Structure
Article

Random Single Amino Acid Deletion Sampling Unveils Structural Tolerance and the Benefits of Helical Registry Shift on GFP Folding and Structure

James A.J. Arpino,[1,3] Samuel C. Reddington,[1] Lisa M. Halliwell,[1] Pierre J. Rizkallah,[2] and D. Dafydd Jones[1,*]
[1]School of Biosciences, Main Building, Park Place, Cardiff University, Cardiff CF10 3AT, UK
[2]School of Medicine, Cardiff University, WHRI, Main Building, Heath Park, Cardiff CF14 4XN, UK
[3]Present address: Centre for Synthetic Biology and Innovation, Imperial College London, London SW7 2AZ, UK
*Correspondence: jonesdd@cf.ac.uk
http://dx.doi.org/10.1016/j.str.2014.03.014

SUMMARY

Altering a protein's backbone through amino acid deletion is a common evolutionary mutational mechanism, but is generally ignored during protein engineering primarily because its effect on the folding-structure-function relationship is difficult to predict. Using directed evolution, enhanced green fluorescent protein (EGFP) was observed to tolerate residue deletion across the breadth of the protein, particularly within short and long loops, helical elements, and at the termini of strands. A variant with G4 removed from a helix (EGFP$^{G4\Delta}$) conferred significantly higher cellular fluorescence. Folding analysis revealed that EGFP$^{G4\Delta}$ retained more structure upon unfolding and refolded with almost 100% efficiency but at the expense of thermodynamic stability. The EGFP$^{G4\Delta}$ structure revealed that G4 deletion caused a beneficial helical registry shift resulting in a new polar interaction network, which potentially stabilizes a *cis* proline peptide bond and links secondary structure elements. Thus, deletion mutations and registry shifts can enhance proteins through structural rearrangements not possible by substitution mutations alone.

INTRODUCTION

Protein backbone mutations or amino acid insertion/deletion (InDel) events are an important part of the natural evolutionary process (de Jong and Rydén, 1981; Leushkin et al., 2012; Taylor et al., 2004; Tóth-Petróczy and Tawfik, 2013) and affect protein structure in a manner distinct to that of side chain substitution (Pascarella and Argos, 1992; Shortle and Sondek, 1995). InDels are now thought to be key contributors to the evolutionary process by instigating major leaps in the protein fitness landscape (Leushkin et al., 2012; Tóth-Petróczy and Tawfik, 2013). Thus, InDels provide a new route to increase sequence and structure sampling space during the protein engineering process (Shortle and Sondek, 1995). However, whether site-directed, computationally or directed evolution-driven, the main focus of protein engineering is the generation of amino acid substitutions. The absence of InDel mutagenesis as part of the routine protein engineering toolbox is partly due to the difficultly in predicting the local and global structural influence of altering the protein backbone; dogma suggests such mutations are likely to be detrimental due to, for example, disruptive registry shifts in organized secondary structure and perturbing folding pathways (Pascarella and Argos, 1992; Shortle and Sondek, 1995). Consequently, there have been relatively few studies concerning the structural impact of engineered InDel mutations (Arpino et al., 2012a; Heinz et al., 1993; O'Neil et al., 2000; Stott et al., 2009; Vetter et al., 1996), especially regarding how any beneficial effects are exerted at the molecular level (Arpino et al., 2012a). Most of these studies have focused on site-directed introduction of InDels, so little information is available on the general tolerance of proteins to InDels and their beneficial effect.

One of the most common backbone mutations observed among protein homologs is deletion of single amino acids (de Jong and Rydén, 1981; Taylor et al., 2004), leading to, for example, expansion of the antibody repertoire (de Wildt et al., 1999), emergence of HIV drug resistance (Imamichi et al., 2001; Wood et al., 2009), herbicide resistance (Patzoldt et al., 2006), and resistance to third-generation β-lactam antibiotics (Jones, 2005; Simm et al., 2007). The potential benefits of sampling single amino acid deletions as part of protein engineering endeavors has recently emerged predominately through the advent of directed-evolution approaches (Bershtein and Tawfik, 2008). Such directed-evolution approaches that sample single amino acid deletions (Fujii et al., 2006; Jones, 2005; Murakami et al., 2002) has allowed broader sampling across the whole protein backbone, which removes any perceived bias concerning tolerance and impact on the target protein. This in turn allows retrospective structural analysis to understand the molecular basis of action.

Here, we have applied a transposon-mediated directed-evolution trinucleotide deletion (TND) approach (Jones, 2005; Simm et al., 2007) to the commercially important and widely used enhanced green fluorescent protein (EGFP), an engineered variant of the original *Aequorea victoria* GFP (Tsien, 1998). EGFP is an important tool in cell biology (Tsien, 1998; Zhang et al., 2002) but still has some limitations resulting in further protein engineering endeavors to improve properties such as stability, folding efficiency, and solubility during cellular expression. Despite their relatively high thermodynamic stability (Tsien,

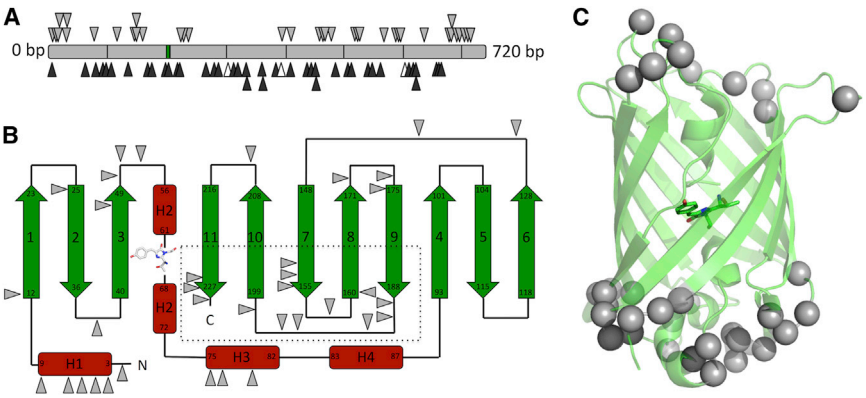


Figure 1. Mapping Deletion Mutations with Respect to EGFP Primary, Secondary, and Tertiary Structure

(A) Gene sequence analysis of fluorescent (gray triangles) and nonfluorescent variants (black triangles) selected during the screening process identified the position of the triplet nucleotide deleted from *egfp* (gray bar). Nonfluorescent variants due to a TND and subsequent introduction of a premature stop codon are highlighted by white triangles.

(B) The secondary structure arrangement and overall topology of EGFP shows the arrangement of β strands (green), α helices (red), and loops (black). Tolerated single amino acid deletions are indicated by gray triangles with an area particularly tolerant to deletion mutations surrounded by a dotted line.

(C) Map of single amino acid deletions onto the tertiary structure of EGFP. Cartoon representation of EGFP (green) with tolerated deletions indicated by gray spheres.

1998), GFPs suffer from off-pathway aggregation due to their slow folding and maturation times (Fukuda et al., 2000). The transposon-based tool has been used previously to sample various domain insertion and codon replacement events relating to EGFP (Arpino et al., 2012a; Baldwin et al., 2009) but without a thorough analysis of the impact of TNDs. Through the construction and screening of a TND library, the general tolerance and impact of single amino acid deletions were explored. A variant with a single amino acid deletion was identified that conferred a brighter fluorescence phenotype on *Escherichia coli*. Rather than altering the spectral characteristics, the deletion mutation caused local structural rearrangements in a 3_{10} helix, resulting in new long-range polar interactions, including with the sole *cis* proline peptide bond.

RESULTS

Tolerance of EGFP to Single Amino Acid Deletion

The EGFP TND library was constructed essentially as described elsewhere (Baldwin et al., 2009; Simm et al., 2007) and screened for a green fluorescence *E. coli* phenotype upon irradiation with near UV light. Only correctly folded EGFP variants bestow the fluorescent green phenotype on *E. coli*. Of all the colonies screened, 10% displayed green fluorescence after extended growth, with 2.5% displaying noticeable green fluorescence after 24 hr at 37°C. A total of 153 colonies were chosen based on their observable color phenotype (88 fluorescent and 65 nonfluorescent) and the *egfp* gene sequenced. Of the 88 fluorescent variants sequenced, 42 different TNDs were identified and from the 65 nonfluorescent variants, 45 were unique TNDs; the total unique TNDs observed was 87. No wild-type EGFP was observed and no additional point mutations or frameshifts were observed in any of the sequenced variants. The distribution of the TNDs is shown in Figure 1A, with more detailed sequence information in Tables S1 and S2 available online. Observed mutations were distributed throughout the *egfp* gene, allowing thorough analysis of EGFP tolerance to single amino acid deletions. Due to the mechanism by which the library is constructed, the TND can span two codons and may give rise to a single amino acid deletion and an adjacent substitution mutation (Jones, 2005; Simm et al., 2007). Cross-codon TNDs can therefore introduce premature stop codons depending on the sequence surrounding the TND; this was observed for four variants and is likely to result in truncated nonfluorescent protein (Figure 1 and Table S2).

The position of the tolerated mutations in relation to the secondary and tertiary structure of EGFP is shown in Figures 1B and 1C. The majority of tolerated single amino acid deletions are found in loops connecting organized secondary structure (60%). The rest are equally distributed across helices (19%) and β strands (21%). The majority tolerated within β strands were found toward the strand termini, with the C-terminal ends of strands 7–11 particularly tolerant (Figure 1B). In relation to the tertiary structure, these sites translate to the two ends of the β barrel (Figure 1C).

This survey highlights the loops of EGFP as more tolerant to single amino acid deletions whereas β strands are the least tolerant. The proportion of EGFP comprising loops and β strands is 43% and 46%, respectively, which contrasts to the observed frequency of 60% and 21% of tolerated sites. In comparison, the observed frequency of tolerated positions in helical regions (19%) is ~2 fold higher than helical contribution to the composition of EGFP secondary structure (11%). The N-terminal helix H1 and barrel capping helix H3 were particularly tolerant (Figure 1B). No tolerated sites were found in the core helical structure housing the chromophore. Of the 45 unique mutations giving rise to nonfluorescent variants, the majority (71%) are located in the middle of β strands, with 18% located in loops and 11% in helices (Figure S1 and Table S2). These included mutations that affected the three chromophore-forming residues, T65, Y66, and G67 (Table S2). Loops were not wholly tolerant to deletions but sensitive to the residue deleted. For example, removal of L137 in the long loop linking strands 6 and 7 was tolerated but deletion of residues D133 or G138 was not. Therefore, the residue removed in loops will dictate the structural rearrangements that occur rather than a general "whole loop" effect being observed.

The solvent-accessible surface area (SASA), an indicator of residue burial, of residues tolerant to deletion was relatively well distributed across the whole range with higher solvent

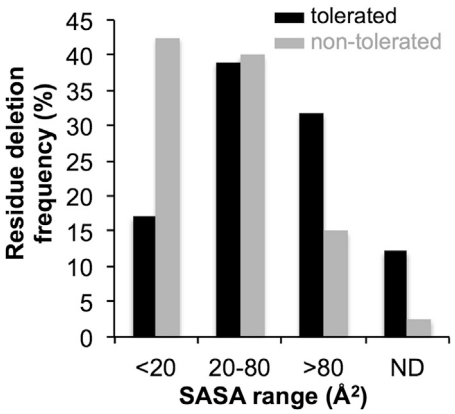


Figure 2. Relationship between EGFP and SASA

Frequency of tolerance (black) and nontolerance (gray) of EGFP residues to deletion and their SASA. ND, not determined because their value could not be calculated because either the residue lies toward the N or C termini and are not part of the determined structure (PDB code: 4EUL; in the case of tolerated deletions) or form part of the chromophore (in the case of the nontolerated deletions).

exposed residues more tolerant (Figure 2 and Table S1). However, there was a trend toward lower SASA values for residues not tolerant to deletion (Figure 2 and Table S2). This stands to reason given that deletions like substitutions, of an amino acid buried in the core of a protein can be disruptive to protein structure and function. The relationship between tolerance to deletion and SASA has a link to the type of residue. Glycine and threonine residues are frequently observed to be tolerant deletions, both of which have relatively low inherent SASA. Larger residues such as leucine and tyrosine are less tolerant to deletion. Proline residues, including P89 involved in the sole *cis* peptide bond, appear to be tolerant to deletion; four of the five variants with a proline deleted still conferred a fluorescence phenotype on *E. coli* (Figure 1 and Tables S1 and S2).

Identification and Fluorescence Properties of EGFP$^{G4\Delta}$

Certain colonies appeared brighter than the general background level and sequencing revealed that the predominant mutation was G4Δ (where Δ refers to a deletion) resident in the N-terminal 3_{10} helix (H1; Figure 3A). Removal of G4 is likely to alter the registry (or relative side chain position) of the helix quite dramatically, as indicated by the helical wheel representation (Figure 3A). Cells expressing EGFP$^{G4\Delta}$ were visibly brighter than those expressing EGFP when grown in parallel on agar plates at 37°C (Figure 3B and Figure S2). The beneficial effects were found to be transferable as introduction of G4Δ to enhanced yellow fluorescent protein improved cellular fluorescence (Figure 3B). Analysis of whole cell fluorescence of cultures grown at 37°C (standardized to absorption at 600 nm of 0.1) induced for the same time revealed that those harboring EGFP$^{G4\Delta}$ were more fluorescent (~2-fold) than EGFP; increased fluorescence intensity was even more pronounced (~4-fold) for cultures grown at 25°C (Figure 3C). In vitro analysis of pure protein revealed that G4Δ did not affect EGFP fluorescence parameters (EGFP versus EGFP$^{G4\Delta}$) because the quantum yields (0.60 versus 0.59), mM extinction coefficients (55 versus 53), and fluorescence lifetimes (2.5 ns versus 2.6 ns) were very similar. This suggests that increased apparent brightness is the result of more efficient production of stable fluorescing protein in the cell rather than inherent changes to fluorescence.

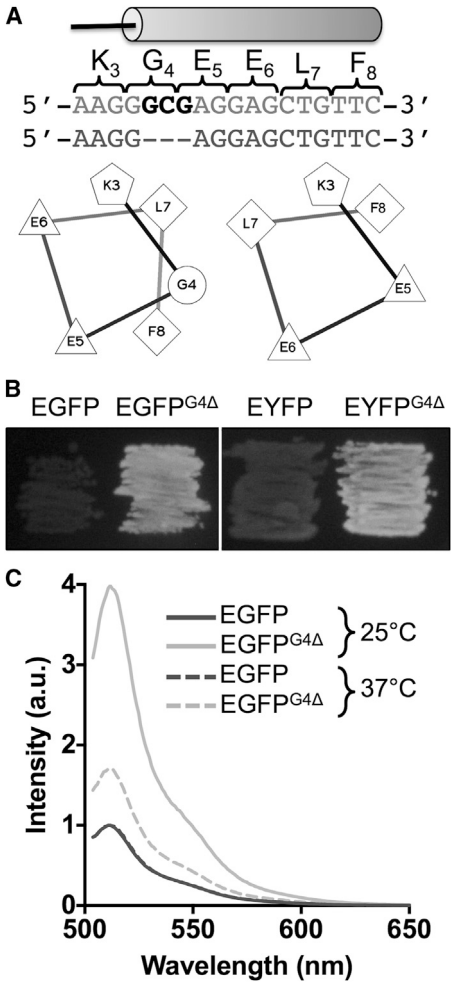


Figure 3. Fluorescence Properties of EGFP$^{G4\Delta}$

(A) The trinucleotide deletion giving rise to the G4Δ mutation and the potential helical register shift as represented by a helical wheel (hydrophobic, diamond; acidic, triangle; basic, pentagon).

(B) Cellular fluorescence of the EGFP and EYFP G4Δ variants (color version available in Figure S2).

(C) Whole cell fluorescence emission (excited at 488 nm) spectra for cultures grown at either 25°C (solid dashed lines) or 37°C (dashed lines). Cell cultures were standardized to an optical density 600 of 0.1 and the spectra normalized to EGFP fluorescence intensity.

Folding Properties of EGFP$^{G4\Delta}$

Because there was little effect on intrinsic fluorescence, the influence of the G4Δ mutation on EGFP folding and stability was probed. It should be noted that EGFP is typical of similarly structured fluorescent proteins in that they are very resilient to chemical denaturation (Tsien, 1998) and take considerable time to reach equilibrium (>1 week; Figure S2; Hsu et al., 2009). Therefore, all reported equilibrium values are considered apparent even though after 250 hr incubation (time used in these studies)

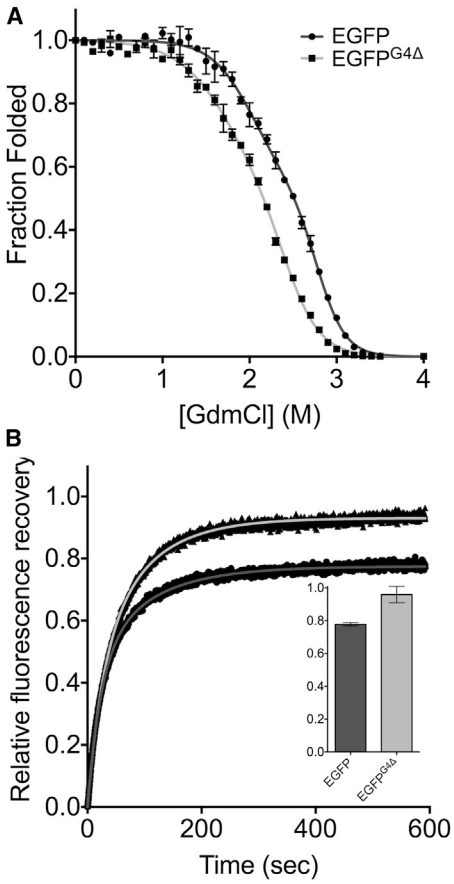


Figure 4. Folding Properties of EGFP$^{G4\Delta}$

(A) Equilibrium unfolding of EGFP (circles) and EGFP$^{G4\Delta}$ (squares). The curves were fit to a three-state model as outlined in the Experimental Procedures. Error bars represent SD of three replicates.

(B) Refolding kinetics of EGFP (dark gray) and EGFP$^{G4\Delta}$ (light gray). The curves were fit to a double exponential as outlined in the Experimental Procedures and Supplemental Experimental Procedures. The final fluorescence recovery yield in refolding is shown in the inset graph.

unfolding was approaching equilibrium based on the change in $[GdmCl]_{50\%}$ (Figure S3). The unfolding curves fit best to a three-state model (Figure 4A; see Figure S4 for fit to two-state model). The observed three-state unfolding suggests a folding intermediate is formed, as has been suggested for related GFPs (Hsu et al., 2009). The $[GdmCl]_{50\%}^{app}$ for the native protein to intermediate transition (N-I) and intermediate to the denatured state transition (I-D) were similar, with EGFP$^{G4\Delta}$ being ~0.3 M lower than EGFP at both transitions (Table 1). However, the dependencies of ΔG on [GdmCl] (or m values) for both transitions show a significant difference (Table 1). The m value for the N-I transition for EGFP is larger than that for EGFP$^{G4\Delta}$ by ~0.65 kcal mol^{-1} M^{-1}; this difference is more pronounced for the I-D transition with EGFP being ~0.81 kcal mol^{-1} M^{-1} larger than EGFP$^{G4\Delta}$. The differences in m value thus have a major impact on the apparent stability ($\Delta G^{H_2O}_{N\text{-}D}$; Table 1), with EGFP$^{G4\Delta}$ unexpectedly destabilized by 5.19 kcal mol^{-1} compared to EGFP. However, $\Delta G^{H_2O}_{N\text{-}D}$ for EGFP$^{G4\Delta}$ is still relatively high (10.7 kcal mol^{-1}) and far from being at the margin of stability. The melting temperature for thermal denaturation also remains high for EGFP$^{G4\Delta}$ (83°C) and is similar to that of EGFP (84°C; Figure S5).

While deleting G4 from EGFP appears to result in an overall decrease in stability, the m values provide insight into the structure of the intermediate and unfolded state. Because equilibrium unfolding m values are strongly correlated with a change in SASA, it is possible to estimate the change in surface area for the N-I and I-D transitions (calculated using http://www-clarke.ch.cam.ac.uk/BPPred.php; Geierhaas et al., 2007; Table 2). The calculated SASA for native EGFP is only slightly higher than that for EGFP$^{G4\Delta}$, differing by ~550 Å^2. For both transitions, the ΔSASA for EGFP$^{G4\Delta}$ is less than that for EGFP ($\Delta\Delta SASA_{N\text{-}I}$ = −1,620 Å^2 and $\Delta\Delta SASA_{I\text{-}D}$ = −2,330 Å^2). Because the calculated SASAs for native EGFP and EGFP$^{G4\Delta}$ are very similar, it appears that both the intermediate and the denatured forms of EGFP$^{G4\Delta}$ retain a greater degree of residue burial, implying a more compact structure compared to the same states for EGFP.

There were significant differences in the refolding kinetics of EGFP$^{G4\Delta}$ compared to EGFP. Importantly, up to 100% of the fluorescence signal observed before denaturation was recovered (Figure 4B), suggesting that a near full population of the protein molecules regained their native structure, comparable with the extensively mutated superfolder GFP (Andrews et al., 2007; Pédelacq et al., 2006). In comparison, 77% of EGFP fluorescence was recovered, consistent with previous observations for related "improved" GFP folding variants (Fukuda et al., 2000; Steiner et al., 2008). In both cases, refolding fit best to a double exponential with an initial fast phase (k_{fast}) followed by a slow phase (k_{slow}). The fast refolding phase for EGFP$^{G4\Delta}$ was marginally slower than EGFP but the slow refolding phase was marginally faster (Table 1). *Cis/trans* isomerization has been shown to be a rate-limiting step in protein folding (Wedemeyer et al., 2002) and is thought to be the reason for the slow refolding phase in GFP (Steiner et al., 2008). Therefore, the barrier to *cis/trans* isomerization appears to be slightly smaller for EGFP$^{G4\Delta}$.

Structural Impact of G4Δ

The crystal structure of EGFP$^{G4\Delta}$ was determined to 1.5 Å resolution (Table 3 for statistics table) and compared to the recently determined high-resolution structure of EGFP (Arpino et al., 2012b; Royant and Noirclerc-Savoye, 2011). Size exclusion chromatography confirmed that like EGFP, EGFP$^{G4\Delta}$ is monomeric (Figure S6). The overall structures of EGFP and EGFP$^{G4\Delta}$ are very similar with backbone and all atom root-mean-square deviations of 0.6 Å and 1.2 Å, respectively (Figure S7), suggesting that deletion of G4 is having a subtle effect on structure, predominantly at the local level. E222, a critical residue in determining the protonated state of the chromophore phenol moiety (thus the spectral characteristics) (Tsien, 1998) and chromophore maturation (Sniegowski et al., 2005), exists in one of two distinct conformations in EGFP (Arpino et al., 2012b; Royant and Noirclerc-Savoye, 2011). However, the electron density for E222 in EGFP$^{G4\Delta}$ was best satisfied when modeled as a single distinct conformer equivalent to the major form in EGFP (Figure S8).

In EGFP, G4 is located in the first organized structural element, a 3_{10} helix, which is relatively distant from the chromophore (Figure 5A and Figure S7). Deletion of G4 results in a significant local

Table 1. Equilibrium Unfolding, Unfolding, and Refolding Kinetic Parameters

Variant	$D_{N\text{-}I}$ (M)[a]	$m_{N\text{-}I}$ (kcal mol^{-1} M^{-1})[b]	$D_{I\text{-}D}$ (M)[c]	$m_{I\text{-}D}$ (kcal mol^{-1} M^{-1})[b]	$\Delta G_{N\text{-}I}$ (kcal mol^{-1})[d]	$\Delta G_{I\text{-}D}$	$\Delta G_{N\text{-}D}$	k_U (min^{-1})[e]	k_{fas} (10^{-2} s^{-1})[f]	k_{slow} (10^{-2} s^{-1})[f]
EGFP	2.04	2.55 ± 0.23	2.79	3.73 ± 0.27	5.18	10.37	15.55	2.09 ± 0.00	4.32 ± 0.08	1.00 ± 0.02
EGFP$^{G4\Delta}$	1.76	1.96 ± 0.1	2.43	2.89 ± 0.14	3.46	7.01	10.47	2.15 ± 0.01	3.78 ± 0.25	1.15 ± 0.05

[a]Concentration of GdmCl at which 50% of the protein sample is in the native and intermediate state.
[b]Measure of dependence of ΔG on denaturant concentration, m value. Error bars represent 1 SD from three replicates.
[c]Concentration of GdmCl at which 50% of the protein sample is in the intermediate and denatured state.
[d]Change in free energy for native to intermediate (N-I), intermediate to denatured (I-D), and native to denatured (N-D) transitions.
[e]Rate constant from single exponential fit of unfolding progress curves (data not shown). Error bars represent 1 SD from three replicates.
[f]Rate constants from two exponential fits of refolding progress curves (Figure 2B). Error bars represent 1 SD from three replicates.

rearrangement of residues in the 3_{10} helix and the helix itself (Figure 5B). K3 rotates by ~120° around the axis of the 3_{10} helix to reside at the position previously occupied by G4. Modeling the K3 side chain to two conformations that differ slightly (root-mean-square deviation 0.66 Å for A versus B) best satisfied the electron density during structure refinement (Figure S9). The side chain positioning of E5 and E6 are also significantly perturbed, but register is restored from L7 and F8 onward (Figure 5B).

The shift of K3 positioning in EGFP$^{G4\Delta}$ appears to have two main effects on the local structure. First, rotation brings the K3 side chain into the vicinity of the *cis* peptide bond between M88 and P89 with the N^Z group being within hydrogen bonding distance of the backbone carbonyl O of M88 (Figures 4D and 5C). This in turn could have implications in stabilization of the *cis* peptide bond. Second, the shift in the 3_{10} helix to accommodate the K3 side chain repositions E5 and generates a new polar network (Figures 5C and 5D). The carboxylate group of E5 is now within electrostatic bonding distance (2.7 Å) of the N^Z amine group of K79 in the adjacent 3_{10} helix, which in turn is within hydrogen bond distance of the backbone carbonyl group of Y74 (Figure 5D). This new polar bond network absent from EGFP now links different secondary structure elements in EGFP$^{G4\Delta}$.

There are also changes to the arrangement of structured water molecules (Figures 5C and 5D). E5 side chain rearrangement appears to result in the displacement of a water molecule normally observed in EGFP (W_{528} in Figure 5C). The position of a second water molecule (W_{492} in EGFP and W_{616} in EGFP$^{G4\Delta}$) is similar between the two but is now capable of forming a hydrogen bond with the E5 side chain in EGFP$^{G4\Delta}$.

DISCUSSION

Predicting the effects of backbone mutations such as single amino acid deletions and the implications in terms of protein structure is currently very difficult and generally avoided as part of the protein design process. This is because not only do deletion mutations alter side-chain placement of adjacent residues, but also locally confine structure. Using directed-evolution approaches, a survey of tolerated deletion mutations can be conducted, which in turn can unearth mutations not obvious on initial inspection that enhance certain properties of a protein in unpredicted ways. InDel events involving GFP, whether they are small alterations such as those here or more drastic events such as insertions of whole protein domains (Arpino et al., 2012a; Baird et al., 1999; Biondi et al., 1998; Doi and Yanagawa, 1999) can lead to new and/or improved functionality so should not be ignored as useful routes to protein engineering. This in turn provides us with mechanistic insights concerning how InDel mutations exert their effect on protein structure during the natural evolutionary process and allow feedback to the protein design process.

Fluorescent proteins represent an important and current target for protein engineering (Pakhomov and Martynov, 2008; Tsien, 1998). Only limited deletion mutations targeted at the termini (Dopf and Horiagon, 1996; Flores-Ramírez et al., 2007; Li et al., 1997) and selected loops (Flores-Ramírez et al., 2007; Li et al., 1997) has been performed on proteins related to GFP. This study is more comprehensive in terms of assessing tolerance and impact throughout the protein. Of the 87 unique deletion events observed, 42 (48%) were tolerated. This is in contrast to previous amino acid deletion studies, which suggest GFP to be largely intolerant to amino acid deletions (Flores-Ramírez et al., 2007; Li et al., 1997). Indeed, such a high frequency of tolerance suggests protein structure is suitably plastic to incorporate backbone alternations without complete loss of stability and function. Recent bioinformatics studies of proteomes suggests that InDel mutations, especially deletion mutations, are major instigators of leaps in the fitness landscape of a protein but largely require local substitution mutations to elicit an effect (Leushkin et al., 2012; Tóth-Petróczy and Tawfik, 2013); this does not appear to always be the case in the more directed approach taken here where single deletions can not only be tolerated, but also be beneficial when incorporated alone (vide infra).

Analysis of the tolerated and nontolerated amino acid deletion positions by mapping to the secondary structure topology and tertiary structure of EGFP (Figure 1 and Figure S1) showed a clear divide between regions tolerant and nontolerant 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 (Pascarella and Argos, 1992). However, helical segments appear more tolerant to deletion than strands (Figures 1B and 1C). Deletion within a strand may cause registry shifts and given EGFP fluorescence is reliant on its tertiary structure, will have obvious detrimental effect on function, the primary screening property. Termini of strands appear to be more tolerant to residue removal, with 21% of tolerated deletions in these regions (Figure 1B). This is in line with previous observations with TEM β-lactamase (Simm et al., 2007). As observed previously with TEM β-lactamase (Simm et al., 2007), helical structures appear more resilient to deletions (Figures 1B

Table 2. Solvent-Accessible Surface Area Changes on Unfolding

Protein	Native SASA (Å^2)[a]	Fully Unfolded SASA (Å^2)[a]	$\Delta SASA_{N-I}$ (Å^2)[b]	$\Delta SASA_{I-D}$ (Å^2)[b]	$\Delta SASA_{N-D}$ (Å^2)[b]
EGFP	9,919	31,849	7,050 ± 340	10,320 ± 440	17,370 ± 390
EGFP$^{G4\Delta}$	9,366	31,953	5,430 ± 300	7,990 ± 370	13,420 ± 350

[a]Calculated using the calc-surface program (http://helixweb.nih.gov/structbio/basic.html). Only residues K3–L231 were considered for SASA calculations. Unfolded state refers to fully unfolded peptide.
[b]Calculated with the determined m values using BPPred (http://www-clarke.ch.cam.ac.uk/BPPred.php; Geierhaas et al., 2007).

and 1C). In EGFP, H1 and H3 were particularly tolerant. Whereas the general tolerance of H1 to deletion may not be entirely surprising given that removal of the first five amino acids can be tolerated to a degree (Li et al., 1997), at first glance the beneficial effects are unexpected: one of the enhanced variants, EGFP$^{G4\Delta}$, has residue from H1 removed (vide infra). Thus, deletions of residues within helices may not be as disruptive as within strands, which may be a consequence of the "stand-alone" nature of helices compared to a strand that forms one element of a β sheet system. This is especially pertinent in the case of EGFP whereby strands form a critical structural feature of the protein (the β barrel) with the two faces of the strands differing markedly in their chemical composition. Structural analysis revealed that helices soon regain residue register (Figure 5B). The relative surface burial of a residue may not be such a critical factor in defining tolerance because deletion of residues with a wide range of SASA values was allowed. However, residues with low solvent exposure have a higher propensity to be nontolerant to deletion (Figure 2). For example, the buried central core helix was not tolerant but it is unknown whether these deletions proved disruptive to the β barrel structure as a whole or to chromophore maturation (and thus fluorescence) because two deletion mutations removed chromophore-forming residues (Figure S1 and Table S2).

Whereas loops are the most tolerant structure to deletions in terms of EGFP, many of these deletion mutations where observed in short loops (five residues or less) that could be considered as turns between organized secondary structure elements. Most notable were short loops connecting S2-S3, S3-H2, S7-S8, and S8-S9 (Figure 1B). Shortening of these already constrained loops could be considered deleterious, but this does not appear to be the case. This is backed up to an extent by the few observed deletion mutations in short loops of nonfluorescent variants (Figure S1), suggesting they are more tolerant than would be expected. These short turn loops in EGFP are also tolerant to larger InDel events such as domain insertion to generate new protein scaffolds with coupled activities (Arpino et al., 2012a).

The context of the deletion appears to be a more important determinant than the secondary structure it occupies. This is illustrated by H1, a 3_{10} helix largely tolerant to deletions. However, only deletion of G4 results in the local structural rearrangement resulting in improved cellular fluorescence through potential stabilization of the *cis* M88-P89 peptide bond (Figure 5). Deletion of the adjacent E5 and E6 residues has little effect on fluorescence phenotype (Table S1) and loss of K3 with the G4S mutation renders the protein nonfluorescent (Table S2). Even replacement of K3 with N removes the beneficial properties of G4Δ (Table S1 and Figure S10), probably by eliminating the favorable interaction the K3 amine group makes with the *cis* M88-P89 peptide bond (vide infra). Key to the beneficial mutational mechanism is the side-chain "flipping" or registry shift (Figure 5B), an event considered to be deleterious, in making new long-range interactions.

Deletion of G4 promotes increased production of functional protein in the cell. Rather than affecting brightness or apparent thermodynamic stability of EGFP, the influence of G4Δ is likely to be manifested through changes in/optimization of the folding process and avoiding potential off-pathway aggregation. Apparent thermodynamic stability of EGFP$^{G4\Delta}$ is curiously lower than that of EGFP (Figure 4A and Table 1), but functional recovery after denaturation is higher (Figure 4B). Unlike many stabilized GFP derived variants containing multiple substitutions (e.g., GFPmut2, Cannone et al., 2005; and cycle 3 GFP, Fukuda et al., 2000 variants), up to 100% of EGFP$^{G4\Delta}$ refolds to a functional state after denaturation (Figure 4B). Given the importance of the folding process to proteins in general, deletion mutations need not be considered harmful and thus be used generally as a mechanism to improve proteins. This has been demonstrated to a degree here by transplanting the G4Δ mutation to EYFP and increasing observed cellular fluorescence (Figure 3B).

The unfolding and refolding properties for EGFP presented here are in good agreement with previous work that uses the *p*-hydroxybenzylidene-imidazolinone (HBI) chromophore as a probe to monitor (un)folding (Hsu et al., 2009; Steiner et al., 2008; Stepanenko et al., 2004). Both EGFP and EGFP$^{G4\Delta}$ equilibrium unfolding fitted to a three-state model, suggesting that folding of the mature protein occurs via an intermediate, which has been observed previously for other related GFPs (Andrews et al., 2008, 2009; Hsu et al., 2009). Although EGFP$^{G4\Delta}$ is less stable by an apparent $\Delta\Delta G_{N-D}$ of 5.1 kcal/mol due predominantly to the change in m value for both transitions (native to intermediate and intermediate to denatured), the m values themselves suggest a change in the folding process itself especially regarding the degree of accessible surface area (Myers et al., 1995). The predicted changes in SASA for EGFP$^{G4\Delta}$ on unfolding are lower than those for EGFP, suggesting that the deletion variant retains a more compact structure in the intermediate and denatured forms. The intermediate state is already considered highly structured in EGFP (Table 1 and Table S3) and in other engineered GFPs (Andrews et al., 2007, 2008, 2009; Huang et al., 2007; Xie and Zhou, 2008), with deletion of G4 potentially increasing it. The apparent increased residue burial/structure in the intermediate and denatured state on introduction of G4Δ may be a function of the long-range polar interactions observed for mature, native EGFP$^{G4\Delta}$ (Figure 5D). The putative hydrogen bond via the K3 amine with the carbonyl group of the M88-P89 peptide bond may stabilize the sole *cis* proline peptide bond

Table 3. Crystallographic Statistics

Data Reduction Statistics	EGFP$^{G4\Delta}$
Beamline	I03
Wavelength (Å)	0.97630
Space group	C121
a, b, c (Å)	91.9, 66.7, 45.3, β = 108.76°
Resolution range (Å)	43.49–1.58
Total reflections measured	125,853
Unique reflections	34,564
Completeness (%) (last shell)	97.5 (98.2)
I/σ (last shell)	8.9 (2.1)
R(merge)[a] (%) (last shell)	8.4 (53.8)
B(iso) from Wilson (Å^2)	12.1
Refinement Statistics	
Protein atoms excluding H	1,987
Solvent molecules	270
R-factor[b] (%)	17.3
R-free[c] (%)	21.10
Rmsd bond lengths (Å)	0.024
Ramachandran Plot Statistics	
Rmsd angles (°)	2.3
Core region (%)	96.7
Allowed region (%)	3.3
Additionally allowed region (%)	0
Disallowed region (%)	0
PDB code	4KA9

Rmsd, root-mean-square deviation.

[a] $R_{merge} = \sum_h \sum_j (I_{hj} - \langle I_j \rangle) / \sum_h I_{hj}$.

[b] $R_{factor} = (\sum_h |F_{h,obs} - F_{h,calc}|) / (\sum_h F_{h,obs})$.

[c] R_{free} is calculated from a set of 5% randomly selected reflections that were excluded from refinement.

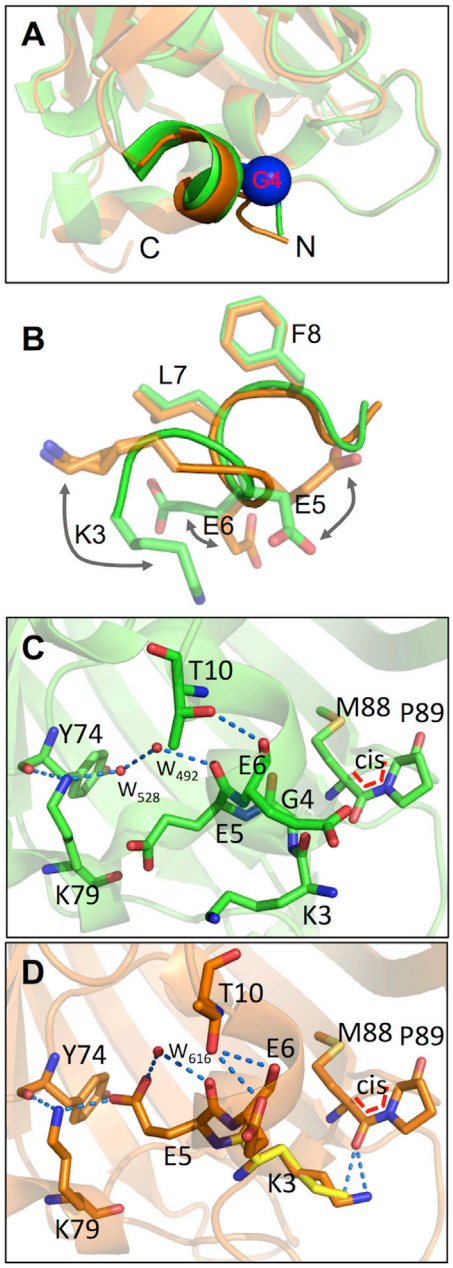


Figure 5. Structural Effects of the G4Δ Mutation on EGFP

(A) Overlap of EGFP (green) and EGFP$^{G4\Delta}$ (orange) with the G4 residue in EGFP highlighted as a blue sphere.

(B) The effect of G4 deletion on the side chain positioning of adjacent residues in the N-terminal 3_{10} helix.

(C and D) Local structure of EGFP centered around G4 (C) and the rearrangements on deletion of G4 in EGFP$^{G4\Delta}$ (D) The alternative conformations for K3 are shown, with the second side chain conformer shown as yellow. The alternate conformations for K3 are highlighted in more detail in Figure S8.

and promote conversion to the native state. Stabilization of the *cis* proline peptide bond has implications in terms of backbone trajectory flux during folding that assists transition to the native state. *Cis-trans* isomerization around the M88-P89 peptide bond is known to be important to the folding process (Enoki et al., 2004) and is considered to be the rate-limiting step in folding (Hsu et al., 2009). *Cis-trans* isomerization around a X-Pro peptide bond is in general one of the rate-determining steps in protein folding and can deviate a protein away from its native folding trajectory to a kinetically trapped aggregative state if the incorrect forms persist (Steiner et al., 2008; Stepanenko et al., 2004). If stabilization of the M88-P89 *cis* peptide bond is perpetuated in the intermediate (and even to a degree in the denatured) state, it may explain the decreased SASA inferred from the m values for EGFP$^{G4\Delta}$ due to a higher degree of residual and/or transient structure. Both G4Δ and the M88-P89 *cis* proline are located in regions that contribute toward the structured element of the equilibrium folding intermediate (Huang et al., 2007; Reddy et al., 2012).

The G4Δ mutation only slightly sped up the slow phase of EGFP folding, but must be put in context of an increase in folding efficiency (Figure 3B). Recent folding simulations of a related GFP, citrine, revealed that misplacement of a loop connecting strands S9 and S10 resulted in the protein becoming stuck in a misfolded kinetic trap (Reddy et al., 2012). This loop lies adjacent to the *cis* M88-P89 peptide bond, which is in turn stabilized through interactions with the repositioned K3 on G4 deletion.

Furthermore, residues in sfGFP forming the lid of the β barrel (including K3 and G4), an element that plays a key role in driving the intermediate to the fully folded fluorescent protein, display conformation heterogeneity in the native state (Andrews et al., 2009). Thus, the role of G4Δ may lie in optimization of the folding efficiency and mechanism.

The shift in register of H1 on deletion of G4 has additional effects in conjunction with repositioning K3 through the formation of an extended polar interaction network (Figure 5). Accompanying the changes involving K3 is the repositioning of E5 resulting in direct interaction with K79 on an adjacent secondary structure element. The generation of such networks has been observed before in fast and efficient folding versions of GFP (Pédelacq et al., 2006).

The one curious and counterintuitive observation is the significant decrease in apparent thermodynamic stability, especially given the generally stabilizing interactions formed on deleting G4. The increased structure in the intermediate and denatured states relative to the native state as implied by the m values may have an impact on the apparent thermodynamic stability of mature EGFP$^{G4\Delta}$ through changes in absolute free energy levels. Thus, the role of G4Δ may lie in optimization of the refolding efficiency and mechanism at the expense of an improved $\Delta G_{N\text{-}D}$. However, it cannot be discounted that the HBI chromophore used to probe folding is more sensitive to changes in EGFP$^{G4\Delta}$ structure, which may be affecting values related to stability such as $[GdmCl]_{50\%}$. Although there is generally good correlation between different probes to monitor GFP unfolding, there can be slight discrepancies in the $[GdmCl]_{50\%}$ value (Huang et al., 2007). The G4Δ mutation may have a greater impact on the de novo folding of the nascent GFP prior to chromophore formation than the refolding/unfolding of the mature protein. Folding of GFP is known to be dependent on the chromophore and has been suggested as the cause for the ''hysteresis'' folding phenomenon observed for mature GFP (Hsu et al., 2009) due to a ''dual basin'' folding landscape comprising a native-like intermediate and the ''locked'' native, fluorescent state (Andrews et al., 2008). Proline isomerization, especially with regards to P89, together with formation of the β barrel lid (vide supra) is thought to be an important contributor to the barrier between the two states. While de novo folding has not been explored here, it may explain why G4Δ results in higher cellular fluorescence in situ (Figure 3). Chromophore maturation occurs after folding and depends on formation of a correctly folded protein (Reid and Flynn, 1997). Furthermore, the N-terminal region is thought to facilitate the folding process of the nascent polypeptide on release from the ribosome (Uemura et al., 2008). Therefore, G4Δ may be having a more significant impact on folding immediately after release from the ribosome in the complex mixture of the cell where off-pathway aggregative folding events are more likely.

Contrary to current dogma, deletion of a single amino acid is generally well tolerated throughout a protein, including helical elements, and can be beneficial as exemplified by the G4Δ mutation. The thought that amino acid deletions hinder protein folding conflicts with observations here as the basis for improved cellular fluorescence imparted by EGFP$^{G4\Delta}$ is thought to be due to changes in the folding process. The residue repositioning through a helical registry shift is critical to generating a new interaction network and is unlikely to have risen through substitution mutations alone, highlighting the ability of deletion mutations to sample structural space not accessible through exchanging only side chains. Indeed, the influence of deletion mutations can go beyond structural stability and also influence functional properties (Fujii et al., 2006; Simm et al., 2007). The identification of G4, which is so close to the start of the structured region of EGFP (Arpino et al., 2012b) highlights the sometimes nonintuitive nature of useful deletion mutations and bestows the benefits of a directed-evolution approach. Together with work presented here, the recent idea of InDel mutations instigating major leaps in the protein fitness landscape during evolution, with compensating (or enabling) substitution mutations (mostly local to the InDel event) helping to improve overall fitness (Leushkin et al., 2012; Tóth-Petróczy and Tawfik, 2013), potentially provides a template for future protein engineering strategies.

EXPERIMENTAL PROCEDURES

TND Library Construction

Insertion of the engineered transposon MuDel (Jones, 2005) into the *egfp* gene encoding EGFP residing within the pNOM-XP3 plasmid was performed using an in vitro transposition and selection procedure described elsewhere (Baldwin et al., 2009) to generate the library *egfp*Δ^{2504}. This is described in more detail in the Supplemental Experimental Procedures.

Protein Production and Purification

The production and subsequent purification of EGFP and EGFP$^{G4\Delta}$ was performed essentially as described elsewhere (Arpino et al., 2012b). A detailed procedure is provided in the Supplemental Experimental Procedures. The production of EGFP and EGFP$^{G4\Delta}$ for whole cell fluorescence analysis was performed as follows. Luria-Bertani (LB) broth (20 ml) supplemented with 100 μg/ml ampicillin and 1 mM isopropyl-beta-D-thiogalactopyranoside (IPTG) was inoculated with a single *E. coli* BL21-Gold (DE3) colony containing a relevant plasmid (pNOM-XP3 containing the *egfp* gene or *egfp*$^{G4\Delta}$ genes) and incubated overnight at either 25°C or 37°C. The production of EGFP, EGFP$^{G4\Delta}$, EYFP, and EYFP$^{G4\Delta}$ in colonies streaked out on LB agar plates was performed as follows. A single BL21-Gold (DE3) colony containing the relevant plasmid (pNOM-XP3 containing the *egfp, egfp*$^{G4\Delta}$, *eyfp*, or *eyfp*$^{G4\Delta}$ genes) was resuspended in LB broth (200 μl) supplemented with 100 μg/ml ampicillin and incubated at 37°C shaking (200 revolutions per minute) for 2 hr. The cultures were streaked out onto LB agar plates supplemented with 100 μg/ml ampicillin and 150 μM IPTG. The plates were incubated overnight at 37°C and depicted with a transilluminator.

Fluorescence Spectroscopy

Excitation and emission spectra were measured using a Cary Eclipse fluorescence spectrophotometer (Varian) in a cuvette of dimensions 5 × 5 mm with a 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. Whole cell fluorescence spectroscopy was performed on *E. coli* BL21-Gold (DE3) cell cultures after expression of EGFP and EGFP$^{G4\Delta}$ at either 25°C or 37°C. Expression cultures were harvested by centrifugation (1,500 × *g* for 10 min) 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 (TNG Buffer) to an optical density 600 of 0.1 in a 1 cm path length cuvette. The resuspended cells were transferred to a cuvette with 5 × 5 mm dimensions and excitation and emission spectra measured as described above. Fluorescence measurements using purified protein samples were performed in 50 mM Tris-HCl, pH 8.0 at 25°C, and 150 mM NaCl. The calculation of quantum yield and fluorescence lifetimes were performed as described elsewhere (Arpino et al., 2012a, 2012b).

Guanidine Hydrochloride Equilibrium Unfolding

Purified protein (1 μM) was prepared in TNG Buffer and guanidine hydrochloride (0–6 M GdmCl) and incubated for up to 250 hr at 37°C. Protein unfolding

was monitored by fluorescence at 520 nm after excitation at 480 nm using a FLUOstar Omega microplate reader (BMG Labtech). To estimate the apparent [GdmCl] at which 50% of the protein is folded and 50% of the protein is unfolded, samples were incubated in 96-well plates at 37°C and measured after 2.5, 5, 22, 48, 120, and 250 hr.

Equilibrium unfolding data measured from samples incubated at 37°C in Eppendorf tubes for 250 hr were fit to a three-state model (Equation 1) formulated using the following equations:

$$K_{N-I} = exp^{\left(\frac{m_{N-I}[D]-\Delta G_{N-I}}{RT}\right)},\ K_{I-D} = exp^{\left(\frac{m_{I-D}[D]-\Delta G_{I-D}}{RT}\right)}$$

$$Fr_N = \frac{1}{1+K_{N-I}+K_{N-I}K_{I-D}},\ Fr_I = \frac{K_{N-I}}{1+K_{N-I}+K_{N-I}K_{I-D}},$$
$$Fr_D = \frac{K_{N-I}K_{I-D}}{1+K_{N-I}+K_{N-I}K_{I-D}}$$

$$F = Y_N + Fr_I(Y_I - Y_N) + Fr_D(Y_D - Y_N), \qquad \text{(Equation 1)}$$

where $\Delta G_{N\text{-}I}$ is the difference in Gibbs' free energy between native and intermediate states and $\Delta G_{I\text{-}D}$ is the difference between intermediate and denatured states. $m_{N\text{-}I}$ is a constant that describes the dependence of ΔG on denaturant concentration, [D], between the native and intermediate states, whereas $m_{I\text{-}D}$ is the same for the intermediate to denatured states. Fr_N, Fr_I, and Fr_D are fractions of the partition function in a three-energy-state system, and the plot of fractional populations of different states against denaturant concentration can be generated from these equations.

Protein Refolding Kinetics

Protein samples were unfolded by dilution (1/100) to a final concentration of 1 μM in 6.4 M GdmCl and refolded by rapid dilution (1/10) into fresh TNG Buffer supplemented with 1 mM dithiothreitol so that the final protein concentration was 100 nM and GdmCl was 0.64 M. In both cases, fluorescence was monitored for 20 min at 510 nm after excitation at 488 nm in a 5 × 5 mm dimension cuvette with an excitation and emission band pass of 2.5 and 5 nm, respectively. Unfolding data were fit with a single exponential decay (Equation 2) and refolding data fit with a double exponential (Equation 3):

$$Y = (Y_0 - P) \times e^{-kt} + P \qquad \text{(Equation 2)}$$

$$Y = Y_0 + \left(F_1 \times \left(1 - e^{-k_{fast}t}\right)\right) + \left(F_2 \times \left(1 - e^{-k_{slow}t}\right)\right), \qquad \text{(Equation 3)}$$

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_{fast}; and F_2 is the proportional value for the second rate constant, k_{slow}.

Protein Crystallization and Structure Determination

Purified EGFP$^{G4\Delta}$ (15 mg/ml in 50 mM Tris-HCl, pH 8.0, and 150 mM NaCl) was screened for crystal formation by the sitting drop vapor diffusion method with incubation at 18°C. Drops were set up with equal volumes of protein and precipitant solutions (0.5 μl each). Crystals of EGFP$^{G4\Delta}$ were obtained from 0.1 M HEPES, pH 7.0, 0.01 M $ZnCl_2$, and 20% (w/v) PEG 6000. Data were collected on beamline I03 at the Diamond Light Source, Harwell, UK. Usable diffraction was recorded up to a resolution of 1.58 Å. Data were reduced with the XIA2 package (Winter, 2009), space group assignment was done by POINTLESS (Evans, 2006), and scaling and merging were completed with SCALA (Evans, 2006) and TRUNCATE (CCP4, 1994). Initial molecular replacement for the EGFP$^{G4\Delta}$ variant structure was performed using a previously determined EGFP structure (Protein Data Bank [PDB] entry 4EUL; Arpino et al., 2012a) as the search model, using PHASER (McCoy et al., 2007). The structure for EGFP$^{G4\Delta}$ was adjusted manually using COOT (Emsley and Cowtan, 2004) and refinement of the completed molecule was carried out using REFMAC (Murshudov et al., 1997). Protein atoms were refined isotropically and anisotropically. All nonprotein atoms were refined isotropically. The above routines were used within the CCP4 package (CCP4, 1994; http://www.ccp4.ac.uk). Graphical representations were made with PyMOL Molecular Graphics System (Schrödinger).

ACCESSION NUMBERS

The PDB accession number for the EGFP$^{G4\Delta}$ structure is 4KA9.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, ten figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.str.2014.03.014.

ACKNOWLEDGMENTS

This work was supported by BBSRC grants BB/E001084 and BB/FOF/263 and a Cardiff Partnership Award to D.D.J. J.A.J.A. was supported by a BBSRC CASE studentship in collaboration with Merck KGaA. L.M.H. is supported by a KESS studentship in partnership with 3M. The authors thank the staff at the Diamond Light Source for the supply of facilities and beam time, especially Beamline I03 and I04 staff. We thank Dr. Roger Chittock for advice on fluorescence lifetime measurements and Nadiatul Zulkifli and Hua Kang for technical support.

Received: November 28, 2013
Revised: March 8, 2014
Accepted: March 10, 2014
Published: May 22, 2014

REFERENCES

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.

Andrews, B.T., Gosavi, S., Finke, J.M., Onuchic, J.N., and Jennings, P.A. (2008). The dual-basin landscape in GFP folding. Proc. Natl. Acad. Sci. USA *105*, 12283–12288.

Andrews, B.T., Roy, M., and Jennings, P.A. (2009). Chromophore packing leads to hysteresis in GFP. J. Mol. Biol. *392*, 218–227.

Arpino, J.A., Czapinska, H., Piasecka, A., Edwards, W.R., Barker, P., Gajda, M.J., Bochtler, M., and Jones, D.D. (2012a). Structural basis for efficient chromophore communication and energy transfer in a constructed didomain protein scaffold. J. Am. Chem. Soc. *134*, 13632–13640.

Arpino, J.A., Rizkallah, P.J., and Jones, D.D. (2012b). Crystal structure of enhanced green fluorescent protein to 1.35 Å resolution reveals alternative conformations for Glu222. PLoS ONE *7*, e47132.

Baird, G.S., Zacharias, D.A., and Tsien, R.Y. (1999). Circular permutation and receptor insertion within green fluorescent proteins. Proc. Natl. Acad. Sci. USA *96*, 11241–11246.

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.

Bershtein, S., and Tawfik, D.S. (2008). Advances in laboratory evolution of enzymes. Curr. Opin. Chem. Biol. *12*, 151–158.

Biondi, R.M., Baehler, P.J., Reymond, C.D., and Véron, M. (1998). Random insertion of GFP into the cAMP-dependent protein kinase regulatory subunit from Dictyostelium discoideum. Nucleic Acids Res. *26*, 4946–4952.

Cannone, F., Bologna, S., Campanini, B., Diaspro, A., Bettati, S., Mozzarelli, A., and Chirico, G. (2005). Tracking unfolding and refolding of single GFPmut2 molecules. Biophys. J. *89*, 2033–2045.

CCP4; Collaborative Computational Project, Number 4 (1994). The CCP4 suite: programs for protein crystallography. Acta Crystallogr. D Biol. Crystallogr. *50*, 760–763.

de Jong, W.W., and Rydén, L. (1981). Causes of more frequent deletions than insertions in mutations and protein evolution. Nature *290*, 157–159.

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.

Doi, N., and Yanagawa, H. (1999). Insertional gene fusion technology. FEBS Lett. *457*, 1–4.

Dopf, J., and Horiagon, T.M. (1996). Deletion mapping of the Aequorea victoria green fluorescent protein. Gene *173* (1 Spec No), 39–44.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132.

Enoki, S., Saeki, K., Maki, K., and Kuwajima, K. (2004). Acid denaturation and refolding of green fluorescent protein. Biochemistry *43*, 14238–14248.

Evans, P. (2006). Scaling and assessment of data quality. Acta Crystallogr. D Biol. Crystallogr. *62*, 72–82.

Flores-Ramírez, G., Rivera, M., Morales-Pablos, A., Osuna, J., Soberón, X., and Gaytán, P. (2007). The effect of amino acid deletions and substitutions in the longest loop of GFP. BMC Chem. Biol. *7*, 1.

Fujii, R., Kitaoka, M., and Hayashi, K. (2006). RAISE: a simple and novel method of generating random insertion and deletion mutations. Nucleic Acids Res. *34*, e30.

Fukuda, H., Arai, M., and Kuwajima, K. (2000). Folding of green fluorescent protein and the cycle3 mutant. Biochemistry *39*, 12025–12032.

Geierhaas, C.D., Nickson, A.A., Lindorff-Larsen, K., Clarke, J., and Vendruscolo, M. (2007). BPPred: a Web-based computational tool for predicting biophysical parameters of proteins. Protein Sci. *16*, 125–134.

Heinz, D.W., Baase, W.A., Dahlquist, F.W., and Matthews, B.W. (1993). How amino-acid insertions are allowed in an alpha-helix of T4 lysozyme. Nature *361*, 561–564.

Hsu, S.T., Blaser, G., and Jackson, S.E. (2009). The folding, stability and conformational dynamics of beta-barrel fluorescent proteins. Chem. Soc. Rev. *38*, 2951–2965.

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.

Imamichi, T., Murphy, M.A., Imamichi, H., and Lane, H.C. (2001). Amino acid deletion at codon 67 and Thr-to-Gly change at codon 69 of human immunodeficiency virus type 1 reverse transcriptase confer novel drug resistance profiles. J. Virol. *75*, 3988–3992.

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.

Leushkin, E.V., Bazykin, G.A., and Kondrashov, A.S. (2012). Insertions and deletions trigger adaptive walks in Drosophila proteins. Proc. Biol. Sci. *279*, 3075–3082.

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.

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.

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.

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.

Myers, J.K., Pace, C.N., and Scholtz, J.M. (1995). Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. Protein Sci. *4*, 2138–2148.

O'Neil, K.T., Bach, A.C., 2nd, and DeGrado, W.F. (2000). Structural consequences of an amino acid deletion in the B1 domain of protein G. Proteins *41*, 323–333.

Pakhomov, A.A., and Martynov, V.I. (2008). GFP family: structural insights into spectral tuning. Chem. Biol. *15*, 755–764.

Pascarella, S., and Argos, P. (1992). Analysis of insertions/deletions in protein structures. J. Mol. Biol. *224*, 461–471.

Patzoldt, W.L., Hager, A.G., McCormick, J.S., and Tranel, P.J. (2006). A codon deletion confers resistance to herbicides inhibiting protoporphyrinogen oxidase. Proc. Natl. Acad. Sci. USA *103*, 12329–12334.

Pédelacq, 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.

Reddy, G., Liu, Z., and Thirumalai, D. (2012). Denaturant-dependent folding of GFP. Proc. Natl. Acad. Sci. USA *109*, 17832–17838.

Reid, B.G., and Flynn, G.C. (1997). Chromophore formation in green fluorescent protein. Biochemistry *36*, 6786–6791.

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.

Shortle, D., and Sondek, J. (1995). The emerging role of insertions and deletions in protein engineering. Curr. Opin. Biotechnol. *6*, 387–393.

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.

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.

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.

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.

Stott, K.M., Yusof, A.M., Perham, R.N., and Jones, D.D. (2009). A surface loop directs conformational switching of a lipoyl domain between a folded and a novel misfolded structure. Structure *17*, 1117–1127.

Taylor, M.S., Ponting, C.P., and Copley, R.R. (2004). Occurrence and consequences of coding sequence insertions and deletions in Mamm. Genomes. Genome Res *14*, 555–566.

Tóth-Petróczy, A., and Tawfik, D.S. (2013). Protein insertions and deletions enabled by neutral roaming in sequence space. Mol. Biol. Evol. *30*, 761–771.

Tsien, R.Y. (1998). The green fluorescent protein. Annu. Rev. Biochem. *67*, 509–544.

Uemura, S., Iizuka, R., Ueno, T., Shimizu, Y., Taguchi, H., Ueda, T., Puglisi, J.D., and Funatsu, T. (2008). Single-molecule imaging of full protein synthesis by immobilized ribosomes. Nucleic Acids Res. *36*, e70.

Vetter, I.R., Baase, W.A., Heinz, D.W., Xiong, J.P., Snow, S., and Matthews, B.W. (1996). Protein structural plasticity exemplified by insertion and deletion mutants in T4 lysozyme. Protein Sci. *5*, 2399–2415.

Wedemeyer, W.J., Welker, E., and Scheraga, H.A. (2002). Proline cis-trans isomerization and protein folding. Biochemistry *41*, 14637–14644.

Winter, G. (2009). xia2: an expert system for macromolecular crystallography data reduction. J. Appl. Cryst. *43*, 196–190.

Wood, N., Bhattacharya, T., Keele, B.F., Giorgi, E., Liu, M., Gaschen, B., Daniels, M., Ferrari, G., Haynes, B.F., McMichael, A., et al. (2009). HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. PLoS Pathog. *5*, e1000414.

Xie, J.B., and Zhou, J.M. (2008). Trigger factor assisted folding of green fluorescent protein. Biochemistry *47*, 348–357.

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.

Structure, Volume *22*

Supplemental Information

Random Single Amino Acid Deletion Sampling Unveils Structural Tolerance and the Benefits of Helical Registry Shift on GFP Folding and Structure

James A. J. Arpino, Samuel C. Reddington, Lisa M. Halliwell, Pierre J. Rizkallah, and D. Dafydd Jones

Random single amino acid deletion sampling unveils structural tolerance and the benefits of helical registry shift on GFP folding and structure.

James A. J. Arpino[1], Sam C. Reddington[1], Lisa M. Halliwell[1], Pierre J. Rizkallah[2] & D. Dafydd Jones.[1]

Supporting Information.

Supporting Methods.

EGFP TND library construction.

Insertion of the engineered transposon MuDel into the *egfp* gene encoding enhanced green fluorescent protein (EGFP) residing within the pNOM-XP3 plasmid was performed using an *in vitro* transposition and selection procedure described previously (Baldwin et al., 2009) to generate the library *egfp*Δ^{2504}. *Mly*I restriction digestion was performed on *egfp*Δ^{2504} DNA (3 μg) to remove MuDel from the pooled plasmid library and analysed by 1.0% (w/v) agarose gel electrophoresis. The linear library DNA was purified from the agarose gel using a QIAquick® gel purification kit (QIAGEN). The purified linear library DNA (50 ng) was recircularised by intramolecular ligation with Quick T4 DNA ligase and the reaction cleaned up with a MinElute reaction cleanup kit (QIAGEN). The ligation reaction mixture (1 μl) was used to transform electrocompetent *E. coli* BL21-Gold (DE3) cells. The transformed cells were grown on LB agar plates supplemented with 100 μg/ml ampicillin and 150 μM IPTG and incubated at 37°C overnight 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 with primers pEXP-F and DDJ013. The PCR products produced (2 μl) were analysed by agarose gel electrophoresis and the rest (23 μl) purified using a QIAquick PCR purification kit (QIAGEN) for DNA sequence analysis, to identify the nature of the triplet nucleotide deletions.

Protein production and purification

The production and subsequent purification of EGFP and EGFP$^{G4\Delta}$ was performed as follows. LB Broth (15 ml) supplemented with 100 μg/ml ampicillin was inoculated with a single *E. coli* BL21-Gold (DE3) colony containing a relevant plasmid (pNOM-XP3 (Baldwin et al., 2009) containing the *egfp* or *egfp*$^{G4\Delta}$ gene) to generate a starter culture and incubated overnight at 37°C. A 1/200 dilution of the starter culture was used to inoculate 1l of LB broth supplemented with 100 μg/ml ampicillin and grown at 37 °C until an O.D.600 of 0.4-0.8 was achieved. Protein expression was induced by the addition of 1 mM IPTG and incubated for 24 hrs at 37 °C. The 1l 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 (Buffer A) and supplemented with 1 mM phenylmethanesulfonylfluoride (PMSF) and 1 mM ethyldiaminetetraacetic acid (EDTA). The cells were lysed by French press using a chilled pressure cell. The lysate was then centrifuged (20000 rpm in a Beckman JA20 rotor for 30 mins) to pellet any cell debris and the supernatant was decanted and stored at 4°C. The cell lysate was subjected to fractionation with ammonium sulphate precipitation. An initial ammonium sulphate concentration of 45% (w/v) was used to precipitate unwanted proteins from solution. After clearance of unwanted precipitate by centrifugation (20000 rpm in a Beckman JA20 rotor for 40 mins) further addition of ammonium sulphate to a final concentration of 75% (w/v) was carried out to precipitate EGFP or EGFP$^{G4\Delta}$. The precipitate was resuspended in 5 ml Buffer A. The sample was buffer exchanged into fresh Buffer A by dialysis in a 10000 MWCO membrane 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 (GE Healthcare) 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 5 column volumes with elution monitored at 280 nm and 488 nm. Pooled fractions were buffer exchanged into fresh Buffer A supplemented with 150 mM NaCl (Buffer B) with Amicon® Ultra centrifugal concentrators. Buffer exchanged protein samples were applied to a SP Superdex™ 200 gel filtration column (GE Healthcare) with elution monitored at 280 nm and 488 nm. The purified protein sample was finally stored in Buffer B. Protein concentration was determined with the DC Protein assay kit (Bio-Rad) using bovine serum albumin (BSA) as a protein standard. The assay was performed as to the manufactures guidelines for use in a microplate assay.

Size exclusion chromatography

Gel filtration standards (Biorad) were applied to a Superdex™ 75 column (20 ml bed volume, 0.5 ml/min flow rate). As per the manufacturers guidelines with protein elution monitored at 280 nm. A standard curve was generated from the plot LogMw against K_{av}, where $K_{av} = (V_e-V_o)/(V_t-V_o)$, V_e is the elution volume, V_t is the total volume and V_o is the void volume. Protein samples were prepared in Buffer B to final concentrations of 25, 50 or 100 uM and applied to a Superdex™ 75 column with protein elution monitored by absorbance at 488 nm. Elution volumes were determined for each sample and K_{av} values calculated. Using the standard curve estimated molecular weights could be determined for each protein sample.

Fit to 2 state unfolding.

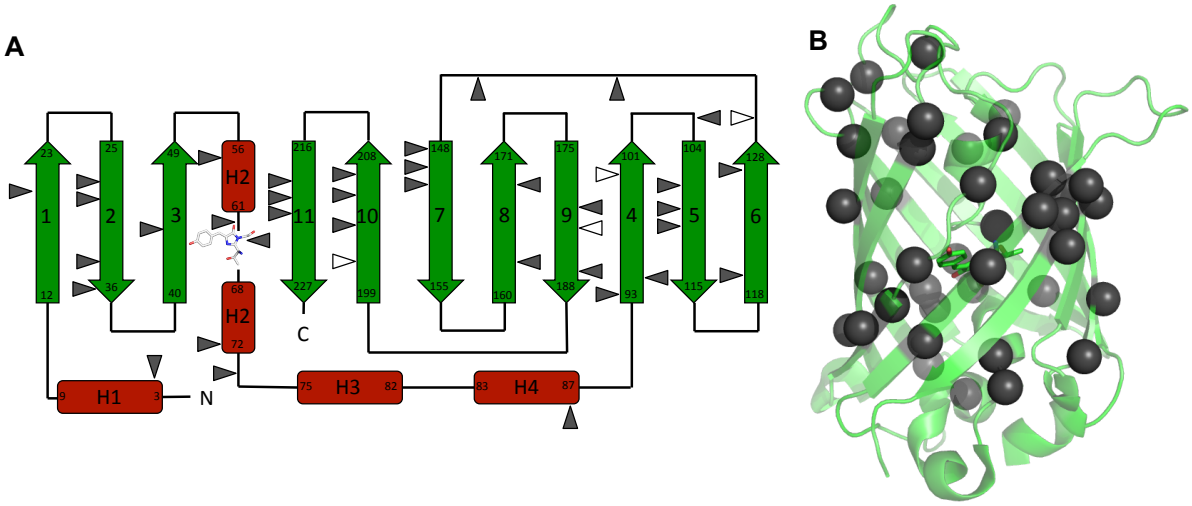
Equilibrium unfolding was fit to a 2-state model in the GraphPad Prism software (*equation 1*) to estimate approach to equilibrium (see Supporting Methods).

$$Y_N = \alpha_N + \beta_N[\mathrm{D}],\ Y_D = \alpha_D + \beta_D[\mathrm{D}]$$

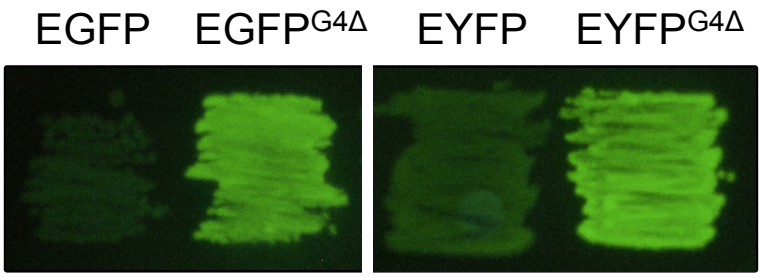
$$F = Y_N - (Y_N - Y_D)\frac{exp\left(\frac{m_{N-D}([D]-[D]_{50\%})}{RT}\right)}{1+exp\left(\frac{m_{N-D}([D]-[D]_{50\%})}{RT}\right)} \qquad \textit{equation 1}$$

Where F is the fraction of folded protein, Y_N and Y_D are intensities of native and denatured states, respectively. To take into account sloping baselines for the fluorescence data, Y_N and Y_D are described as a function of α_N, β_N, α_D and β_D, respectively. Where α_N and α_D are the fluorescence intensities of the native and denatured states, respectively, and β_N and β_D are the slopes of the native and denatured baselines. $m_{N\text{-}D}$ is a constant that describes the dependence of ΔG on denaturant concentration, [D], between the native and denatured states. $[D]_{50\%}$ is the estimated midpoint of the unfolding transition and represents the concentration of denaturant at which 50% of the protein is folded and 50% is unfolded.

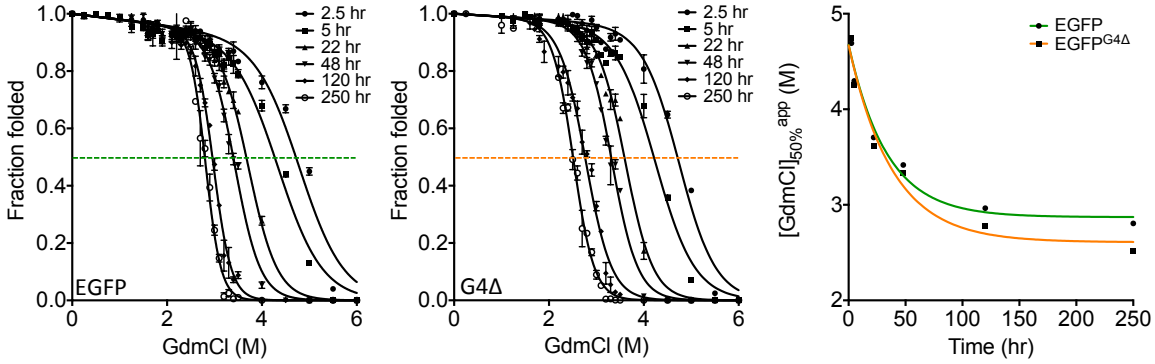
Supporting Figures.



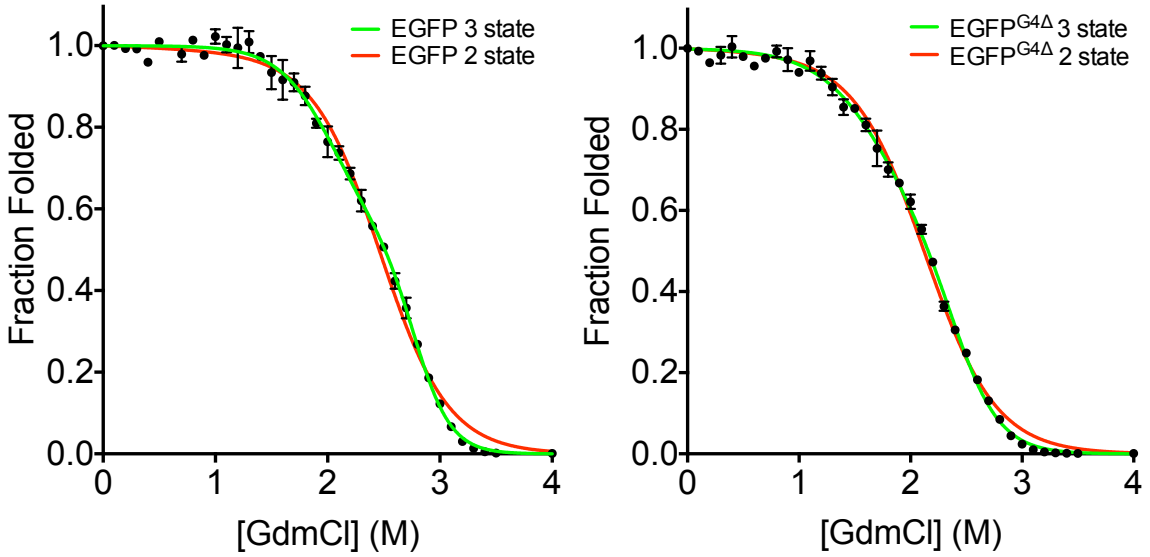
Supporting Figure S1, related to Figure 1. Mapping non-tolerated single amino acid deletion mutations with respect to EGFP (A) secondary and (B) tertiary structure. (A). The secondary structure arrangement and overall topology of EGFP shows the arrangement of β-strands (green), α-helices (red) and loops (black). Disruptive single amino acid deletions identified in this study are indicated by black triangles and trinucleotide deletions generating stop codon are shown as white triangles. (B) Map of single amino acid deletions onto the tertiary structure of EGFP. Cartoon representation of EGFP (green) with disruptive deletions indicated by black spheres.



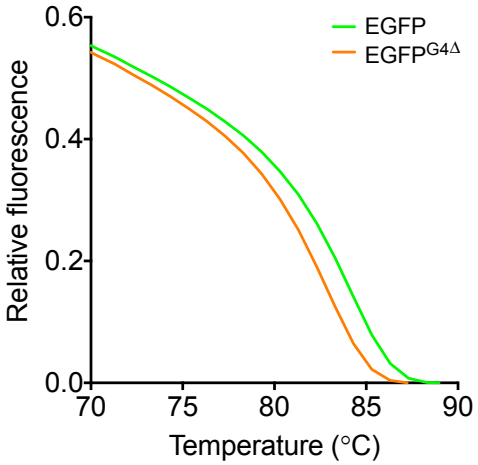
Supporting Figure S2, related to Figure 3. Colour version of cellular fluorescence of the EGFP and EYFP, and the corresponding G4Δ variants presented in Figure 3 in the main manuscript.



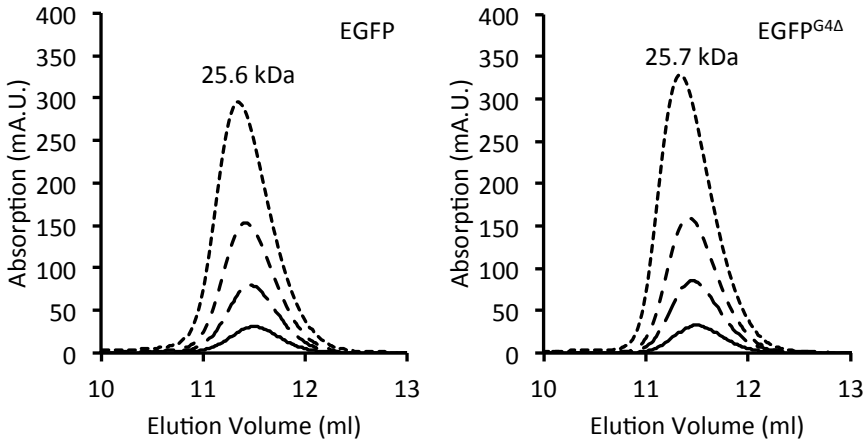
Supporting Figure S3, related to Figure 4 and Table 1. Guanidinium chloride induced equilibrium unfolding and equilibrium kinetics. Fluorescence emission at 520 nm after excitation at 480 nm was monitored for (A) EGFP and (B) EGFP$^{G4\Delta}$, over 250 hrs (as indicated in the figures) and data were fit to a two state model (GraphPad Prism). **C,** 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 close approach to equilibrium.



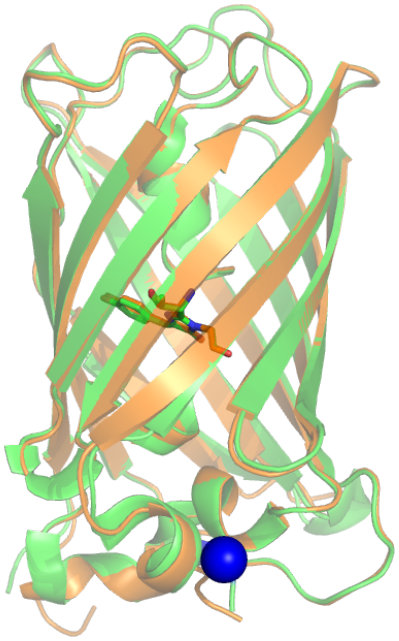
Supporting Figure S4, related to Figure 4. Two state and three state model fits to equilibrium unfolding data. Equilibrium unfolding data for EGFP (left panel) and EGFP$^{G4\Delta}$ (right panel) fit to a two state (red) or three state (green) model highlights the poor fit of the data to a two state model.



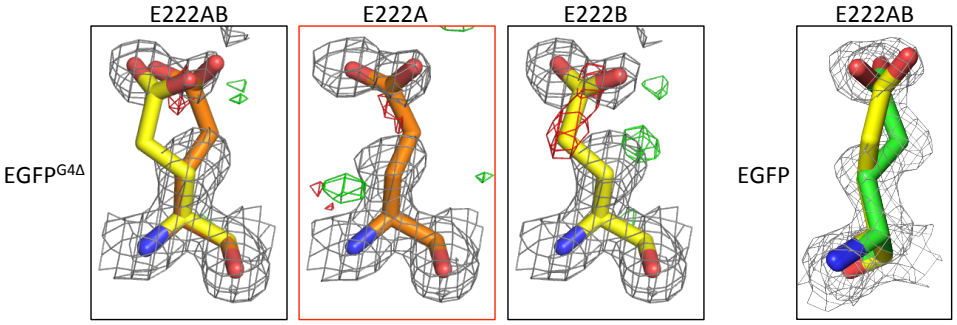
Supporting Figure S5, related to Figure 4. Thermal melting curves for EGFP and EGFP$^{g4\Delta}$. Melting temperatures (Tm) of EGFP and EGFP$^{G4\Delta}$ were determined by monitoring fluorescence with an Opticon 2 qPCR thermal cycler (MJ Research) while ramping the temperature from 25-98°C. Protein 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 was used to determined an apparent melting temperature.



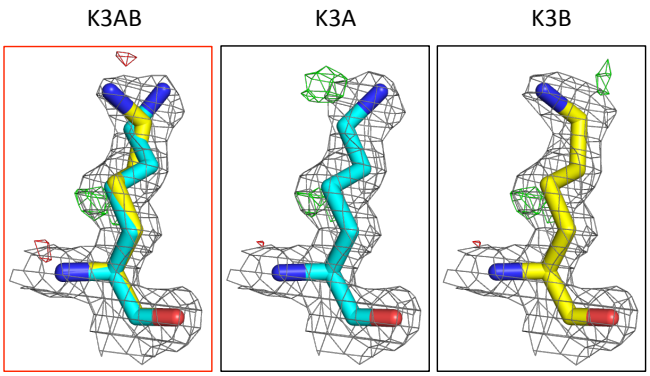
Supporting Figure S6, related to Figure 5 and Table 3. Size exclusion chromatography of EGFP$^{G4\Delta}$. The elution profiles of (A) EGFP and (B) EGFP$^{G4\Delta}$ at 10 µM (black line), 25 µM (long dash), 50 µM (medium dash) and 100 µM (short dash). The estimated molecular weight based on the peak elution volume is shown on the graph.



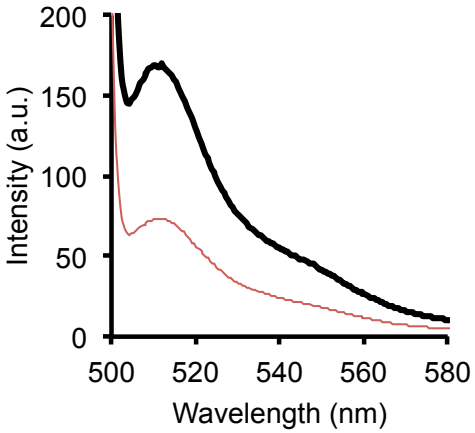
Supporting Figure S7, related to Figure 5. Overlap of EGFP (green) with EGFP$^{G4\Delta}$ (orange) with the G4 residue in EGFP highlighted as a blue sphere and the chromophore shown as stick representation. The RMSDs between the two structures in terms of backbone and all atoms was 0.6Å and 1.2Å respectively.



Supporting Figure S8, related to Figure 5. Rationale behind modelling of E222 as a single conformer in EGFP$^{G4\Delta}$. Modelling of residue E222 as either the single conformer A (E222A), the single conformer B (E222B) or as a double conformer (E222AB). The electron density does not fully support the modelling of E222 in EGFP$^{G4\Delta}$ as a double conformer. The model used in final crystal structure refinement is highlighted in the red box (E222A).



Supporting Figure S9, related to Figure 5. Rationale behind modelling of K3 as a double conformer in EGFP$^{G4\Delta}$. Modelling of residue K3 as either the single conformer A (K3A) or conformer B (K3B) does not fully satisfy the electron density. Modelling of residue K3 by both conformers does satisfy the electron density. The model used in final crystal structure refinement is highlighted in a red (K3AB) box.



Supporting Figure S10, related to Figure 3. Whole cell fluorescence emission (excited at 488 nm) spectra for cultures grown at 37°C expressing EGFP (black line) or EGFP$^{K3N\text{-}G4\Delta}$. Cell cultures were standardised to an OD_{600} of 0.1.

Supporting Table S1, related to Figure 1. Tolerated TNDs in *egfp* and subsequent amino acid mutations

Nucleotide deletion[a]	Amino acid Mutation[b]	Frequency	Secondary structure[c]	SASA (Å^2)	% SASA
$_{3}$G**TG A**GC$_{10}$	V1Δ S2G	2	N-terminus	ND	ND
$_{9}$AA**G GG**C$_{16}$	K3N G4Δ	4	H1	2.77	13
$_{12}$G**GC GA**G$_{19}$	G4Δ	8	H1	2.77	13
$_{12}$GG**C GA**G$_{19}$	E5Δ	2	H1	57.09	42
$_{18}$**GAG**$_{22}$	E6Δ	1	H1	84.96	42
$_{27}$A**CC GG**G$_{34}$	T9Δ G10R	6	H1	102.95	70
$_{27}$AC**C GG**G$_{34}$	G10Δ	2	Loop H1-S1	37.91	38
$_{36}$**GTG**$_{40}$	V12Δ	1	S1	9.20	12
$_{75}$**CAC**$_{79}$	H25Δ	2	S2	79.74	54
$_{114}$**ACC**$_{118}$	T38Δ	3	Loop S2-S3	65.55	37
$_{144}$**TGC**$_{148}$	C48Δ	1	S3	3.92	9
$_{147}$**ACC**$_{151}$	T50Δ	1	Loop S3-H2	81.50	50
$_{150}$A**CC GG**C$_{157}$	T50Δ G51S	2	Loop S3-H2	81.50	50
$_{159}$C**TG CC**C$_{166}$	L53Δ	1	Loop S3-H2	2.58	11
$_{225}$C**CC GA**C$_{232}$	P75Δ D76H	2	H3	21.59	17
$_{225}$**GAC**$_{229}$	D76Δ	2	H3	118.40	73
$_{237}$**AAG**$_{241}$	K79Δ	1	H3	58.66	24
$_{396}$GA**G GA**C$_{403}$	E132D D133Δ	1	Loop S6-S7	108.24	72
$_{411}$**GGG**$_{415}$	G138Δ	2	Loop S6-S7	26.72	21
$_{459}$A**TG GC**C$_{466}$	M153Δ A154T	2	S7	69.42	37
$_{462}$G**CC GA**C$_{469}$	A154Δ	5	S7	30.50	23
$_{465}$**GAC**$_{469}$	D155Δ	4	S7	22.16	22
$_{474}$AA**G AA**C$_{481}$	K158Δ	1	Loop S7-S8	106.96	57
$_{480}$**GGC**$_{484}$	G160Δ	1	S8	11.54	10
$_{513}$AT**C GA**G$_{520}$	I171M E172Δ	3	Loop S8-S9	88.73	39
$_{522}$**GGC**$_{526}$	G174Δ	2	Loop S8-S9	68.18	52
$_{525}$**AGC**$_{529}$	S175Δ	1	Loop S8-S9	59.04	34
$_{567}$G**GC GA**C$_{574}$	G189Δ	1	Loop S9-S10	22.96	36
$_{570}$G**AC GG**C$_{577}$	D190Δ	1	Loop S9-S10	152.83	100
$_{576}$C**CC GT**G$_{583}$	P192Δ V193L	3	Loop S9-S10	130.44	95
$_{588}$**CCC**$_{592}$	P196Δ	1	Loop S9-S10	5.11	16
$_{591}$**GAC**$_{595}$	D197Δ	1	Loop S9-S10	54.34	62
$_{594}$**AAC**$_{598}$	N198 Δ	1	Loop S9-S10	100.68	71
$_{633}$C**CC AA**C$_{640}$	P211Δ N212H	3	Loop S10-S11	112.07	58
$_{678}$G**CC GCC GG**G$_{687}$	A226Δ A227Δ	1	S11	30.10 / 28.62	12 / 20
$_{681}$G**CC GG**G$_{688}$	A227Δ	5	S11	28.62	20
$_{681}$GC**C GG**G$_{688}$	G228Δ	2	C-terminus	48.44	38
$_{690}$AC**T CT**C$_{697}$	L231Δ	1	C-terminus	178.68	93
$_{699}$A**TG GA**C$_{706}$	M233Δ D234N	2	C-terminus	ND	ND
$_{702}$GA**C GA**G$_{709}$	D234E E235Δ	2	C-terminus	ND	ND
$_{705}$**GAG**$_{709}$	E235Δ	1	C-terminus	ND	ND
$_{711}$**TAC**$_{715}$	Y237Δ	1	C-terminus	ND	ND

[a] Numbers refer to gene sequence numbering for *egfp* (GFPmut1)

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

[c] Secondary structure elements as defined by Fig 1, helices (H), strands (S).

Supporting Table S2, related to Figure 1. Non-tolerated TNDs in *egfp* and subsequent amino acid mutations

Nucleotide deletion[a]	Amino acid Mutation[b]	Frequency	Secondary structure[c]	SASA (Å^2)
$_{9}$A**AG G**GC$_{16}$	K3Δ G4S	1	H1	178.25
$_{60}$G**GC G**AC$_{67}$	G20Δ	3	S1	5.93
$_{81}$T**TC A**GC$_{88}$	F27Δ S28C	1	S2	5.40
$_{90}$T**CC G**GC$_{97}$	S30Δ G31C	3	S2	31.28
$_{99}$GG**C GA**G$_{106}$	E34Δ	2	S2	89.14
$_{105}$GG**C GA**T$_{112}$	D36Δ	1	S2	26.72
$_{135}$A**AG T**TC$_{142}$	K45Δ F46I	1	S3	45.42
$_{168}$CC**C TG**G$_{175}$	W57Δ	1	H2	12.84
$_{171}$**TGG**$_{174}$	W57Δ	3	H2	12.84
$_{189}$AC**C CT**G$_{196}$	L64Δ	1	Loop H2-H3	0.00
$_{192}$C**TG AC**C$_{199}$	L64Δ T65P	2	Loop H2-H3/Cro	0.00
$_{198}$T**AC G**GC$_{205}$	Y66Δ G67C	1	Cro	ND
$_{216}$**AGC**$_{220}$	S72Δ	1	H3	2.38
$_{219}$**CGC**$_{223}$	R73Δ	1	Loop H3-H4	87.13
$_{261}$**GCC**$_{265}$	A87Δ	2	H5	5.30
$_{279}$G**TC CA**G$_{286}$	V93Δ Q94E	1	S4	19.40
$_{282}$**CAG**$_{286}$	Q94Δ	1	S4	5.31
$_{300}$T**TC A**AG$_{307}$	F100Δ K101STOP	1	S4	3.91
$_{309}$G**AC G**GC$_{316}$	D103Δ	1	Loop S4-S5	28.42
$_{321}$A**AG AC**C$_{328}$	K107Δ	1	S5	98.33
$_{330}$G**CC G**AG$_{337}$	A110Δ	3	S5	5.59
$_{330}$GC**C GA**G$_{337}$	E111Δ	1	S5	53.03
$_{360}$**GTG**$_{364}$	V120Δ	3	S6	8.67
$_{360}$G**TG A**AC$_{367}$	V120Δ N121D	1	S6	8.67
$_{381}$G**GC AT**C$_{388}$	G127Δ I128V	1	S6	0.42
$_{390}$T**TC A**AG$_{397}$	F130Δ K131STOP	1	Loop S6-S7	10.87
$_{411}$**CTG**$_{415}$	L137Δ	1	Loop S6-S7	22.36
$_{435}$**TAC**$_{439}$	Y145Δ	1	Loop S6-S7	23.93
$_{444}$**CAC**$_{448}$	H148Δ	3	S7	9.18
$_{450}$G**TC TA**T$_{457}$	V150Δ Y151D	3	S7	0.01
$_{450}$GT**C TA**T$_{457}$	Y151Δ	2	S7	103.92
$_{486}$**AAG**$_{490}$	K162Δ	1	S8	64.53
$_{507}$**CAC**$_{511}$	H169Δ	3	S8	8.20
$_{510}$A**AC AT**C$_{516}$	N170Δ	1	S8	50.12
$_{540}$**GAC**$_{544}$	D180Δ	1	S9	44.22
$_{546}$TA**C CA**G$_{553}$	Y182STOP Q183Δ	2	S9	0.00
$_{561}$**CCC**$_{565}$	P187Δ	1	S9	17.65
$_{600}$TA**C CT**G$_{607}$	Y200STOP L201Δ	1	S10	0.55
$_{609}$AC**C CA**G$_{616}$	Q204Δ	1	S10	101.34
$_{615}$TC**C GC**C$_{622}$	A206Δ	1	S10	55.05
$_{618}$GC**C CT**G$_{625}$	L207Δ	1	S10	22.63
$_{621}$C**TG AG**C$_{628}$	L207Δ S208R	1	S10	22.63
$_{654}$AT**G GT**C$_{661}$	M218I V219Δ	1	S11	27.30
$_{660}$**CTG**$_{664}$	L220Δ	1	S11	0.00
$_{663}$**CTG**$_{667}$	L221Δ	1	S11	65.26

[a] Numbers refer to gene sequence numbering for *egfp* (GFPmut1)

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

[c] Secondary structure elements as defined by Fig 1, helices (H), strands (S).

Supporting References

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. Molecular BioSystems *5*, 764-766.