Prediction of residue-specific contributions to binding and thermal stability using yeast surface display

Shahbaz Ahmed[1], Kavyashree Manjunath[2], Raghavan Varadarajan[1*]

[1]Molecular Biophysics Unit, Indian Institute of Science, Bangalore-560012, India

[2]Institute for Stem Cell Science and Regenerative Medicine, Bangalore 560065, India

*Author for correspondence

Telephone: +91-80-2293-. 2612

Fax: +91-80-23600535

Email: varadar@iisc.ac.in

This file includes:

Main Text

Figures 1 to 7

Supplementary information

Abstract

Quantitative prediction of residue-specific contributions to protein stability and activity is challenging, especially in the absence of experimental structural information. This is important for prediction and understanding of disease causing mutations, and for protein stabilization and design. Using yeast surface display of a saturation mutagenesis library of the bacterial toxin CcdB, we probe the relationship between ligand binding and expression level of displayed protein, with *in vivo* solubility in *E.coli* and *in vitro* thermal stability. We find that both the stability and solubility correlate well with the total amount of active protein on the yeast cell surface but not with total amount of expressed protein. We coupled FACS and deep sequencing to reconstruct the binding and expression mean fluorescent intensity of each mutant. The reconstructed mean fluorescence intensity (MFI_{seq}) was used to differentiate between buried site, exposed non active-site and exposed active-site positions with high accuracy. The MFI_{seq} was also used as a criterion to identify destabilized as well as stabilized mutants in the library, and to predict the melting temperatures of destabilized mutants. These predictions were experimentally validated and were more accurate than those of various computational predictors. The approach was extended to successfully identify buried and active-site residues in the receptor binding domain of the spike protein of SARS-CoV-2, suggesting it has general applicability.

Keywords: Protein stability, mutational scanning, residue burial, free energy, saturation mutagenesis.

Abbreviations:

YSD, Yeast surface display; SSM, Site saturation mutagenesis; FACS, Fluorescence-activated cell sorting; DFE, Distribution of fitness effects; RBD, Receptor binding domain; SARS-CoV-2, Severe acute respiratory syndrome coronavirus 2; ACE-2, Angiotensin-converting enzyme 2.

Introduction

Mutagenesis is often used to generate variants of proteins with improved biophysical properties such as solubility and activity and to understand protein function. The advancement of high-throughput mutagenesis techniques has enabled the generation of a large number of variants of a protein in a short span of time, in a massively parallelizable manner [1–3]. If an appropriate functional assay to score protein activity *in vivo* exist, it is possible to infer the relative activity of each variant in the library, through library screening coupled to next generation sequencing [4–6]. However, there is a dearth of efficient, high-throughput methods to measure the solubility and stability of multiple protein variants in parallel, and to discriminate between buried and active-site residues solely using mutational data [7].

Yeast surface display (YSD) is commonly used as a tool to identify protein variants with improved biophysical properties [8,9]. YSD is preferable to bacterial expression for disulfide containing or glycosylated proteins. Agglutinin based Aga2p is the most widely used system to display proteins on the yeast cell surface [10]. Aga2p is a small protein (7.5 kDa), covalently linked via disulphide linkages to the yeast cell surface protein Aga1p [11]. Previous studies have shown that the amount of protein displayed on the yeast cell surface is directly correlated to the amount of protein secreted by the cells, as well as the thermal stability of the protein [12]. However, in other studies where the secretion efficiency [13] or yeast cell surface expression of proteins was measured, no such correlation was observed [14,15]. Proteolysis of yeast surface displayed proteins has also been used to differentiate properly folded, stable variants from unstructured variants or molten globules, as a proxy for stabilization [16–18]. However, this has primarily been applied to relatively small proteins [16–19]

A previous study which showed correlation between stability and expression levels was carried out on a limited number of mutants, that were studied individually. In addition, the WT protein itself had a very low T_m [12]. It has also been suggested that if the stability of a protein crosses a certain threshold, its expression does not increase linearly with increase in stability and it is therefore difficult to distinguish stable mutants from less stable ones, using only expression as the criterion [20]. With a very high level of yeast surface expression for unstable variants, the yeast quality control system may not be able to differentiate between properly folded, unfolded or molten globule like proteins. However, once displayed on the yeast cell surface such mutants may unfold or aggregate and hence will not bind to a tertiary structure specific ligand or cognate partner.

To verify the above hypothesis, we used *Escherichia.coli* (*E.coli*) CcdB as a model protein. CcdB is the toxin component of the CcdAB toxin-antitoxin (TA) module which binds both free DNA Gyrase and the DNA Gyrase-DNA complex, these are referred to as inhibition and poisoning respectively. Formation of the poisoned CcdB:DNA Gyrase:DNA ternary complex stalls replication and causes cell death [21]. The other component of this TA module codes for an antitoxin CcdA, which neutralizes the toxicity of the CcdB toxin upon binding to CcdB. A mutation of Arginine to Cysteine in the DNA Gyrase subunit A (GyrA) at residue 462 can abolish the binding of Gyrase to CcdB [21]. The CSH501 *E.coli* strain carries this mutation in the gene of the *gyrA* subunit which makes it insensitive to CcdB [22]. In a previous study, a single-site saturation mutagenesis library of CcdB was generated and the mutants were scored based on their *in vivo* growth phenotype (MS_{seq} score) [4]. In *E.coli*, a good correlation was found between the MS_{seq} score of ~70 mutants with either ΔT_m of purified protein (r =0.65) or *in vivo* solubility in *E.coli* (r =0.69) [23]. In contrast to plate based phenotypes, YSD provides greater flexibility and improved quantitation. We therefore wished to explore the correlation between the amount of surface expression or ligand binding seen with YSD, with thermal stability and *E.coli in vivo* solubility using this large set of characterized mutants, which had a range of *in vitro* thermal stability and *in vivo* solubility.

We initially examined 30 different variants of CcdB, which have varying solubility (when expressed in *E.coli), in vitro* thermal stability, accessibility and residue depth. The *in vivo* solubility of these mutants ranged from completely soluble to insoluble. We did not find a good correlation between total expressed protein amount on the yeast cell surface and either *in vivo* solubility in *E.coli,* or *in vitro* determined thermal stability. However, a better correlation was observed between the amount of active protein on the yeast cell surface (i.e., the amount of bound ligand) with *in vivo* solubility/thermal stability. In the yeast cell surface display system [24], activity was monitored by measuring the extent of binding of yeast cell surface displayed CcdB to a FLAG tagged fragment of GyrA14 as described previously [25].

Multiple rounds of sorting enrich mutants which have the highest expression and binding on the yeast cell surface. Sorting in such a way may lead to the identification of mutants with better biophysical properties, however, it does not give any information about the relative activity of all the mutants in a library. We coupled FACS and deep sequencing to reconstruct the MFI (MFI_{seq}) of each mutant in the Site Saturation Mutagenesis (SSM) library of CcdB, using single round FACS sorting methodology. We use this parameter MFI_{seq}, to rank all the mutants based on their activity to generate the mutational landscape or distribution of fitness effects (DFE). We found that the DFE generated using binding was more accurate than the DFE generated using expression. Overall, our MFI_{seq} scoring parameter could readily discriminate between stable and destabilized mutants of CcdB in a highly multiplexed manner.

It is well known that mutations that affect activity occur primarily at either surface exposed residues directly involved in binding or catalysis or at buried residues important for folding and stability. It has been difficult to distinguish between these two classes of residues, solely from mutational data [7]. We show here that by examining the effects of charged substitution on surface expression we can discriminate between the two classes of residues. To further validate the approach described above, we analyzed previously published saturation mutagenesis YSD expression and binding data for the receptor binding domain (RBD) of SARS-CoV-2 to its ligand ACE-2 [26]. We could successfully predict both binding-site and buried residues solely from the mutational data in this system as well.

Results

YSD of CcdB mutants

Yeast surface display (YSD) has become an increasingly popular tool for protein engineering and library screening applications [27]. Aga2p mating adhesion receptor of *Saccharomyces cerevisiae* is used as a fusion protein for yeast surface display. For surface expression, we used a vector in which CcdB is fused at the C-terminus of Aga2 [25]. We generated (Supplementary Figure S1) and individually characterized 30 CcdB variants on the yeast cell surface. Most CcdB mutants had similar levels of expression to the WT protein (Figure 1A). However, the mutants showed different amounts of active protein as assayed by binding to the FLAG tagged GyrA14 compared to the WT protein (Figure 1B). Previously, we have characterized the *in vitro* thermal stability and *in vivo* solubility of several CcdB mutants [23]. The correlation coefficient (r) between amount of total protein on the yeast cell surface with *in vivo* solubility or T_m of the corresponding purified protein were 0.44 and 0.29 respectively (Figure 2A-B). It is unclear why mutants which have very low solubility in *E.coli* are highly expressed on the yeast cell surface. It was previously hypothesized that the protein folding quality control system in yeast is not as effective as in mammalian systems, therefore partially folded/molten globule/aggregated protein may exist on the surface of yeast [14]. A correlation of r=0.81 was found between the amount of active protein on the yeast cell surface with its *in vivo* solubility determined in *E.coli* (Figure 2C). We also found a better correlation (r=0.69) between amount of active CcdB protein on the yeast cell surface and its *in vitro* thermal stability (Figure 2D), compared to that between total CcdB protein on the yeast cell surface and thermal stability.

Deep sequencing analysis of CcdB library and MFI calculation for CcdB mutants

To extend these results, an SSM library of ccdB was expressed on the yeast cell surface. Different populations based on extent of binding to gyrase or cell surface expression were sorted. A total of 32 different populations were sorted at two different concentrations of GyrA14 (100 nM, 5 nM) as a function of either surface expression level or the extent of binding to GyrA14 (Supplementary Figure S2). MFI was calculated for each mutant as explained in the Methods section. The MFI was calculated at different stringencies (where the stringency refers to the sum of reads for a given mutant over each gate of the histogram), namely 25, 50, 100, 150 and 200 reads. All mutants with a total read number less than the stringency value were removed from the analysis. As the stringency increased, the pairwise correlation between the biological replicates increased (Supplementary Table 1). The data was analysed with a stringency of 50 reads, since at higher stringencies, correlation did not improve significantly, but the number of mutants reduced. Reconstructed Binding and Expression MFI from deep sequencing data are hereafter referred to as MFI_{seq} (bind) and MFI_{seq} (expr) respectively.

MFI reconstruction and its correlation with stability, solubility and residue burial.

A few published studies have described estimation of MFI values using deep sequencing of sorted populations and are therefore similar to our experimental strategy. However, the procedure for MFI reconstruction in these reports was relatively complicated compared to that used here. [28–31]. In the present study, we calculated an absolute mean MFI instead of a relative MFI. A good correlation was found between the MFI of individually analysed mutants and their corresponding MFI_{seq} values, validating our approach of MFI reconstruction (Supplementary Figure S3A, 3B). Individually analysed mutants showed a good correlation between the amount of active protein on the cell surface and *in vitro* measured thermal stability of the purified protein. Similarly, we also found a good correlation between MFI_{seq} (bind) of mutants inferred from deep sequencing, and thermal stability as well as *in vivo* solubility for the selected mutants (Supplementary Figure S3C, 3D).

For the exposed (>10% accessibility) and active-site residues (from PDB ID:1X75), mutations did not affect the degree of surface expression (Figure 3A). However, many buried site mutants showed very low expression, possibly because of aggregation and degradation inside cells or during export (Figure 3C). In the case of binding, a very high mutational sensitivity was found both at buried and active-site residues (Figure 3D) similar to the previous report of CcdB mutants in *E.coli* [23]. We also found a very high mutational sensitivity of binding for a few non-interacting residues in the loop connecting beta strands S2 and S3 at both 5 nM and 100 nM GyrA14 concentration (Supplementary Figure S4). The residues I24, I25 and D26 in this loop are directly involved in interacting with Gyrase and mutation at non-interacting residues (22, 23 and 27) in the loop might restrict or alter the conformation of the loop, thus reducing the affinity of CcdB mutants to GyrA14. However, there was no effect on the expression of the mutants in this loop, indicating that the mutant proteins are not destabilized (Supplementary Figure S4). We did not find a high correlation between MFI$_{seq}$ (bind) and either accessibility or depth, because many mutations at both buried and active-site residues have high mutational sensitivity (Supplementary Table 2). The previously described parameter RankScore, is a measure of mutant activity in *E.coli* [4] with high RankScore denoting lower activity. We found a poor correlation between the MFI$_{seq}$ (bind) values of CcdB mutants at both exposed non active-sites as well as active-site residues, and RankScore. In *E.coli*, most of the exposed non active-site residues do not show any mutational sensitivity, i.e. they have the same RankScore values as WT. However, in the present case many such CcdB mutants show lower binding to GyrA14 compared to WT. The loss of binding could be attributed to the decrease in the affinity between CcdB and Gyrase, or destabilization due to mutation. We defined a new parameter MrMFI (mean residue MFI) which is the mean of the MFI values of all the mutants at a certain position. MrMFI (expr) and MrMFI (bind) at 100 nM GyrA14, show a good correlation with RankScore. (Supplementary Table 2). MrMFI (expr) also showed good correlation with Depth which is a structural measure of residue burial [32]. However, in the case of binding at 5 nM, a weaker correlation of MrMFI (bind) with the aforementioned parameters was observed (Supplementary Table 2). In previous studies, identification of the active-site residues solely from the deep sequencing data was not very efficient [4,7], this is presumably because *in vivo* activity is often governed by threshold effects, and because mutations at buried residues also affect activity. The current methodology removes such drawbacks. We could distinguish between buried and active-site resides by comparing the MFI_{seq} (bind) and MFI_{seq} (expr). Most buried site residues showed low values of both MFI_{seq} (bind) and MFI_{seq} (expr) compared to WT. However, the active-site residues showed low MFI_{seq} (bind) but similar MFI_{seq} (expr) compared to WT. We found that the average MFI_{seq} values of charged residues are a good predictor to discriminate between buried and active-site residues. For calculating $MrMFI_{charged}$ of charged WT residues, we only consider mutants with opposite charge. For some mutants at buried positions, we found a very low $MrMFI_{charged}$ (expr) but the mutants were absent in $MrMFI_{charged}$ (bind). We found that such mutants had very high reads, suggesting that the values of $MrMFI_{charged}$ (expr) are correct. We anticipated that such mutants lack binding and are therefore present only in the bin which had a background level of binding signal, the presence of mutant in only that gate led to the removal of such mutants due to the stringency set for the analysis. Hence, such mutants were assigned a $MrMFI_{charged}$ (bind) similar to other buried positions. $MrMFI_{charged}$ had a bimodal distribution (Supplementary Figure S5), so k-means clustering was performed to identify the mean (μ) and standard deviation (σ) of each distribution. The distributions were named D1 (higher $MrMFI_{charged}$) and D2 (lower $MrMFI_{charged}$). Buried site residues were assigned to be those which have $MrMFI_{charged}$ (bind)) and MFI_{seq} (expr) less than the set threshold (μ+0.5*σ) for distribution D2. Active-site residues were assigned as those which had $MrMFI_{charged}$ (bind)) less than (μ+σ) of the D2 distribution and MFI_{seq} (expr) higher than (μ-2*σ) of distribution D1 (Figure 4). The accuracy, specificity and sensitivity of prediction of exposed non active-site, buried and exposed active-site residues are mentioned in Supplementary Table 3.

Selection and characterization of putative stabilized mutants from deep sequencing data.

In the previous section, we discussed the correlation between protein biophysical properties like thermal stability and *in vivo* solubility with either the amount of active protein or the ratio of active protein to total protein on the yeast cell surface for a few (30) mutants. However, most of these mutants were destabilized with respect to the WT protein. To confirm whether this correlation also holds for mutants that have stability similar or greater than WT, we selected a few CcdB mutants based on either the $\mathrm{MFI_{seq}}$ (bind) or $\mathrm{MFI_{seq}}$ (ratio) ($\mathrm{MFI_{seq}}$ (bind)/ $\mathrm{MFI_{seq}}$ (expr)) for *in vitro* characterization of thermal stability. We examined the average and standard deviation of expression for all mutants and selected only those mutants which cross a minimum cut-off ($\mu+\sigma$) for $\mathrm{MFI_{seq}}$ (expr) to remove the bias created by mutants which have very low expression. No threshold for expression was set for selection of mutants based on their $\mathrm{MFI_{seq}}$ (bind). No selection of the mutants was performed based solely on the $\mathrm{MFI_{seq}}$ (expr).

Seven mutants were characterized using the criteria $\mathrm{MFI_{seq}}$ (bind) at 5 nM GyrA14, none of them showed a higher T_m than WT (Figure 5A); whereas only one of the mutant selected on the basis of $\mathrm{MFI_{seq}}$ (ratio) showed a significantly higher T_m than WT (Figure 5B). A subset of seven mutants was selected based on $\mathrm{MFI_{seq}}$ (bind) at 100 nM GyrA14. Two of the mutants were more stable, two were similar to WT and the remaining three showed a lower T_m than WT CcdB (Figure 5C). Seven mutants were selected based on $\mathrm{MFI_{seq}}$ (ratio) and characterized, two showed higher stability, three mutants were similar to WT and the remaining three were less stable than WT CcdB (Figure, 5D). We therefore hypothesize that if the stability of a mutant crosses a threshold then its expression will not increase further. To confirm this hypothesis, we measured the amount of active protein on the yeast cell surface for seven individual mutants which had T_m's ranging from 60 °C to 70 °C, and found that the expression and binding for these mutants are similar to each other and to WT (Supplementary Figure S6).

Prediction of thermal stabilities of putative destabilized mutants.

For destabilized mutants we observed a good correlation between MFI_{seq} (bind) and T_m of individual mutants (Supplementary Figure S3D). Using this correlation, we next predicted the T_m of each mutant for an additional set of (n=28) previously described CcdB mutants [23] based on their MFI_{seq} (bind). We found a good correlation (r=0.83) between predicted and *in vitro* measured T_m for this set of CcdB mutants as well (Supplementary Figure S7A). This now allows us to identify putative destabilized mutants and accurately predict the extent of destabilization for all such mutants in the CcdB YSD library. We also predicted the thermal stability of CcdB mutants using the *in silico* tool HoTMuSiCv1.0 [33], however, we did not find a good correlation between measured and predicted T_m (Supplementary Figure S7B). It has been shown that *in vitro* protein thermal stability and free energy of unfolding are correlated [23,34,35]. We therefore predicted the free energy of unfolding for CcdB mutants using SDM [36], mCSM [37], PoPMuSiC [38], DynaMut [39], DUET [40], MAESTROweb [41], DeepDDG [42], CUPSAT [43], PremPS [44] and INPS-MD [45]. We found moderate correlations, with DeepDDG performing the best (r=0.59), but still poorer compared to our prediction from YSD data (r=0.83). For a more detailed comparison we analysed the predictions of stability by DeepDDG since this showed the highest correlation with measured stability of individual mutants at non active-site residues. We excluded residues 21, 22, 23 and 27 as these positions behaved like active-site residues. We found that trends for ΔΔG predicted by DeepDDG for exposed non active-site residues are similar to those obtained from MFI_{seq} (bind) (Figure 6A, 6B). However, we observed some mutant specific differences at residues 8, 16, 50, 53 and 96. Mutations at residues 50 and 96 have highly deleterious effects which reduced GyrA14 binding to yeast surface displayed protein, these are only partially predicted by DeepDDG. In the case of charged and polar mutations at residue 8, 16 and 53 we did not observe a reduction in binding, but the software predicted them to be destabilizing. In the case of buried positions, we found mutation specific effects at 35, 52 and 94 where DeepDDG predicted changes were significantly smaller than the experimentally observed ones. We also found that most of the phenylalanine, tryptophan and arginine mutations were highly destabilizing and the mutants did not bind to GyrA14, however the software gave a lower stability penalty for these substitutions (Figure 6C, 6D). Our MFI based measurements suggested greater destabilization for several mutants relative to DeepDDG prediction. While the overall trends were similar, as discussed above, there are several differences between MFI based and DeepDDG based stability predictions.

Deep mutational scanning of SARS COV-2 receptor binding domain (RBD)

To examine the generality of our approach, we also analyzed recently reported deep mutational scanning data of the SARS-CoV-2 receptor binding domain [26]. In this study two separate libraries were generated and individually sorted based on expression and binding to ACE-2. The binding (Sortseq (bind)) or expression ((Sortseq (expr)) MFIs relative to WT for barcoded mutants were calculated from the deposited NGS data as explained in the Methods section. Additionally, we analyzed binding at only one concentration of ACE-2 (100 pM, TiteSeq_09) at which the binding started to saturate. Buried residues were those with <10% side chain accessibility in chain C of PDB ID 7KMH [46]. ACE-2 binding (active-site) residues were assigned as those contacting ACE-2 [47]. To identify the active-site and buried residues from Sortseq data, we calculated the $MrMFI_{charged}$ for each position. Similar to CcdB, we observed a bimodal distribution for both $MrMFI_{charged}$ (bind) and $MrMFI_{charged}$ (expr) (Supplementary Figure S8) and k-means and standard deviation were calculated for both the distribution D1 (higher $MrMFI_{charged}$) and D2 (lower $MrMFI_{charged}$). As described above for CcdB, buried residues were identified as those which had $MrMFI_{charged}$ (bind) and $MrMFI_{charged}$ (expr) less than the set threshold (μ+0.5*σ) for distribution D2. The active-site positions were identified as those which had MrMFI$_{\text{charged}}$ (bind) lower than the set threshold (μ+σ) for population D2 and MrMFI$_{\text{charged}}$ (expr) values higher then (μ-2*σ) for population D1. We accurately identified most of the buried residues, however there were some false positive and false negative predictions relative to the crystal structure information (Figure 7). We found 21 positions to be false negative buried positions. We categorized these false negatives into two categories, namely, glycine and the side chains which are pointing towards the surface. The accessibility calculated by DEPTH server for glycine was zero and we therefore expected glycine to fall into the false negative buried category. Thirteen positions out of twenty-one false negative were glycine. Another six positions, 336, 348, 361, 443 and 480 had their side chains pointing towards the protein surface. We also found similar false negative buried residues in CcdB where the side chain hydrophilic group was pointing towards the protein surface. Position 363 and 365 had accessibility <10% and were pointing towards the core of the protein in the PDB (7KMH) used to calculate accessibility. However, we found that these positions have high accessibility (>30%) in another structure (PDB ID 7D2Z). All the available RBD structures are in complex with other molecules this might be responsible for variation in the accessibility of residues in different RBD structures. We found 17 false positive buried residue predictions, seven of them were aromatic, seven are charged or polar, two are prolines and one is an aliphatic residue. The specificity, sensitivity and accuracy of prediction is mentioned in Supplementary Table 3. Active site residues were identified with very high accuracy (Supplementary Table 3), though there were a few false negative and false positive predictions. Additionally, we found several positions which had Sortseq (expr) like WT, however, they had very low Sortseq (bind) (Supplementary Figure S9A, S9D). We hypothesize that these positions are also assisting in the maintenance of proper RBM conformation and enabling its binding to ACE-2.Residues 447, 448, 473 and 476 which gave false positive results, 447 and

476 are part of the receptor binding motif (RBM) and contain glycine in a conformation which is available only for glycine. Hence mutation to a non-Gly residue will likely disrupt the conformation of the RBM thus decreasing binding to ACE-2. Mutations at positions 446, 453, 493 and 498 gave false negative results. Of these false negative positions, 446 is again glycine. We found that the Arg mutants at N493 and N498 positions have very little effect on binding (supplementary Figure S9F). We hypothesized that these positions may not have the most optimal WT residue, or they may show no mutational penalty for binding to ACE-2. A recent report showed that the affinity of Q498R to ACE-2 is higher than WT RBD [48] and was enriched as double mutant Q498R/N501Y when selected for RBD mutants having high affinity towards ACE-2 [49]. It has also been reported that when chimeric virus evolved in the presence of neutralizing antibodies C121 and C141, this enriched for the Q493R mutation. The mutant virus grows to high PFU titers similar to WT, and infectivity is also inhibited by a chimeric ACE-2 analog, similar to WT [50]. The specificity, sensitivity and accuracy of prediction is mentioned in Supplementary Table 3.

Discussion

With the advancement of mutagenesis and directed evolution methodologies, proteins with modified traits and function can be developed in a relatively short duration of time [51–53]. *E.coli* remains an expression host of choice for many proteins and high level, soluble *E.coli* expression is a desirable attribute. When eukaryotic or unstable prokaryotic proteins are overexpressed in bacteria, they often tend to form insoluble aggregates called inclusion bodies (IB). Formation of IBs often results in low yields of purified soluble protein. Designing improved variants of a protein by increasing half-life, stability and activity is an ongoing requirement of most pharmaceutical and biotechnology industries. However, a reliable, high-throughput, efficient and rapid method is required for solubility and stability analysis of engineered proteins. Previously, several high-throughput methods to select for soluble expression have been developed based on fusion to a reporter protein. These rely on the reporter activity, which is perturbed if an aggregation prone protein is fused [54–57]. These methods can be used to isolate protein variants with enhanced solubility but cannot reveal if the fused protein is properly folded. In some cases, such unstable proteins may also form soluble aggregates [23]. Since many of these reporter screens employ cytoplasmic expression and use bacterial hosts, disulphide rich or glycosylated proteins, or those binding to complex ligands cannot be studied. Yeast surface display coupled to FACS, has been widely used to evolve such targets. Typically, populations are sorted for multiple rounds to enrich for stable binders to a target of interest [58–61]. While this approach readily selects for high affinity binders, selecting for stable proteins is more difficult. In some cases, this methodology has also been used to isolate stable variants of proteins [27] and a good correlation was observed between surface expression and improved biophysical parameters. However, other studies in different systems did not find such a correlation [14,15].

In the present work we utilize YSD to measure the amount of total protein as well as total active protein displayed on the yeast cell surface. A good correlation was found between the amount of active CcdB mutant on the yeast surface and corresponding *in vivo* solubility in *E.coli* (r=0.82) or T_m (r=0.70). A recent report also suggests that the amount of active protein on the yeast cell surface can be used as a criterion to isolate stable mutants [20]. In the present study, no correlation was found between the amount of total protein on the yeast cell surface and the biophysical properties of mutants. A few mutants which have very low solubility in *E.coli* showed very high expression, but there was a negligible amount of active protein on the yeast surface. It has been previously suggested that the quality control system in yeast is not able to discriminate these mutants from properly folded ones or alternatively that the folded conformation is maintained by chaperones in the ER [62]. Once these mutants are exported to the cell surface they may start to unfold. This could be one reason why some groups including ours did not find a good correlation of surface expression with the stability or solubility of these proteins. In previous studies [12], a very limited number of proteins were used for surface expression studies, it is possible that in this small number, mutants which had high surface expression or secretion but lower stability than WT were not observed.

Yeast surface display coupled to FACS typically requires multiple rounds of sorting to enrich variants with desired activity and phenotype. Here, we have performed a single round of sorting and developed a rapid, uncomplicated procedure of estimating MFI's of individual mutants of CcdB combining FACS and deep sequencing. This MFI_{seq} was shown to correlate well with the corresponding experimentally measured MFIs for several individual mutants. The MFI_{seq} was used to generate the mutational landscape of expression and binding of a mutant library. We showed that such data can be used to accurately discriminate between buried, exposed non active-site and exposed active-site residues both for CcdB and an unrelated protein, RBD of the spike protein of SARS-CoV-2. Highly destabilizing charged mutations in the core of the protein decreased both expression and binding, while the active-site residues showed reduction in binding alone for charged mutations. Relative to an earlier study which assayed *in vivo* activity in *E.coli* [4], the present methodology is better able to identify and distinguish between the two categories of mutationally sensitive residues, namely buried and exposed active-site residues. In general, mutations that affect total activity *in vivo* can do so by affecting specific activity without changing the amount of folded protein, decrease the amount of folded protein without affecting specific activity or a combination of the above. The present analysis distinguishes between the above possibilities, and is therefore able to distinguish buried from exposed, active-site positions. This is useful for applications that attempt to use saturation mutagenesis data for protein model discrimination and structure prediction [63,64] as well as interpreting clinical data on disease causing mutations [65,66]

MFI_{seq} (bind) was also used to predict the T_m of CcdB mutants. We found a good correlation between predicted and measured ΔT_m for a subset of CcdB mutants. We also compared the accuracy of *in silico* tools used to predict the stability of mutants and found that these tools had lower accuracy relative to our approach. We used experimental stability measurements for a small number of destabilized mutations, combined with MFI_{seq} measurement to predict stabilities of all destabilized mutants in the saturation mutagenesis library. We could readily identify destabilized mutants of CcdB, however, the recovery of mutants more stable than WT was lower, but still significant, considering the rarity of such mutations. This is likely due to the possibility that if the stability of the protein crosses a threshold, additional increments in stability do not result in enhanced expression or binding.

A limitation of the present approach is that it requires an epitope tagged or fluorescently labelled conformation specific binding partner. Another limitation could be differential relative stability of proteins upon yeast cell surface display compared to expression in the native host and/or intracellular expression. For glycosylated proteins, the stability of mutants may also be altered because of hyper glycosylation of protein on the yeast cell surface compared to proteins expressed in mammalian systems or prokaryotic systems where glycosylation is absent. The presence of glycosylation may also affect the binding to a cognate partner which in turn may give rise to false results. This does not appear to be the case for the SARS-CoV-2 RBD which contains a single glycan at residue 343, but may be an issue for protein with multiple glycosylation site. We are examining these possibilities in ongoing studies. Despite these caveats, the present study suggests that the proposed methodology can accurately distinguish buried from active-site residues, quantitatively estimate thermal stabilities of destabilized mutants in large libraries and also be used to identify stabilized mutants.

Materials and Methods

Bacterial strains, yeast strains and plasmids

E.coli CSH501 strain carries a mutation in the gyrA gene which abolishes inhibition and poisoning by CcdB [67]. The EBY100 strain of Saccharomyces cerevisiae has the aga1 gene under the Gal1 promoter for inducible expression and a TRP1 auxotrophic mutation. The strain lacks the aga2 gene, so only Aga2p fused protein expressed from the plasmid, will form a complex with the Aga1p for yeast cell surface display [68]. The ccdB gene was cloned in the pBAD24 plasmid for controllable expression in *E.coli*. ccdB mutants were cloned in the pPNLS shuttle vector for yeast cell surface expression [69].

Cloning of WT and mutant ccdB in *E.coli*.

ccdB mutants in pBAD24 were generated using three fragment Gibson assembly. Briefly, ccdB was amplified in two fragments using two sets of oligos (Supplementary Figure S1). For each fragment one of the oligos binds to the vector and the other binds to the gene. The primer of both fragments which bind to the gene were completely overlapping and contained the desired mutation. The fragments were gel extracted and Gibson assembled with NdeI and HindIII digested pBAD24 vector. The Gibson assembled product was electroporated in *E.coli* CSH501 strain and positive transformants were selected on LB agar media containing ampicillin (100 µg/mL). The sequence was confirmed by Sanger sequencing. Sequence confirmed WT or mutant ccdB in pBAD24 vector was used as a template for PCR to amplify the ccdB gene by Vent DNA polymerase. The PCR amplified product was co-transformed with SfiI digested pPNLS vector in the EBY100 strain of S*accharomyces cerevisiae* using LiAc/SS carrier DNA/PEG method for *in vivo* recombination [70]. Positive transformants were selected on SDCAA Tryptophan dropout media plates and the sequence was confirmed by Sanger sequencing.

Protein Purification

WT and mutant CcdB was purified as described previously [71]. Briefly, an overnight culture was diluted 100-fold in LB media containing ampicillin (100µg/ml) and induced with L-arabinose (0.2% w/v) at an OD_{600} of ~0.5. Following induction for 3 hours, cells were harvested and lysed by sonication. The soluble fraction was separated using centrifugation and incubated with CcdA peptide (residues 45-72^{n}) coupled to Affigel-15 at 4 ^{o}C. The unbound fraction was removed and the column was washed with bicarbonate buffer (50 mM $NaHCO_3$, 500 mM NaCl, pH 8.5). The bound protein was eluted with 200 mM glycine (pH 2.5) and collected in an equal volume of 400 mM HEPES buffer (pH 8) to neutralize the acidity of glycine.

GyrA14 was purified as described previously [72]. Briefly, an overnight culture was diluted 100-fold in LB media containing ampicillin (100µg/ml) and induced with IPTG (1 mM) at an OD_{600} of ~0.5. Following induction for 3 hours, cells were harvested and resuspended in TES buffer (0.2 M Tris, pH 7.5, 0.5 mM EDTA, 0.5 M sucrose and 1 mM PMSF). Cells were lysed and the soluble fraction was separated using centrifugation. The soluble fraction was incubated with pre-equilibrated Ni-NTA beads for 2 hours at 4 ^{o}C. The unbound fraction was removed, and the column was washed with 100 column volumes of wash buffer (50 mM imidazole in 0.05 M Tris, pH 8, 0.5 M NaCl). The protein was eluted with 500 mM imidazole in 0.05 M Tris, pH 8, 0.5 M NaCl and dialysed against 1x PBS.

Estimation of solubility of WT and mutant CcdB in *E.coli.*

E.coli CSH501 strain, transformed with pBAD24 plasmid containing WT or mutant ccdB, was grown in media containing ampicillin for 16 hours at 37 ^{o}C and 180 RPM. A secondary culture was grown by diluting overnight grown culture 100-fold. Upon reaching an OD_{600} of 0.4-0.5, CcdB variants were induced with Arabinose at a final concentration of 0.2%(w/v) for 3 hours. The cells were harvested from 1.5 ml culture and lysed in 500µL 1X PBS, using sonication.

Supernatant and pellet fractions were separated by centrifugation at 13000 RPM at 4 °C. The pellet fraction was resuspended in 500 µL 1X PBS and equal volumes of pellet and supernatant fractions were loaded on Tricine-SDS-PAGE to measure the relative amounts of protein in each fraction.

Protein thermal stability measurement using Thermal shift assay (TSA)

The thermal shift assay was conducted in an iCycle iQ5 Real Time Detection System (Bio-Rad, Hercules, CA). A solution of total volume 20 µL containing 10 µM of the purified CcdB protein and 2.5X Sypro orange dye in suitable buffer (200 mM HEPES, 100 mM glycine), pH 7.5 was added to a well of a 96-well iCycler iQ PCR plate. The plate was heated from 15 °C to 90 °C with a 0.5 °C increment every 30 seconds. The normalized fluorescence data was plotted against temperature and T_m measured as described [23,73].

Yeast surface expression of WT and mutant CcdB proteins in EBY100 cells and flow cytometric analysis.

Saccharomyces cerevisiae EBY100 cells containing WT ccdB or mutant in pPNLS plasmids were grown in three ml SDCAA media (glucose 20g/L, yeast nitrogen base 6.7g/L, casamino acid 5g/L, citrate 4.3g/L, sodium citrate dihydrate 14.3g/L) for sixteen hours. Grown cells were diluted to an OD_{600} of 0.2 in three ml SDCAA media and grown till the OD_{600} reached two. Thirty million cells were harvested using centrifugation and resuspended in three ml SGCAA induction media (galactose 20g/L, yeast nitrogen base 6.7g/L, casamino acid 5g/L, citrate 4.3g/L, sodium citrate dihydrate 14.3g/L) for sixteen hours at 30 °C, 250 RPM [24]. One million cells were used for flow cytometric analysis. The amount of total protein expressed on the yeast cell surface was estimated by incubating the induced cells in 20 µL FACS buffer (1X PBS and 0.5% BSA), containing chicken anti-HA antibodies from Bethyl labs (1:600 dilution) for 30 minutes at 4 °C. This was followed by washing the cells twice with 100 µL FACS buffer at 4 °C. Washed cells were incubated with 20 µL FACS buffer containing goat anti-chicken antibodies conjugated to Alexa Fluor 488 (1:300 dilution), for 20 minutes at 4 ^{o}C. Fluorescence of yeast cells was measured by flow-cytometric analysis. The total amount of active protein on the yeast cell surface was estimated by incubating the induced cells in 20 µL FACS buffer containing 100 nM GyrA14 for 45 minutes at 4 ^{o}C. Cells were washed and incubated with 20µL mouse anti-FLAG antibodies (1:300). This was followed by washing the cells twice with FACS buffer, followed by incubating with 20 µL rabbit anti-mouse antibodies conjugated to Alexa Fluor 633 (1:1600 dilution). The flow-cytometric analysis was carried out on BD Accuri or BD Aria III instruments.

Yeast surface expression and sorting of CcdB Single-site saturation mutagenesis (SSM) library.

Previously, an SSM library of ccdB was generated in the pBAD24 vector [4,23]. The library was PCR amplified using primers having homology to the pPNLS vector. The PCR amplified library was gel extracted and cloned in pPNLS vector using yeast *in vivo* recombination.

A similar protocol was used for sample preparation of the library for FACS as described above for the single mutants with slight modifications. Briefly, ten million cells were taken for FACS sample preparation and the reagents were used in 10X higher volumes compared to the earlier flowcytometric analysis. Two different concentrations of GyrA14 (100 nM, 5 nM) were used for sorting CcdB mutants based on the binding in the 1D histogram. The cells were sorted in 11 and 10 different populations (bins) in case of binding with GyrA14 at concentrations of 100 nM and 5 nM respectively. Additionally, 11 different populations (bins) were sorted from the expression histogram. The experiment was repeated in a biological replicate. The sorting of CcdB libraries was performed using a BD Aria III cell sorter.

Sample preparation for deep sequencing

Sorted populations were grown on SDCAA agar plates for 48 hours. Colonies were scraped and plasmids were extracted from the cells. The ccdB gene was PCR amplified using primers which bind upstream and downstream of the ccdB sequence and had multiplex identifier (MID) sequence to segregate the reads from different sorted bins. The DNA was amplified for 15 cycles using PCR and the amplified product was gel extracted and purified. Equal amounts of DNA from each sorted population were pooled, and the library was generated using the TruSeq™ DNA PCR-Free kit from Illumina. The sequencing was done on an Illumina HiSeq 2500 250PE platform at Macrogen, South Korea after incorporating 20% ϕX174 DNA in the library.

Analysis of deep sequencing data

Deep sequencing data for the ccdB mutants obtained from the Hiseq 2500 platform was processed using a pipeline developed by adopting certain aspects from an already existing in-house protocol (https://github.com/skshrutikhare/cys_library_analysis). The latter method involved the alignment with wild type sequence followed by merging of the paired-end reads, while in the modified protocol, the reads are first merged and then aligned with the wild-type sequence. The present methodology consists of the following steps: assembling the paired end reads, quality filtering, binning, alignment and mutant identification. All these steps were incorporated in a pipeline and made executable from a single command using a parameter file unique to a given data-set. In the first step, paired end reads were assembled using the PEAR v0.9.6 (Paired-End Read Merger) tool [74]. The "quality filtering" step involved deletion of terminal "NNN" residues in the reads, and removal of reads, not containing the relevant MID and/or primers, along with the reads having mismatched MID's. Finally, only those reads having bases with Phred score ≥ 20 are retained. A binning step involved further filtering, which eliminated all those reads having incorrectly placed primers, truncated MIDs/primers (due to quality filtering) and shorter/longer sequences than the length of the wild type sequences. The remaining reads were binned according to the respective MIDs. In the alignment step, reads were aligned with the wild type ccdB sequence using the Water v6.4.0.0 program [75] and reformatted. The default values of all parameters, except the gap opening penalty, which was changed to 20, was used. In the final step of "substitution", reads were classified based on insertions, deletions and substitutions (single, double etc mutants).

MFI reconstruction from deep sequencing data

Reads of each mutant were normalized across different bins individually (Equation 1), and the fraction of each mutant (*Xi*) distributed amongst the different bins was calculated (Equation 2). The reconstructed MFI for an individual mutant was calculated by the summation of the product, obtained upon multiplying the fraction (*Xi*) of the mutant in a particular bin (*i*) with the MFI of the corresponding bin obtained from the FACS experiment (*Fi*), across the various bins populated by the respective mutant (Equation 3).

Normalized read of mutant in bin *i* (*Ni*) = $\frac{\text{No. of reads of mutant } i \text{ in bin } i}{\sum \text{reads in bin } i}$ **....** Equation 1

Fraction of mutant in each gate (*Xi*) = $\frac{\boldsymbol{Ni}}{\sum_1^n \boldsymbol{Ni}}$ **....** Equation 2

Reconstructed MFI = $\sum_1^n Fi * Xi$ **....** Equation 3

The MFI$_{seq}$ of the biological replicates were different so the MFI$_{seq}$ of one of the replicates was adjusted using “m” and “c” obtained from the correlation between the replicates and then averaged.

Average MFI$_{seq}$ = $\frac{\text{MFIseq (replicate 1)}+(\text{m}*\text{MFIseq (replicate 2)}+C)}{2}$

Depth, accessibility and RankScore calculations.

Depth was calculated using the server DEPTH [32,76]. Accessibility was calculated using the program NACCESS [77]. In both cases, the input co-ordinates were homodimeric CcdB (PDB ID 3VUB). RankScore and MS_{seq} are measures of mutational sensitivity in *E.coli*. Values were obtained from Adkar et al [4]. Buried residues were those with <10% accessibility in 3VUB. Active-site residues were those with ΔASA>0. ΔASA difference between the solvent accessible surface area of CcdB residues in the free (3VUB) and GyrA14-bound forms (1X75) respectively [78].

Deep mutational scanning of SARS COV-2 receptor binding domain (RBD)

The deep mutational scanning data was taken from a recent report [26] in which two independent libraries of RBD were generated and sorted in four different bins based on expression or binding to ACE-2. In the MFI of binding and expression for individual mutants was reconstructed in that study using a maximum likelihood method using fitdistrplus R package. The expression MFI (Sortseq (expr)) data was shared by the authors in a repository (https://github.com/jbloomlab/SARS-CoV-2-RBD_DMS). We reconstructed the binding MFI (Sortseq (bind)) at an ACE-2 concentration of 100 pM (TiteSeq_09). For Sortseq (bind) estimation we used the script provided by the authors (https://github.com/jbloomlab/SARS-CoV-2-RBD_DMS/blob/master/results/summary/compute_expression_meanF.md).The authors used data from both single and multiple mutants, together with a model to account for epistatic effects to infer the MFI values for individual mutants. We modified the script to change the input data required to calculate Sortseq (bind). For both Sortseq (bind) and Sortseq (expr), we analyzed only single mutant data to avoid any artifacts that might arise from the epistatic model and took the average of delta Sortseq MFI (log(Sortseq (WT)) – log(Sortseq (mutant))) of mutants which had multiple barcodes. The Sortseq MFI values of mutants were averaged between the two libraries and the antilog was calculated for delta Sortseq MFI to analyse the ratio of Sortseq (bind) or Sortseq (expr) of mutants with respect to WT.

Acknowledgements

This work was funded in part by a grant to RV from the Department of Science and Technology, grant number-EMR/2017/004054, DT.15/12/2018), Government of India, Department of Biotechnology, grant no. BT/COE/34/SP15219/2015 DT. 20/11/2015, Ministry of Science and Technology, Government of India and Bill and Melinda Gates Foundation (USA). We also acknowledge funding for infrastructural support from the following programs of the Government of India: DST FIST, UGC Centre for Advanced study, Ministry of Human Resource Development (MHRD), and the DBT IISc Partnership Program. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. S.A. acknowledges Council of Scientific & Industrial Research for his fellowship (SPM-07/079(0218)/2015-EMR-I). K.M. is thankful to Department of Science and Technology (DST) Science and Engineering Research Board for financial support, sanction order no: PDF/2017/002641. Aparna Ashok is duly acknowledged for FACS.

Author contribution

R.V. and S.A. designed the experiments. S.A. performed all the experiments, R.V. and S.A. analyzed all the data. K.M. wrote the software and carried out the processing of the deep sequencing data. R.V. and S.A. wrote most of the manuscript.

Footnotes

The authors claim no conflict of interest.

References

[1] P.C. Jain, R. Varadarajan, A rapid, efficient, and economical inverse polymerase chain reaction-based method for generating a site saturation mutant library., Anal. Biochem. 449 (2014) 90–8.

[2] L. Zheng, U. Baumann, J.-L. Reymond, An efficient one-step site-directed and site-saturation mutagenesis protocol, Nucleic Acids Res. 32 (2004) e115.

[3] E.E. Wrenbeck, J.R. Klesmith, J.A. Stapleton, A. Adeniran, K.E.J. Tyo, T.A. Whitehead, Plasmid-based one-pot saturation mutagenesis., Nat. Methods. 13 (2016) 928–930.

[4] B. V. Adkar, A. Tripathi, A. Sahoo, K. Bajaj, D. Goswami, P. Chakrabarti, M.K. Swarnkar, R.S. Gokhale, R. Varadarajan, Protein model discrimination using mutational sensitivity derived from deep sequencing, Structure. 20 (2012) 371–381.

[5] D.M. Fowler, C.L. Araya, S.J. Fleishman, E.H. Kellogg, J.J. Stephany, D. Baker, S. Fields, High-resolution mapping of protein sequence-function relationships, Nat. Methods. 7 (2010) 741–746.

[6] K.A. Matreyek, L.M. Starita, J.J. Stephany, B. Martin, M.A. Chiasson, V.E. Gray, M. Kircher, A. Khechaduri, J.N. Dines, R.J. Hause, S. Bhatia, W.E. Evans, M. V. Relling, W. Yang, J. Shendure, D.M. Fowler, Multiplex assessment of protein variant abundance by massively parallel sequencing, Nat. Genet. 50 (2018) 874–882.

[7] M. Bhasin, R. Varadarajan, Prediction of Function Determining and Buried Residues Through Analysis of Saturation Mutagenesis Datasets., Front. Mol. Biosci. 8 (2021) 635425.

[8] L.L. Jones, S.E. Brophy, A.J. Bankovich, L.A. Colf, N.A. Hanick, K.C. Garcia, D.M. Kranz, Engineering and characterization of a stabilized alpha1/alpha2 module of the class I major histocompatibility complex product Ld., J. Biol. Chem. 281 (2006) 25734–44.

[9] R.L. Schweickhardt, X. Jiang, L.M. Garone, W.H. Brondyk, Structure-expression relationship of tumor necrosis factor receptor mutants that increase expression., J. Biol. Chem. 278 (2003) 28961–28967.

[10] E. Shusta, L. Pepper, Y. Cho, E. Boder, A Decade of Yeast Surface Display Technology: Where Are We Now?, Comb. Chem. High Throughput Screen. 11 (2008) 127–134.

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

[12] E. V. Shusta, M.C. Kieke, E. Parke, D.M. Kranz, K.D. Wittrup, Yeast polypeptide fusion surface display levels predict thermal stability and soluble secretion efficiency, J. Mol. Biol. 292 (1999) 949–956.

[13] Y. Hagihara, P.S. Kim, Toward development of a screen to identify randomly encoded, foldable sequences, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 6619–6624.

[14] S. Park, Y. Xu, X.F. Stowell, F. Gai, J.G. Saven, E.T. Boder, Limitations of yeast surface display in engineering proteins of high thermostability., Protein Eng. Des. Sel. 19 (2006) 211–217.

[15] A. Piatesi, S.W. Howland, J.A. Rakestraw, C. Renner, N. Robson, J. Cebon, E. Maraskovsky, G. Ritter, L. Old, K.D. Wittrup, Directed evolution for improved secretion of cancer-testis antigen NY-ESO-1 from yeast, Protein Expr. Purif. 48 (2006) 232–242.

[16] A. Chevalier, D.A. Silva, G.J. Rocklin, D.R. Hicks, R. Vergara, P. Murapa, S.M. Bernard, L. Zhang, K.H. Lam, G. Yao, C.D. Bahl, S.I. Miyashita, I. Goreshnik, J.T. Fuller, M.T. Koday, C.M. Jenkins, T. Colvin, L. Carter, A. Bohn, C.M. Bryan, D.A. Fernández-Velasco, L. Stewart, M. Dong, X. Huang, R. Jin, I.A. Wilson, D.H. Fuller, D. Baker, Massively

parallel de novo protein design for targeted therapeutics, Nature. 550 (2017) 74–79.

[17] G.J. Rocklin, T.M. Chidyausiku, I. Goreshnik, A. Ford, S. Houliston, A. Lemak, L. Carter, R. Ravichandran, V.K. Mulligan, A. Chevalier, C.H. Arrowsmith, D. Baker, Global analysis of protein folding using massively parallel design, synthesis, and testing., Science. 357 (2017) 168–175.

[18] B. Basanta, M.J. Bick, A.K. Bera, C. Norn, C.M. Chow, L.P. Carter, I. Goreshnik, F. Dimaio, D. Baker, An enumerative algorithm for de novo design of proteins with diverse pocket structures, Proc. Natl. Acad. Sci. U. S. A. 117 (2020) 22135–22145.

[19] J. Dou, A.A. Vorobieva, W. Sheffler, L.A. Doyle, H. Park, M.J. Bick, B. Mao, G.W. Foight, M.Y. Lee, L.A. Gagnon, L. Carter, B. Sankaran, S. Ovchinnikov, E. Marcos, P.-S. Huang, J.C. Vaughan, B.L. Stoddard, D. Baker, De novo design of a fluorescence-activating β-barrel., Nature. 561 (2018) 485–491.

[20] M.W. Traxlmayr, E. V. Shusta, Directed evolution of protein thermal stability using yeast surface display, in: Methods Mol. Biol., Humana Press, New York, NY, 2017: pp. 45–65.

[21] P. Bernard, M. Couturier, Cell killing by the F plasmid CcdB protein involves poisoning of DNA-topoisomerase II complexes, J. Mol. Biol. 226 (1992) 735–745.

[22] K. Bajaj, P.C. Dewan, P. Chakrabarti, D. Goswami, B. Barua, C. Baliga, R. Varadarajan, Structural correlates of the temperature sensitive phenotype derived from saturation mutagenesis studies of CcdB, Biochemistry. 47 (2008) 12964–12973.

[23] A. Tripathi, K. Gupta, S. Khare, P.C. Jain, S. Patel, P. Kumar, A.J. Pulianmackal, N. Aghera, R. Varadarajan, Molecular Determinants of Mutant Phenotypes, Inferred from Saturation Mutagenesis Data, Mol. Biol. Evol. 33 (2016) 2960–2975.

[24] G. Chao, W.L. Lau, B.J. Hackel, S.L. Sazinsky, S.M. Lippow, K.D. Wittrup, Isolating and engineering human antibodies using yeast surface display, Nat. Protoc. 1 (2006) 755–768.

[25] A. Sahoo, S. Khare, S. Devanarayanan, P.C. Jain, R. Varadarajan, Residue proximity information and protein model discrimination using saturation-suppressor mutagenesis, Elife. 4 (2015) e09532.

[26] T.N. Starr, A.J. Greaney, S.K. Hilton, D. Ellis, K.H.D. Crawford, A.S. Dingens, M.J. Navarro, J.E. Bowen, M.A. Tortorici, A.C. Walls, N.P. King, D. Veesler, J.D. Bloom, Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding, Cell. 182 (2020) 1295-1310.e20.

[27] L.R. Pepper, Y.K. Cho, E.T. Boder, E. V Shusta, A decade of yeast surface display technology: where are we now?, Comb. Chem. High Throughput Screen. 11 (2008) 127–134.

[28] E. Sharon, Y. Kalma, A. Sharp, T. Raveh-Sadka, M. Levo, D. Zeevi, L. Keren, Z. Yakhini, A. Weinberger, E. Segal, Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters, Nat. Biotechnol. 30 (2012) 521–530.

[29] N. Peterman, E. Levine, Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations., BMC Genomics. 17 (2016) 206.

[30] G. Cambray, J.C. Guimaraes, A.P. Arkin, Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli., Nat. Biotechnol. 36 (2018) 1005–1015.

[31] W.L. Noderer, R.J. Flockhart, A. Bhaduri, A.J. Diaz de Arce, J. Zhang, P.A. Khavari, C.L. Wang, Quantitative analysis of mammalian translation initiation sites by FACS-seq., Mol. Syst. Biol. 10 (2014) 748.

[32] S. Chakravarty, R. Varadarajan, Residue depth: a novel parameter for the analysis of protein structure and stability., Structure. 7 (1999) 723–32.

[33] F. Pucci, J.M. Kwasigroch, M. Rooman, Protein Thermal Stability Engineering Using HoTMuSiC., Methods Mol. Biol. 2112 (2020) 59–73.

[34] J. Chen, Z. Lu, J. Sakon, W.E. Stites, Increasing the thermostability of staphylococcal nuclease: Implications for the origin of protein thermostability, J. Mol. Biol. 303 (2000) 125–130.

[35] R.S. Prajapati, M. Das, S. Sreeramulu, M. Sirajuddin, S. Srinivasan, V. Krishnamurthy, R. Ranjani, C. Ramakrishnan, R. Varadarajan, Thermodynamic effects of proline introduction on protein stability, Proteins Struct. Funct. Genet. 66 (2007) 480–491.

[36] A.P. Pandurangan, B. Ochoa-Montaño, D.B. Ascher, T.L. Blundell, SDM: A server for predicting effects of mutations on protein stability, Nucleic Acids Res. 45 (2017) W229–W235.

[37] D.E.V. Pires, D.B. Ascher, T.L. Blundell, MCSM: Predicting the effects of mutations in proteins using graph-based signatures, Bioinformatics. 30 (2014) 335–342.

[38] Y. Dehouck, J.M. Kwasigroch, D. Gilis, M. Rooman, PoPMuSiC 2.1: A web server for the estimation of protein stability changes upon mutation and sequence optimality, BMC Bioinformatics. 12 (2011) 151.

[39] C.H.M. Rodrigues, D.E.V. Pires, D.B. Ascher, DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability, Nucleic Acids Res. 46 (2018) W350–W355.

[40] D.E.V. Pires, D.B. Ascher, T.L. Blundell, DUET: A server for predicting effects of mutations on protein stability using an integrated computational approach, Nucleic Acids Res. 42 (2014) W314–W319.

[41] J. Laimer, J. Hiebl-Flach, D. Lengauer, P. Lackner, MAESTROweb: a web server for structure-based protein stability prediction, Bioinformatics. 32 (2016) 1414–1416.

[42] H. Cao, J. Wang, L. He, Y. Qi, J.Z. Zhang, DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks, J. Chem. Inf. Model. 59 (2019) 1508–1514.

[43] V. Parthiban, M.M. Gromiha, D. Schomburg, CUPSAT: Prediction of protein stability upon point mutations, Nucleic Acids Res. 34 (2006) W239–W242.

[44] Y. Chen, H. Lu, N. Zhang, Z. Zhu, S. Wang, M. Li, PremPS: Predicting the impact of missense mutations on protein stability, PLOS Comput. Biol. 16 (2020) e1008543.

[45] C. Savojardo, P. Fariselli, P.L. Martelli, R. Casadio, INPS-MD: a web server to predict stability of protein variants from sequence and structure: Table 1., Bioinformatics. 32 (2016) 2542–2544.

[46] B.E. Jones, P.L. Brown-Augsburger, K.S. Corbett, K. Westendorf, J. Davies, T.P. Cujec, C.M. Wiethoff, J.L. Blackbourne, B.A. Heinz, D. Foster, R.E. Higgs, D. Balasubramaniam, L. Wang, R. Bidshahri, L. Kraft, Y. Hwang, S. Žentelis, K.R. Jepson, R. Goya, M.A. Smith, D.W. Collins, S.J. Hinshaw, S.A. Tycho, D. Pellacani, P. Xiang, K. Muthuraman, S. Sobhanifar, M.H. Piper, F.J. Triana, J. Hendle, A. Pustilnik, A.C. Adams, S.J. Berens, R.S. Baric, D.R. Martinez, R.W. Cross, T.W. Geisbert, V. Borisevich, O. Abiona, H.M. Belli, M. de Vries, A. Mohamed, M. Dittmann, M. Samanovic, M.J. Mulligan, J.A. Goldsmith, C.-L. Hsieh, N. V Johnson, D. Wrapp, J.S. McLellan, B.C. Barnhart, B.S. Graham, J.R. Mascola, C.L. Hansen, E. Falconer, LY-CoV555, a rapidly isolated potent neutralizing antibody, provides protection in a non-human primate model of SARS-CoV-2 infection., BioRxiv Prepr. Serv. Biol. (2020). https://doi.org/10.1101/2020.09.30.318972.

[47] S.K. Malladi, R. Singh, S. Pandey, S. Gayathri, K. Kanjo, S. Ahmed, M.S. Khan, P. Kalita, N. Girish, A. Upadhyaya, P. Reddy, I. Pramanick, M. Bhasin, S. Mani, S.

Bhattacharyya, J. Joseph, K. Thankamani, V.S. Raj, S. Dutta, R. Singh, G. Nadig, R. Varadarajan, Design of a highly thermotolerant, immunogenic SARS-CoV-2 spike fragment., J. Biol. Chem. 296 (2020) 100025.

[48] T. Xue, W. Wu, N. Guo, C. Wu, J. Huang, L. Lai, H. Liu, Y. Li, T. Wang, Y. Wang, Single point mutations can potentially enhance infectivity of SARS-CoV-2 revealed by in silico affinity maturation and SPR assay, BioRxiv. (2020). https://doi.org/10.1101/2020.12.24.424245.

[49] J. Zahradník, S. Marciano, M. Shemesh, E. Zoler, J. Chiaravalli, B. Meyer, O. Dym, N. Elad, G. Schreiber, SARS-CoV-2 RBD in vitro evolution follows contagious mutation spread, yet generates an able infection inhibitor, BioRxiv. (2021). https://doi.org/10.1101/2021.01.06.425392.

[50] Y. Weisblum, F. Schmidt, F. Zhang, J. DaSilva, D. Poston, J.C.C. Lorenzi, F. Muecksch, M. Rutkowska, H.-H. Hoffmann, E. Michailidis, C. Gaebler, M. Agudelo, A. Cho, Z. Wang, A. Gazumyan, M. Cipolla, L. Luchsinger, C.D. Hillyer, M. Caskey, D.F. Robbiani, C.M. Rice, M.C. Nussenzweig, T. Hatziioannou, P.D. Bieniasz, Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants., Elife. 9 (2020) e61312.

[51] U.T. Bornscheuer, B. Hauer, K.E. Jaeger, U. Schwaneberg, Directed Evolution Empowered Redesign of Natural Proteins for the Sustainable Production of Chemicals and Pharmaceuticals, Angew. Chemie Int. Ed. 58 (2019) 36–40.

[52] K. Chen, F.H. Arnold, Enzyme engineering for nonaqueous solvents: Random mutagenesis to enhance activity of subtilisin E in Polar Organic Media, Bio/Technology. 9 (1991) 1073–1077.

[53] G. Winter, A.D. Griffiths, R.E. Hawkins, H.R. Hoogenboom, Making antibodies by phage display technology, Annu. Rev. Immunol. 12 (1994) 433–455.

[54] K.L. Maxwell, A.K. Mittermaier, J.D. Forman-Kay, A.R. Davidson, A simple in vivo assay for increased protein solubility, Protein Sci. 8 (1999) 1908–1911.

[55] W.C. Wigley, R.D. Stidham, N.M. Smith, J.F. Hunt, P.J. Thomas, Protein solubility and folding monitored in vivo by structural complementation of a genetic marker protein, Nat. Biotechnol. 19 (2001) 131–136.

[56] A.C. Fisher, Genetic selection for protein solubility enabled by the folding quality control feature of the twin-arginine translocation pathway, Protein Sci. 15 (2006) 449–458.

[57] G.S. Waldo, B.M. Standish, J. Berendzen, T.C. Terwilliger, Rapid protein-folding assay using green fluorescent protein, Nat. Biotechnol. 17 (1999) 691–695.

[58] M.W. Traxlmayr, C. Obinger, Directed evolution of proteins for increased stability and expression using yeast display, Arch. Biochem. Biophys. 526 (2012) 174–180.

[59] O. Esteban, H. Zhao, Directed evolution of soluble single-chain human class II MHC molecules, J. Mol. Biol. 340 (2004) 81–95.

[60] Y.-S. Kim, R. Bhandari, J.R. Cochran, J. Kuriyan, K.D. Wittrup, Directed evolution of the epidermal growth factor receptor extracellular domain for expression in yeast., Proteins. 62 (2006) 1026–1035.

[61] M.C. Kieke, E. V Shusta, E.T. Boder, L. Teyton, K.D. Wittrup, D.M. Kranz, Selection of functional T cell receptor mutants from a yeast surface-display library., Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 5651–5656.

[62] S. Park, Y. Xu, X.F. Stowell, F. Gai, J.G. Saven, E.T. Boder, Limitations of yeast surface display in engineering proteins of high thermostability, Protein Eng. Des. Sel. 19 (2006) 211–217.

[63] S. Khare, M. Bhasin, A. Sahoo, R. Varadarajan, Protein model discrimination attempts

using mutational sensitivity, predicted secondary structure, and model quality information, Proteins Struct. Funct. Bioinforma. 87 (2019) 326–336.

[64] E.M. Jones, N.B. Lubock, A.J. Venkatakrishnan, J. Wang, A.M. Tseng, J.M. Paggi, N.R. Latorraca, D. Cancilla, M. Satyadi, J.E. Davis, M.M. Babu, R.O. Dror, S. Kosuri, Structural and functional characterization of G protein-coupled receptors with deep mutational scanning., Elife. 9 (2020) e61312.

[65] B.J. Livesey, J.A. Marsh, Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations., Mol. Syst. Biol. 16 (2020) e9380.

[66] G.M. Findlay, R.M. Daza, B. Martin, M.D. Zhang, A.P. Leith, M. Gasperini, J.D. Janizek, X. Huang, L.M. Starita, J. Shendure, Accurate classification of BRCA1 variants with saturation genome editing, Nature. 562 (2018) 217–222.

[67] K. Bajaj, P.C. Dewan, P. Chakrabarti, D. Goswami, B. Barua, C. Baliga, R. Varadarajan, Structural correlates of the temperature sensitive phenotype derived from saturation mutagenesis studies of CcdB, Biochemistry. 47 (2008) 12964–12973.

[68] E.T. Boder, K.D. Wittrup, Yeast surface display for directed evolution of protein expression, affinity, and stability, Meth. Enzym. 328 (2000) 430–444.

[69] T.A. Najar, S. Khare, R. Pandey, S.K. Gupta, R. Varadarajan, Mapping Protein Binding Sites and Conformational Epitopes Using Cysteine Labeling and Yeast Surface Display, Structure. 25 (2017) 395–406.

[70] R.D. Gietz, R.H. Schiestl, High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method, Nat. Protoc. 2 (2007) 31–34.

[71] G. Chattopadhyay, R. Varadarajan, Facile measurement of protein stability and folding kinetics using a nano differential scanning fluorimeter, Protein Sci. 28 (2019) 1127–1134.

[72] M.H. Dao-Thi, L. Van Melderen, E. De Genst, L. Buts, A. Ranquin, L. Wyns, R. Loris, Crystallization of CcdB in complex with a GyrA fragment, Acta Crystallogr. Sect. D Biol. Crystallogr. 60 (2004) 1132–1134.

[73] F.H. Niesen, H. Berglund, M. Vedadi, The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability, Nat. Protoc. 2 (2007) 2212–2221.

[74] J. Zhang, K. Kobert, T. Flouri, A. Stamatakis, PEAR: A fast and accurate Illumina Paired-End reAd mergeR, Bioinformatics. 30 (2014) 614–620.

[75] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.

[76] K.P. Tan, R. Varadarajan, M.S. Madhusudhan, DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins., Nucleic Acids Res. 39 (2011) W242–W248.

[77] T.J. Hubbard SJ, NACCESS, Dep. Biochem. Mol. Biol. Univ. Coll. London. (1993).

[78] N.K. Aghera, J. Prabha, H. Tandon, G. Chattopadhyay, S. Vishwanath, N. Srinivasan, R. Varadarajan, Mechanism of CcdA-Mediated Rejuvenation of DNA Gyrase., Structure. 28 (2020) 562-572.e4.

Figures

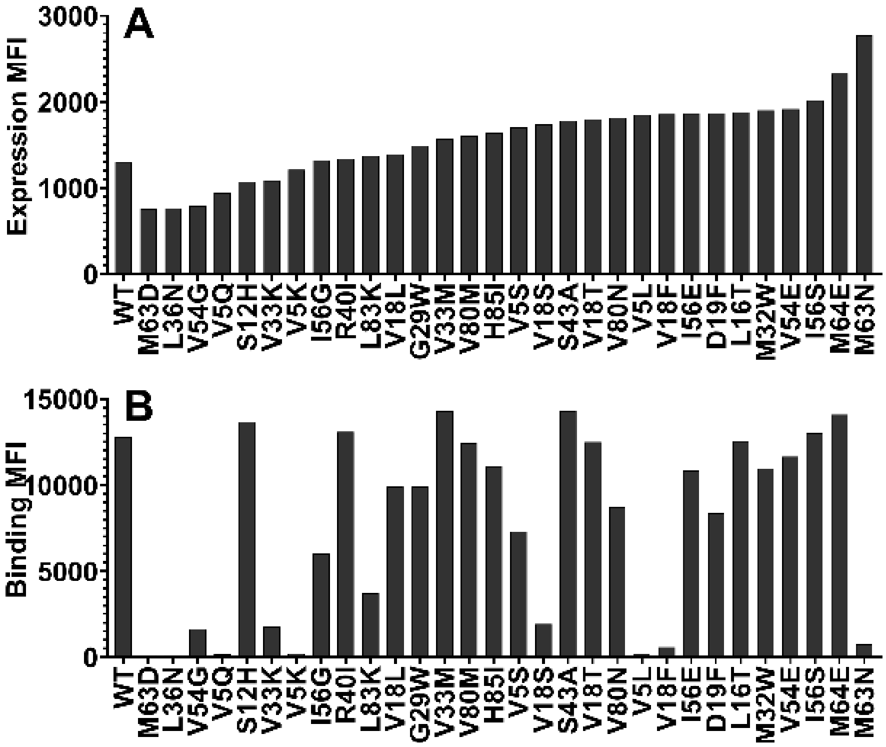

Figure 1: Comparison of the level of expression and binding of CcdB mutants on the yeast cell surface. (A) The expression and (B) binding to GyrA14 of individual mutants. Most mutants expressed at high levels, however, the amount of active protein varied widely. A few mutants which showed a high level of expression did not show any binding to GyrA14. In both panels, mutants are arranged in order of increasing expression level.

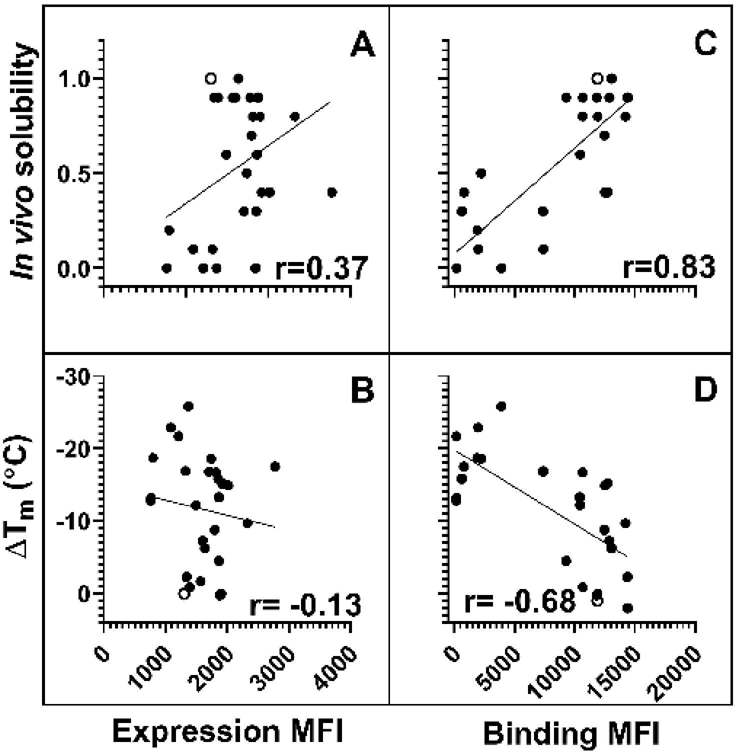

Figure 2: Correlation of *E.coli in vivo* solubility and *in vitro* thermal stability with the amount of total and active protein on the yeast cell surface. For individual mutants, MFI's of expression and binding were estimated by probing the HA tag on surface expressed protein and the FLAG tag on cell surface bound GyrA14 respectively. Correlation of the total amount of protein displayed on the yeast cell surface with (A) *in vivo* solubility or (B) ΔT_m (T_m (mutant)- T_m (WT)) of CcdB mutants. Correlation of the amount of active protein on the yeast cell surface with (C) *E.coli in vivo* solubility or (D) ΔT_m of CcdB mutants. A better correlation was observed between biophysical parameters with binding MFI rather than expression MFI. In the figure, the ΔT_m of WT was increased by 1°C to remove overlap with another point. Data for *E.coli in vivo* solubility and thermal stability was taken from Tripathi et al [23]. WT data is shown in open circles.

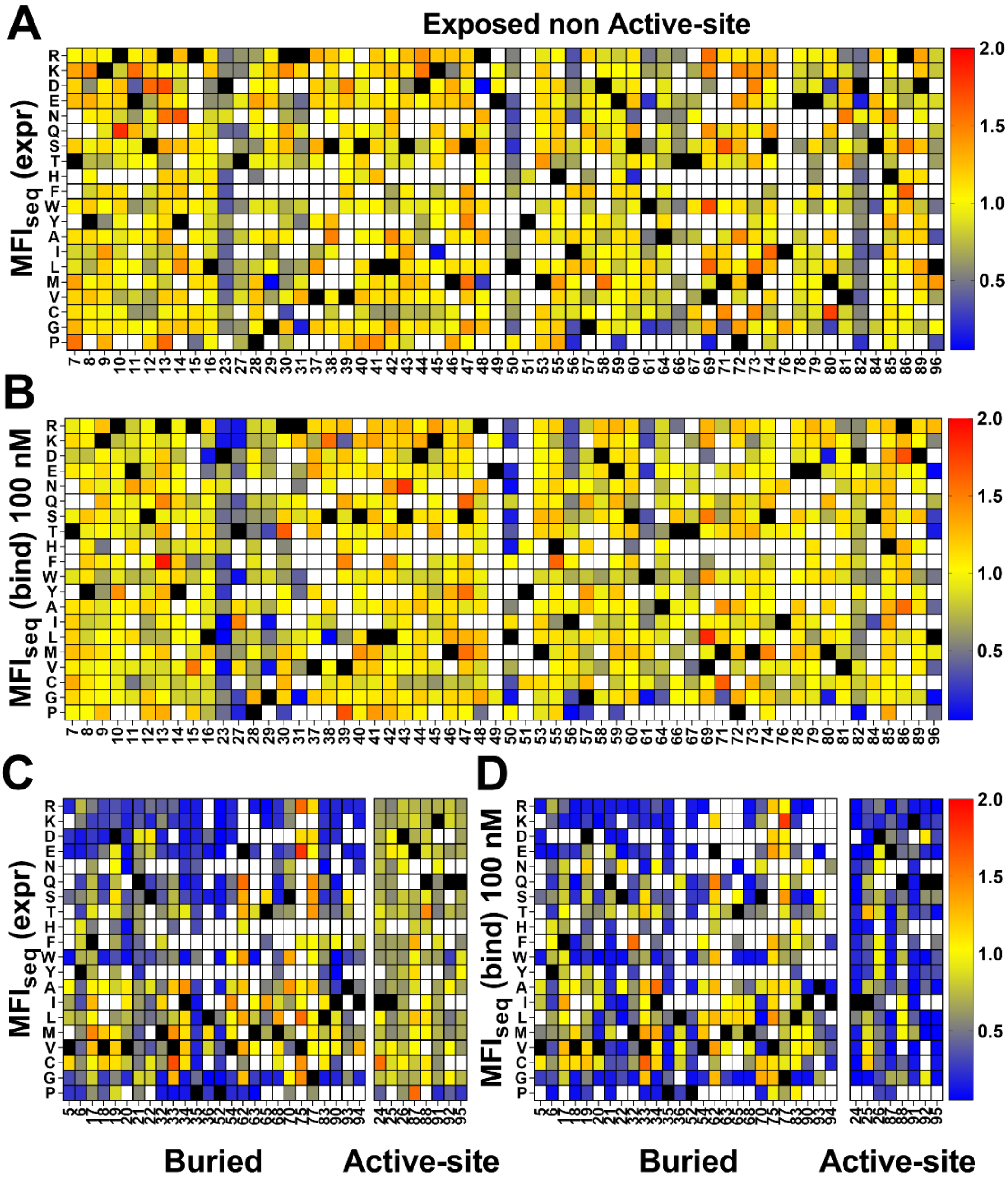

Figure 3: Heatmap of normalized MFI$_{seq}$ values for CcdB mutants. MFI$_{seq}$ value of mutant was divided by the MFI$_{seq}$ value of WT to normalize it. (A) MFI$_{seq}$ (expr) and (B) MFI$_{seq}$ (bind) at 100 nM GyrA14 for exposed non active-site residues. (C) MFI$_{seq}$ (expr) and (D) MFI$_{seq}$ (bind) for buried and active-site residues. Exposed, buried (PDB ID:3VUB) and active-site (PDB ID:1X75) residues are segregated based on the crystal structure. Residues which had accessibility greater than 10% were considered exposed, all remaining residues were considered buried, and active-site mutants in contact with GyrA14 were identified as explained

the Methods section. Blue to red colour represents increasing normalized MFI_{seq} values, black colour shows the WT residue at the corresponding position. White colour indicates that the mutant is not available. The buried site residues have very high mutational sensitivity both in case of expression and binding. The active-site residues show mutational sensitivity only with respect to Gyrase binding. Information about the mutational sensitivity of expression and binding can be used to differentiate exposed, buried and active-site residues.

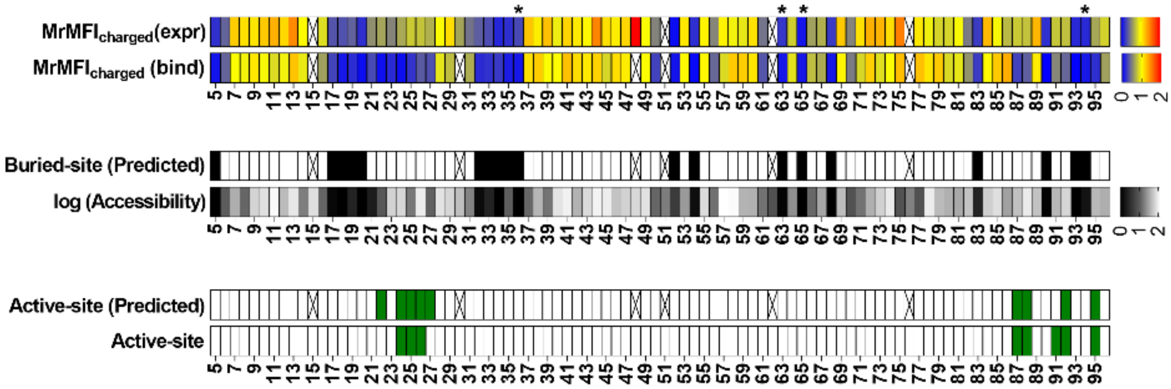

Figure 4: Identification of buried and active-site residues from MrMFI$_{\text{charged}}$ (bind) and MrMFI$_{\text{charged}}$ (expr). Side chain accessibilities in dimeric CcdB (PDB: 3VUB), darker to lighter shade indicate increasing accessibility, accessibility is reported as log accessibility. the mutants were clustered into two bins based on the distribution of MrMFI$_{\text{charged}}$ and k-means and standard deviations were calculated for both distributions. The distributions were named D1 (higher MrMFI$_{\text{charged}}$) and D2 (lower MrMFI$_{\text{charged}}$). Residues which had MrMFI$_{\text{charged}}$ (binding) and MrMFI$_{\text{charged}}$ (expr) lower than (μ+0.5*σ) of distribution D2 were characterized as buried. The false negatives were Y6, D19, Q21, S22, S70, V75 and G77, the polar side chains of these residues are pointing towards the surface. Active-site residues were identified as those in contact with GyyrA14 (PDB ID 1X75). Residues which had MrMFI$_{\text{charged}}$ (binding) less than (μ+σ) of D2 distribution and MrMFI$_{\text{charged}}$ (expr) higher than (μ-2*σ) of distribution D1 were predicted as active-site. We obtained a few putative false positives. However, these residues are likely involved in functional aspects of activity that cannot be inferred from the CcdB:GyrA14 crystal structure. The same residues were seen to be important for CcdB activity *in vivo* in *E.coli* [23]. Some positions could not be categorized due to lack of reads, such positions are indicated with an 'X'. Positions indicated with '*' are the one where MrMFI$_{\text{charged}}$ (expr) was observed and the mutants had high read counts but the mutants were absent in MrMFI$_{\text{charged}}$ (bind), such positions were assigned MrMFI$_{\text{charged}}$ (bind) values similar to other buried positions.

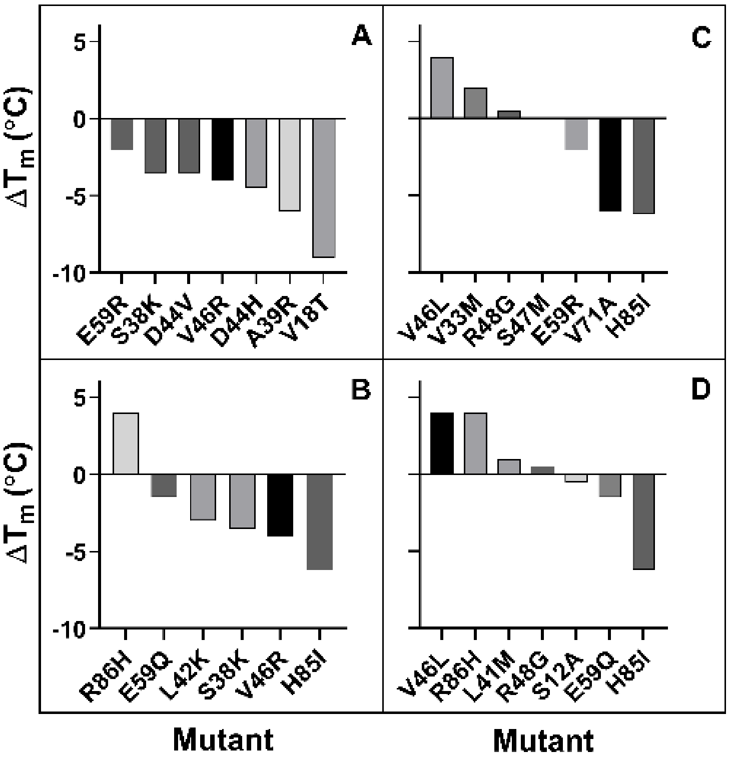

Figure 5: ΔT_m of putative stabilized CcdB mutants. Mutants were identified from A) MFI_{seq} (bind) at 5 nM GyrA14, (B) MFI_{seq} (ratio) at 5 nM GyrA14, (C) MFI_{seq} (bind) at 100 nM GyrA14, (D) MFI_{seq} (ratio) at 100 nM GyrA14.

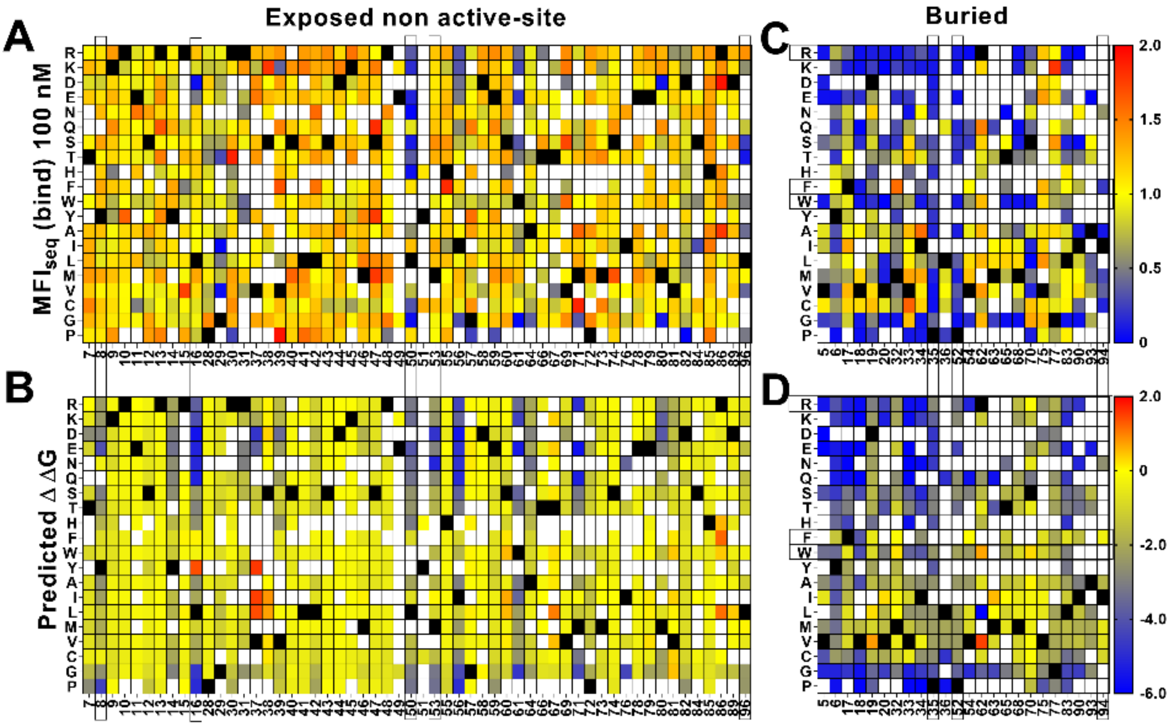

Figure 6: Comparison of stabilities estimated by DeepDDG and yeast surface display. Heat maps for (A,C) MFI$_{seq}$ (bind) normalized to WT and (B, D) ΔΔG predicted by DeepDDG. Residue positions or specific amino acid mutations showing significantly different predicted stabilities by the two methods are highlighted by a box. Blue to red colour corresponds to increasing stability.

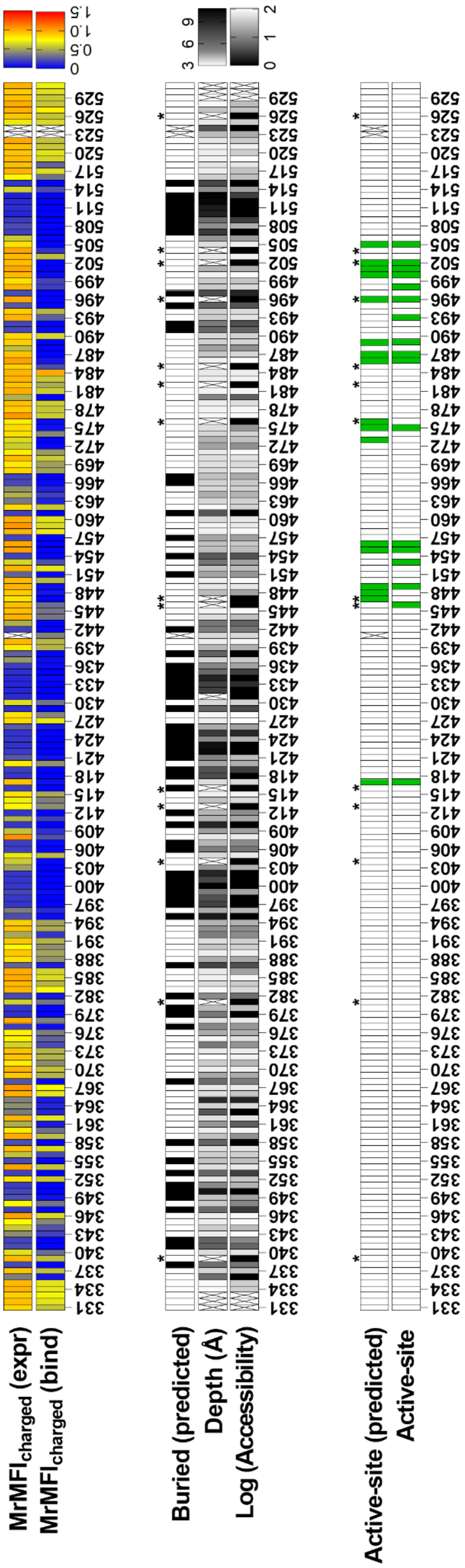

Figure 7: Prediction of buried and active-site positions in SARS-CoV-2 RBD from Sortseq data. Buried residues were identified from chain C of PDB ID 7KMH, residues which

had <10% side chain accessibility were categorized as buried. The accessibility and depth was calculated using DEPTH server [76]. Active-site residues were identified from PDB ID 6M0J as explained earlier [47]. Criteria used to predict buried and active-site positions from MFI data were identical to those used for CcdB. Positions which did not have MrMFI data or could not be assigned to either buried or active-site categories are highlighted with “X”. Accessibility calculated by DEPTH server for glycine is zero and these are marked with a ‘*’.