosovsky, M., Davidov, D. (2005) Probing the microwave radiation effect on the green fluorescent protein luminescence in solution. Synthetic Metals. **155**, 422-425
147 Kiss, C., Fisher, H., Pesavento, E., Dai, M., Valero, R., Ovecka, M., Nolan, R., Phipps, M. L., Velappan, N., Chasteen, L., Martinez, J. S., Waldo, G. S., Pavlik, P. and Bradbury, A. R. (2006) Antibody binding loop insertions as diversity elements. Nucleic Acids Res. **34**, e132
148 Doi, N. and Yanagawa, H. (1999) Insertional gene fusion technology. FEBS Lett. **457**, 1-4
149 Choi, J. W., Nam, Y. S., Park, S. J., Lee, W. H., Kim, D. and Fujihira, M. (2001) Rectified photocurrent of molecular photodiode consisting of cytochrome c/GFP hetero thin films. Biosens Bioelectron. **16**, 819-825
150 Choi, J. W., Nam, Y.S., Lee, B.H., Ahn, D.J., Nagamune, T. . (2006) Charge trap in self-assembled monolayer of cytochrome b_{562}-green fluorescent protein chimera. Current Applied Physics. **6**, 760-765
151 Bogdanov, A. M., Mishin, A. S., Yampolsky, I. V., Belousov, V. V., Chudakov, D. M., Subach, F. V., Verkhusha, V. V., Lukyanov, S. and Lukyanov, K. A. (2009) Green fluorescent proteins are light-induced electron donors. Nat Chem Biol. **5**, 459-461
152 Scruggs, A. W., Flores, C. L., Wachter, R. and Woodbury, N. W. (2005) Development and characterization of green fluorescent protein mutants with altered lifetimes. Biochemistry. **44**, 13377-13384
153 Kumsta, C. and Jakob, U. (2009) Redox-regulated chaperones. Biochemistry. **48**, 4666-4676
154 Dooley, C. T., Dore, T. M., Hanson, G. T., Jackson, W. C., Remington, S. J. and Tsien, R. Y. (2004) Imaging dynamic redox changes in mammalian cells with green fluorescent protein indicators. J Biol Chem. **279**, 22284-22293
155 Lesser, M. P. (1996) Elevated temperatures and ultraviolet radiation cause oxidative stress and inhibit photosynthesis in symbiotic dinoflagellates. American Society of Limnology and Oceanography. **41**, 271-283

156 Sevrioukova, I. F. (2009) Redox-linked conformational dynamics in apoptosis-inducing factor. J Mol Biol. **390**, 924-938
157 Turner, N. A., Xia, F., Azhar, G., Zhang, X., Liu, L. and Wei, J. Y. (1998) Oxidative stress induces DNA fragmentation and caspase activation via the c-Jun NH2-terminal kinase pathway in H9c2 cardiac muscle cells. J Mol Cell Cardiol. **30**, 1789-1801
158 Iles, K. E. and Forman, H. J. (2002) Macrophage signaling and respiratory burst. Immunol Res. **26**, 95-105
159 Barford, D. (2004) The role of cysteine residues as redox-sensitive regulatory switches. Curr Opin Struct Biol. **14**, 679-686
160 Fan, S. W., George, R. A., Haworth, N. L., Feng, L. L., Liu, J. Y. and Wouters, M. A. (2009) Conformational changes in redox pairs of protein structures. Protein Sci. **18**, 1745-1765
161 Kim, S. J., Jeong, D. G., Chi, S. W., Lee, J. S. and Ryu, S. E. (2001) Crystal structure of proteolytic fragments of the redox-sensitive Hsp33 with constitutive chaperone activity. Nat Struct Biol. **8**, 459-466
162 Graumann, J., Lilie, H., Tang, X., Tucker, K. A., Hoffmann, J. H., Vijayalakshmi, J., Saper, M., Bardwell, J. C. and Jakob, U. (2001) Activation of the redox-regulated molecular chaperone Hsp33--a two-step mechanism. Structure. **9**, 377-387
163 Vijayalakshmi, J., Mukhergee, M. K., Graumann, J., Jakob, U. and Saper, M. A. (2001) The 2.2 A crystal structure of Hsp33: a heat shock protein with redox-regulated chaperone activity. Structure. **9**, 367-375
164 Choi, H., Kim, S., Mukhopadhyay, P., Cho, S., Woo, J., Storz, G. and Ryu, S. E. (2001) Structural basis of the redox switch in the OxyR transcription factor. Cell. **105**, 103-113
165 Hanson, G. T., Aggeler, R., Oglesbee, D., Cannon, M., Capaldi, R. A., Tsien, R. Y. and Remington, S. J. (2004) Investigating mitochondrial redox potential with redox-sensitive green fluorescent protein indicators. J Biol Chem. **279**, 13044-13053
166 Lohman, J. R. and Remington, S. J. (2008) Development of a family of redox-sensitive green fluorescent protein indicators for use in relatively oxidizing subcellular environments. Biochemistry. **47**, 8678-8688
167 Cannon, M. B. and Remington, S. J. (2006) Re-engineering redox-sensitive green fluorescent protein for improved response rate. Protein Sci. **15**, 45-57
168 Ostergaard, H., Henriksen, A., Hansen, F. G. and Winther, J. R. (2001) Shedding light on disulfide bond formation: engineering a redox switch in green fluorescent protein. Embo J. **20**, 5853-5862
169 Shaked, Z., Szajewski, R. P. and Whitesides, G. M. (1980) Rates of thiol-disulfide interchange reactions involving proteins and kinetic measurements of thiol pKa values. Biochemistry. **19**, 4156-4166
170 Kneen, M., Farinas, J., Li, Y. and Verkman, A. S. (1998) Green fluorescent protein as a noninvasive intracellular pH indicator. Biophys J. **74**, 1591-1599
171 Waris, G. and Ahsan, H. (2006) Reactive oxygen species: role in the development of cancer and various chronic conditions. J Carcinog. **5**, 14
172 Gomes, A., Fernandes, E. and Lima, J. L. (2005) Fluorescence probes used for detection of reactive oxygen species. J Biochem Biophys Methods. **65**, 45-80

173 Jung, G., Wiehler, J. and Zumbusch, A. (2005) The photophysics of green fluorescent protein: influence of the key amino acids at positions 65, 203, and 222. Biophys J. **88**, 1932-1947
174 Lichtenstein, C. and Brenner, S. (1981) Site-specific properties of Tn7 transposition into the E. coli chromosome. Mol Gen Genet. **183**, 380-387
175 Berry, S. M., Lu, Y. . (2011) Protein structure design and engineering. Encyclopedia of life sciences
176 Lee, B., Takeda, S., Nakajima, K., Noh, J., Choi, J., Hara, M. and Nagamune, T. (2004) Rectified photocurrent in a protein based molecular photo-diode consisting of a cytochrome b_{562}-green fluorescent protein chimera self-assembled monolayer. Biosens Bioelectron. **19**, 1169-1174

Appendices

Appendix A. Superdex™ 75 analytical size exclusion chromatography standards



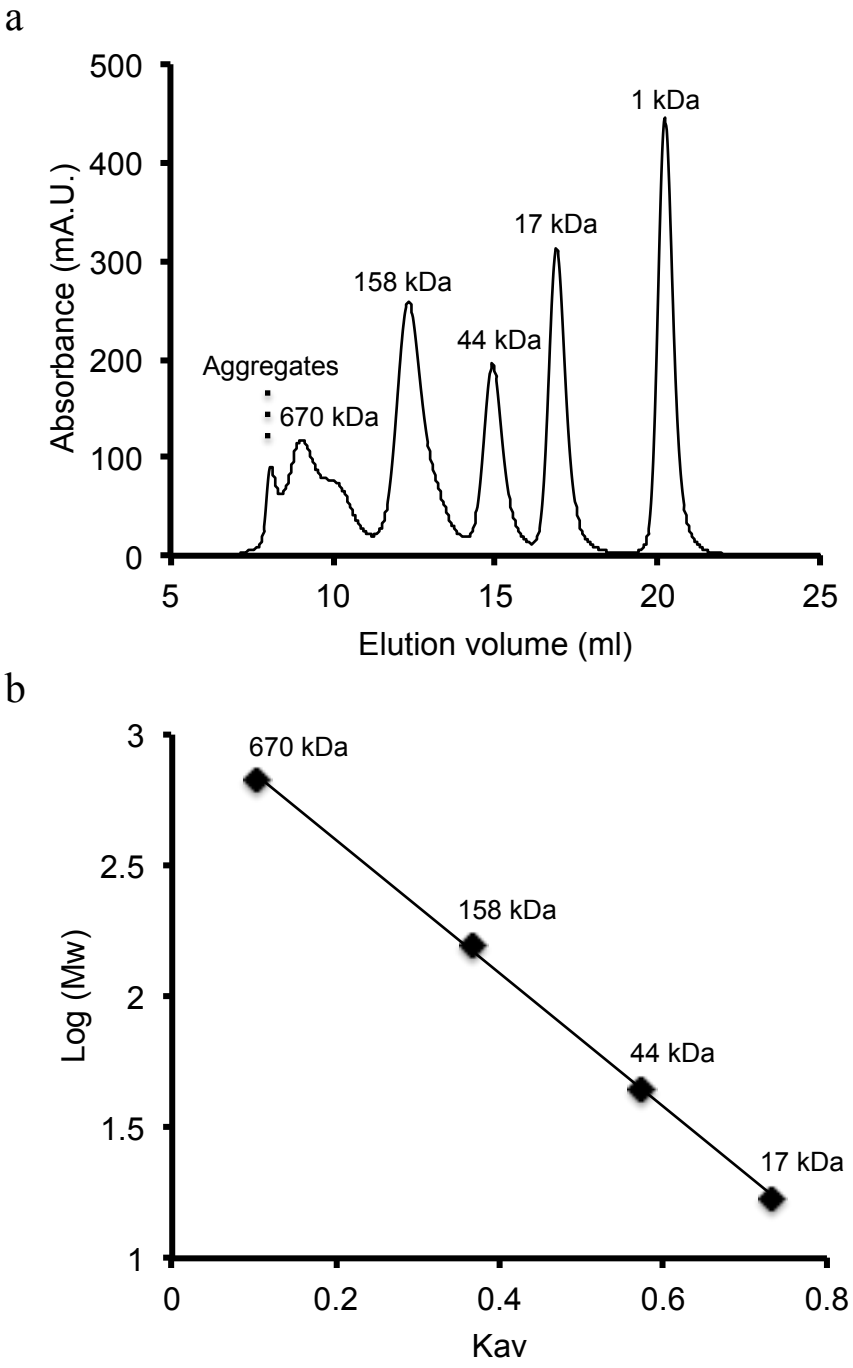
Appendix A. Analytical size exclusion chromatography standards and standard curve. a, Elution profile for BioRad gel filtration standards (Section, 2.6.2.1, Table 2.1). Elution of protein standards from a Superdex™ 75 gel filtration column was monitored by absorbance at 280 nm. **b,** Standard curve of log(Mw) against K_{av}. K_{av} values are calculated using Equation 11 (Section 2.6.2.1).

Appendix B. Triplet nucleotide deletions positions sampled during domain insertion library construction that resulted in out of cyt b_{562} domains.

Variant	TND[a]	Variant	TND[a]
NF196	4 TGA 8	NF29	259 CCG 263
NF120	28 CCG 32	NF54	259 CCG 263
NF15	46 TGG 50	NF174	281 CCA 285
NF179	59 CGG 63	NF03	309 GAC 313
NF48	71 CGG 75	NF52	311 CGG 315
NF12	72 GGC 76	NF113	317 CTA 321
NF183	86 CGT 90	NF166	319 ACA 323
NF51	91 CCG 95	NF152	328 GCG 332
NF86	91 CCG 95	NF121	347 GGG 351
NF159	91 CCG 95	NF44	371 CGA 375
NF138	92 CGG 96	NF87	377 GAA 381
NF154	93 GGC 97	NF163	379 AGG 383
NF79	94 GCG 97	NF30	382 GCA 386
NF43	101 CGA 105	NF06	413 GGG 417
NF158	101 CGA 105	NF143	414 GGG 418
NF34	107 CGA 111	NF47	424 TGG 428
NF11	108 GAT 112	NF94	434 CTA 437
NF33	112 CCA 116	NF164	440 CAG 444
NF55	112 CCA 116	NF04	443 CCA 447
NF103	112 CCA 116	NF146	452 CTA 456
NF170	112 CCA 116	NF68	463 CCG 467
NF119	116 CTA 120	NF141	463 CCG 467
NF134	118 ACG 122	NF132	487 AGG 491
NF137	118 ACG 122	NF175	503 CCG 507
NF41	127 TGA 131	NF131	515 CGA 519
NF112	143 CTG 147	NF207	521 CGG 525
NF194	143 CTG 147	NF147	522 GGC 526
NF18	151 CCG 155	NF206	539 CGA 543
NF202	151 CCG 155	NF27	540 GAC 544
NF126	171 TGG 175	NF169	603 CTG 607
NF190	171 TGG 175	NF203	620 CCT 624
NF208	171 TGG 175	NF185	634 CCA 637
NF101	215 CAG 219	NF93	660 CTG 664
NF182	217 GCC 221	NF76	661 TGC 665
NF70	250 TCT 254	NF88	712 ACA 716

[a] Gene numbering is for the *egfp* gene [107]. The triplet nucleotide displayed is the sequence deleted upon removal of the transposon, MuDel, by restriction digestion. TND stands for triplet nucleotide deletion.

Appendix C. Superdex™ 200 analytical size exclusion chromatography standards



Appendix C. Analytical size exclusion chromatography standards and standard curve. a, Elution profile for BioRad gel filtration standards (Section, 2.6.2.1, Table 2.1). Elution of protein standards from a Superdex™ 200 gel filtration column was monitored by absorbance at 280 nm. **b,** Standard curve of log(Mw) against K_{av}. K_{av} values are calculated using Equation 11 (Section 2.6.2.1).

Publications

Baldwin, A. J., Arpino, J. A., Edwards, W. R., Tippmann, E. M. and Jones, D. D. (2009) Expanded chemical diversity sampling through whole protein evolution. Mol Biosyst. **5**, 764-766

Arpino, J.A.J, Czapinska, H., Piasecka, A., Edwards, W.R., Barker, P., Gajda, M., Bochtler, M., Jones, D.D. (2011) Structural basis for efficient energy transfer in a constructed protein scaffold. Submitted to PNAS