A dual-reporter system for investigating and optimizing protein translation and folding in *E. coli*

Ariane Zutz[1,3], Louise Hamborg Nielsen[1,2], Lasse Ebdrup Pedersen[1], Maher M. Kassem[2], Elena Papaleo[2], Anna Koza[1], Markus J. Herrgård[1], Kaare Teilum[2], Kresten Lindorff-Larsen[2], Alex Toftgaard Nielsen[1#]

[1]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemitorvet, 2800 Kgs. Lyngby, Denmark

[2]Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen N, Denmark

[3]Department of Virology, Max-von-Pettenkofer-Institute, LMU, Pettenkofer-Straße 9A, 80336 Munich, Germany

[#]Corresponding author

Phone: +45 45258010

Postal address: Building 220, Kemitorvet, 2800 Kgs. Lyngby, Denmark

Email: atn@biosustain.dtu.dk

Abstract

Strategies for investigating and optimising the expression and folding of proteins for biotechnological and pharmaceutical purposes are in high demand. Here, we describe a dual-reporter biosensor system that simultaneously assesses *in vivo* protein translation and protein folding, thereby enabling rapid screening of mutant libraries. We have validated the dual-reporter system on five different proteins and find an excellent correlation between reporter intensity signals and the levels of protein expression and solubility of the proteins. We further demonstrate the applicability of the dual-reporter system as a screening assay for deep mutational scanning experiments. The system enables high throughput selection of protein variants with high expression levels and altered protein stability. Next generation sequencing analysis of the resulting libraries of protein variants show a good correlation between computationally predicted and experimentally determined protein stabilities. We furthermore show that the mutational experimental data obtained using this system may be useful for protein structure calculations.

Introduction

Expression of heterologous proteins is essential for a number of purposes including functional and structural characterization, as well as for industrial production of enzymes and biochemicals through metabolic pathway engineering. However, heterologous expression of recombinant proteins in bacteria such as *E. coli* often results in misfolding, aggregation and degradation of the proteins. Therefore, it is of significant importance to improve proteins for efficient expression.

Several strategies for improving expression and folding of heterologous proteins are known, including for example screening and optimisation of environmental factors such as host strain, growth medium and temperature, induction parameters and co-expression of folding chaperones[1]. Other strategies involve the use of protein affinity and solubility tags, which are short peptide or protein tags fused to the N- or C-terminus of a protein. Solubility tags function as folding scaffolds thereby helping to improve translation and folding of proteins with poor folding properties[2]. A small affinity tag is less likely to interfere with the three-dimensional structure of the protein[3], and it has the advantage that it can be used for affinity purification and for detection and quantification by Western blotting.

Another strategy for improving protein expression and folding involves optimisation of the expression plasmid and gene of interest. The natural variation in codon usage often reflects changes in translation speed needed for correct co-translational protein folding.[4] Changes in codon usage or expression host may lead to changes in the translation rate, cause mis-incorporation of amino acid residues and truncations of the protein due to premature termination of translation.[5–8] Structured parts of a protein may demand slower translation to enable co-translational folding, while more unstructured parts allow for more rapid translation.[4] Other factors to optimize include the choice of promoter, different mRNA secondary structures, optimal open reading frames, and avoidance of certain amino acid residues in the N- or C-terminal of the protein as they can be susceptible for proteolysis or prevent initiation of the translation process.[9,10] Furthermore, hydrophobic parts of the polypeptide chain are more prone to aggregation[11], and truncation of unstructured hydrophobic parts

of the protein may thus improve protein folding. Additionally, computational methods can be used to predict protein variants with optimized improved folding and expression properties [12].

A more efficacious way to improve protein expression would be to screen large random mutant libraries for variants of a protein with optimized folding. However, generation of random mutant libraries often results in frequent frame shift mutations and stop codons. When screening for mutants with improved folding, it is therefore necessary to exclude the large number of clones that no longer express the target protein or form aggregates. It would thus be desirable to screen simultaneously for folding and expression. Many such methods for the analysis of protein expression and folding require extraction of proteins from the production organism, separating the proteins into soluble (folded) and insoluble fractions, and analysing these fractions using SDS-PAGE or dot-blot based technologies[13–15]. These time-consuming processes are not amenable for screening of larger libraries of production organisms or protein variants at the single-cell level. Several bacterial systems have been developed for testing and screening variants for expression or stability. Examples include fusion reporter proteins for assessing protein folding and solubility using fluorescence, enzymatic reactions, antibiotic resistance or ligand binding as reporters for the production of soluble and folded proteins.[16–20] However, no system enables simultaneous monitoring of translation and folding at the single cell level.

Although proteins are generally able to fold into their native conformation by themselves, most organisms have evolved mechanisms for controlling and aiding the process, and preventing unproductive misfolding. Molecular chaperones are constitutively expressed and participate in *de novo* protein folding by stabilizing the nascent polypeptide chain on the ribosome, in protein

trafficking and domain assembly, and assist in degradation of partially folded and aggregated proteins. Several bacterial chaperones are induced when misfolded protein is expressed in the cell and their promotors may be used to drive stress induced heterologous protein expression[21]. Thus, chaperone promoters can be used to construct reporters for the presence of e.g. misfolded protein. Previous work has focused on coupling such promoters to the expression of luciferase[22] or beta-galactosidase[23], which both require chemical assays for assessing their activity. Several methods are available for monitoring the level of protein production, such as it has been demonstrated using a translation-coupling system in *E. coli*[24]. None of the current methods, however, are suited for high-throughput approaches with simultaneous but independent *in vivo* monitoring of both translation and protein folding.

Here, we demonstrate a functional dual reporter system that enables single-cell monitoring of both protein translation levels and the degree of protein misfolding. The system can be used to analyse translation levels and folding properties of heterologously expressed proteins in *E.* coli. We demonstrate the use of the system for screening the expression levels of various proteins including the effect of different solubility tags. We further show how the system can be combined with fluorescence activated cell sorting (FACS) and next generation sequencing (NGS) in a deep mutational scanning experiment[25] for generating protein wide identification of mutations important for correct protein translation and folding. We find a good correlation between computationally calculated protein stability of mutant PARP1-BRCT proteins and experimental data. Furthermore, we show that the mutational experimental data obtained in this work can be used to select native-like structures

from a large pool of structures highlighting the usefulness of such systems in protein structure calculation.

Results

Dual-reporter system

To enable high throughput analysis, we have developed a dual-reporter system that simultaneously monitors protein translation and protein folding at the single cell level (Figure 1A). The translation sensor consists of a translation coupling cassette comprised of a strong secondary mRNA structure formed by a C-terminal hexa-histidine tag, a stop codon for the gene of interest and a ribosome binding site (RBS) for a downstream fluorescent reporter protein, mCherry. The cassette has been inserted into a modified pET22b plasmid containing the pBR322 origin of replication (Figure 1B). When the gene of interest is correctly translated the secondary mRNA structure will be unfolded by ribosomal helicase activity, and expose the RBS for the downstream reporter gene enabling RNA polymerase to continue transcription.[24] An untimely termination of the translation will hinder the ribosome reaching the position of the mCherry gene, thus preventing a fluorescent signal to emerge. Correct translation of the gene of interest results in mCherry being expressed in a one to one ratio with the protein of interest.

The protein folding sensor is based on the naturally occurring heat shock response system in *E. coli*. Heat shock proteins (HSPs) are expressed to protect the cell when exposed to high temperatures or other form of stress conditions. HSPs are often molecular chaperones that bind the hydrophobic parts of partially unfolded proteins and assist in refolding and protection against degradation by

proteases. In *E. coli*, the alternative sigma factor, called RpoH, controls the transcription of several cytoplasmic HSPs, including the small inclusion body heat shock proteins, IbpA/B. [26,27] In an unstressed cell RpoH is bound to chaperone DnaK, but during stress RpoH will be released when DnaK binds unfolded protein, thus increasing the level of free RpoH in the cell. RpoH then binds to the RNA polymerase sigma70, which subsequently recognizes heat shock promoters and thus initiate a heat shock response. In our protein folding sensor, the RpoH inducible IbpA promoter is inserted upstream of a GFP reporter gene in a modified pSEVA631 vector[28] with the medium-copy pBBR1 origin of replication (Figure 1B). When the dual reporter system is used, the formation of misfolded protein and inclusion bodies will initiate expression from the IbpA promoter resulting in the expression of GFP.

Effect of plasmid backbone and GFP variant on signal distribution

To test whether the copy number affects the IbpAp-GFP activity, we have analysed the heat shock response from two vector backbones with the pBBR1 or the ColE1 origin of replication (ORI), respectively. The two ORIs were further tested in combination with two GFP-variants, GFP-mut3 and GFP-ASV. GFP-mut3 is a stable GFP-variant with a half-life estimated to more than one day, while a C-terminal degradation tag makes GFP-ASV susceptible to protease degradation and results in a shorter half-life of about 110 min.[10]

Cell cultures in the exponential growth phase were exposed to 42°C for 10 minutes to induce a cellular heat shock response. The heat shock response was followed by monitoring the GFP fluorescence for two hours (Figure 2A). Immediately after the exposure to 42°C, a rapid increase in GFP expressed from plasmids with the pBBR1 ORI was observed reaching a maximal level after 20

minutes. The GFP signal was consistent with the expected change in RpoH synthesis rate observed during a heat shock induced response, where the formation of misfolded protein is known to initiate a spike in the RpoH synthesis rate, that slowly declines to a level higher than before the heat shock.[29] The heat shock promoter under the control of the ColE1 ORI plasmid did not give rise to a heat shock response signal.

Since differentiation between heat shock induced and un-induced GFP responses is crucial for the applicability of the folding sensor, single cell analysis was carried out. FACS profiles of the heat induced and un-induced ColE1 plasmids show broad overlapping peaks, indicating a leaky expression of GFP (Figure 2B). The accumulation of GFP in the cell made it impossible to monitor a heat shock response signal different from the basal GFP level using the ColE1 based plasmids. In contrast, two sharp well-defined peaks were observed from the heat induced and un-induced pBBR1 plasmids with a 3-5-fold increase in the signal-to noise ratio. A significant increase was observed in the signal-to-noise ratio for the GFP-ASV variant combined with the pBBR1 backbone. The highly stable GFP-mut3 slowly accumulates in the cell over time and results in higher GFP signals for both heat induced pBBR1 and the control, thereby resulting in a lower signal-to-noise ratio. The short half-life of GFP-ASV prevented the basal accumulation of GFP in the cell, which enabled the distinction of the heat shock induced response from protein misfolding in single cells.

To test the compatibility of the translation sensor and the protein folding sensor, we chose to analyse two human proteins with differences in expression levels and solubility in *E. coli*, PARP1-BRCT and BRCA1-BRCT. PARP1-BRCT and BRCA1-BRCT contain the BRCT domain of human Poly[ADP-ribose] polymerase 1 (PARP1) and human breast cancer 1, early onset (BRCA1), respectively. The BRCA1-BRCT

166 construct was designed to promote misfolding by making a truncation of the folded BRCT domain.
167 PARP1-BRCT was expressed in high yields as soluble protein in *E. coli* as shown by SDS-PAGE and
168 Western blot analyses, while the BRCA1-BRCT domain was expressed as an insoluble protein in *E. coli*
169 (Figure 2C).

170 Folding of PARP1-BRCT and BRCA1-BRCT was further analysed using pSEVA631(Sp)-IbpAp-GFP-
171 ASV and pSEVA631(Sp)-IbpAp-GFP-mut3 as protein folding sensors and monitored by flow cytometry.
172 As expected, the expression of the soluble PARP1-BRCT did not initiate a GFP response compared to
173 the control carrying only an empty pET22b vector (Figure 2D). Overexpression of the insoluble BRCA1-
174 BRCT, however, promoted binding of RpoH to the IbpA promoter region of the folding sensor,
175 resulting in a 5 to 10-fold increase in GFP-signal compared to PARP1-BRCT. As previously observed,
176 the GFP-ASV variant yielded a higher fluorescent signal, and a better signal-to-noise ratio compared to
177 GFP-mut3 (Figure 2D). These results demonstrate the applicability of the protein folding sensor as a
178 tool to monitor protein folding *in vivo*.

179 ***Effect of protein solubility tags on translation and folding***

180 Overexpression of recombinant proteins in *E. coli* often results in misfolded proteins and the
181 formation of insoluble aggregates. To enhance the solubility, fusion proteins are often linked to the N-
182 terminus of proteins that aggregates during expression. To test the applicability of the dual-reporter
183 system to monitor the effects of linking solubility tags, we investigated the fusion of two commonly
184 used expression tags, the N-utilization A (NusA) and the small ubiquitin related modifier (SUMO), on
185 the translation levels and solubility of four different model proteins. The proteins were chosen based
186 on their different translation levels and tendency to form inclusion bodies when expressed in *E. coli,*

and include PARP1-BRCT[30], a truncated variant of BRCA1-BRCT[31], the human cyclin-dependent kinase inhibitor, p19[32] and the viral oncogene E6 from human *papillomavirus type 16* [33]. Wild-type PARP1-BRCT, BRCA1-BRCT, E6, and p19 were cloned into the translation sensor with either NusA or SUMO linked to the N-terminus of the proteins. The translation and protein folding sensors were co-expressed at 30°C. Translation and protein folding were monitored by flow cytometry, while SDS-PAGE and Western blot analyses were used to compare the levels of expressed and soluble protein. We observed a strong correlation between translation levels monitored by flow cytometry and the expression yield detected by Western blot (Figure 3A). The N-terminally linked solubility tags did not have a large impact on the translation level of PARP1-BRCT, BRCA1-BRCT, and E6, which all expressed also without the tags, and the SUMO tag even decreased the expression level of BRCA1-BRCT. In contrast, under the given conditions wild type p19 did not express, however, fusion with either NusA or SUMO enabled expression, with NusA having a bigger effect.

The total amount of protein expressed and the fraction of soluble protein were quantified by Western blotting and compared to the GFP fluorescence from the protein folding sensor monitored by flow cytometry (Figure 3b). Expression of all PARP1-BRCT and the p19 variants resulted in background GFP fluorescence. In contrast BRCA1-BRCT and E6 expressed as insoluble aggregates and resulted in high GFP fluorescence. Different levels of GFP fluorescence were observed for BRCA1-BRCT and E6 although they both were expressed as insoluble protein. We also note that the fusion of the proteins to NusA or SUMO did not increase the amount of soluble protein, as also quantified by the folding sensor. The results demonstrate that it is possible to combine both biosensors to simultaneously investigate translation and proper folding of proteins in *E. coli.*

Quantitative determination of protein stability and protein misfolding

To test whether the folding sensor can be used to quantitatively measure protein stability and protein misfolding, six variants of the chymotrypsin inhibitor 2 (CI2) with different experimentally-determined thermodynamic stabilities (ΔG_U)[34] were cloned into the translation sensor vector. CI2 is a serine protease that has been extensively used as a model protein in protein folding and stability studies [34,35]. The protein variants were expressed at 30°C using pSEVA631(Sp)-IbpAp-GFP-ASV as protein folding sensor. The GFP fluorescence monitored by flow cytometry was compared to the *in vitro* stability of the His-tagged proteins at 30°C determined by global fitting of temperature and denaturant unfolding. The GFP fluorescence clearly changed with ΔG_U, where more stable proteins resulted in lower GFP signals (Figure 4). These results show that the GFP fluorescence arising from the protein folding sensor can be used as a proxy for the *in vitro* stability of variants in a mutant library. These results also suggest that the dual-reporter system can be used for analysis and sorting of mutant libraries using flow cytometry based on translation levels, and that it may enable the selection of proteins with altered stability.

Screening and sorting of protein wide mutant libraries

Factors affecting proper folding of proteins can be investigated by random mutagenesis. However, stop-codons, frameshifts and indels will often be introduced in a randomly generated mutant library, which render it difficult and time-consuming to screen for new protein variants. We thus demonstrate the use of the dual-reporter system to screen for variants with either reduced or increased stability in a high-throughput mode. First, a randomly generated mutant library of pET22-PARP1-BRCT-mCherry was expressed with the protein folding sensor pSEVA631(Sp)-IbpAp-GFP-ASV to

demonstrate the applicability of the dual-reporter system to screen for new variants with correct translation but altered stability. PARP1-BRCT is a stable protein resulting in a low GFP signal and a signal distribution corresponding to the control. We created a mutant library and used the folding sensor to analyse positions and variation important for proper cellular folding. The mutant library was prepared using the error prone DNA polymerase Mutazyme II that provides a minimal mutational bias. By adjusting the amount of initial target DNA and the number of gene duplications, a mutation rate of 1-3 amino acid substitutions per protein was achieved. GFP and mCherry fluorescence was quantified by FACS one hour after protein expression was induced by IPTG (Figure 5A). PARP1-BRCT WT and the PARP1-BRCT mutant library showed a high translation level with well-defined mCherry signals that were distinct from the background fluorescence from an empty pET22b vector (Figure 5A). Before sorting of the cells, similar GFP signals and distributions were obtained for PARP1-BRCT WT, the PARP1-BRCT mutant library, and the background control. The cell cultures were sorted using FACS for high mCherry signal (P1) alone and for both high mCherry signal (P1) and high GFP signal (P2). Gate 1 (P1) was defined as a mCherry signal higher than the control plasmid background, to ensure that only cells expressing correctly translated proteins were collected. Gate 2 (P2) was defined as the upper 1% of cells with the highest GFP fluorescence, in order to select for variants expressing proteins with decreased stability. The collected cells were grown and sorted again using the same criteria as in the initial sorting. The fluorescence from the final pools of sorted cells was analysed by FACS one hour and three hours after induction of protein expression by IPTG (Figure 5B). The PARP1-BRCT library sorted for high mCherry signal (Lib. P1) represents correctly translated proteins, and show similar GFP intensities and signal distributions as the PARP1-BRCT WT. The PARP1-BRCT library

sorted for both high mCherry signal and a high GFP signal (Lib. P2) shows a clear shift in GFP signal compared to PARP1-BRCT WT and the PARP1-BRCT library that was only sorted for correct translation (Lib. P1) (Figure 5B). The shift in GFP signals indicates that protein variants with impaired folding properties have been enriched. Furthermore, the PARP1-BRCT library sorted for high GFP fluorescence showed a small broadening of the GFP signal three hours after induction due to continuous expression of GFP concurrently with PARP1-BRCT being expressed. Single cells were extracted from the sorted libraries and the PARP1-BRCT gene was amplified by PCR and prepared for DNA amplicon sequencing.

Mutant library sequencing, stability analysis and decoy detection

The PARP1-BRCT mutant library was sequenced by NGS both before and after sorting into the two populations (red (Lib. P1) and green (Lib. P2)) using FACS (Figure 5C). For simplicity, we only investigated single site (amino acid) mutants. For a given mutant protein sequence, we compared its frequency in the green pool (destabilized proteins) with the frequency in the reference pool and used it as a proxy for protein stability. More specifically, for a given mutant protein sequence, we calculated the ratio between the high GFP fluorescence pool and the reference pool using Enrich2 [36] that gives a score based on the normalized ratios. If the score is higher than 0 we consider the mutation to be neutral or stabilizing, and if the score was below 0, we consider it to be destabilizing (Figure 5C bottom). As expected, full saturation mutagenesis was not obtained due to the low mutagenesis rate that makes it unlikely to have more than one nucleotide change per codon.

We then turned to *in silico* calculations of the change in thermodynamic stability ($\Delta\Delta G$) of the BRCT domain using FoldX[37] and a solution NMR structure (PDB ID: 2COK) to assess how well

271 predictions of thermodynamic stability correlate with the experimental data. We performed
272 computational saturation mutagenesis in which we mutated each amino acid to all 19 other possible
273 ones and calculated the change in stability (Figure 5C bottom), and considered $\Delta\Delta G$ values greater
274 than or equal to 3 kcal/mol to be destabilizing (red x in Figure 5C) and the remaining to be either
275 neutral or stabilizing, to match the binary format of the sequencing data. Overall, we find a relatively
276 low agreement between the FoldX calculations and the sequencing data, where destabilizing
277 mutations based on the sequencing data (blue squares) are not always captured by FoldX. Enrich2
278 only ranks the observed mutations and do not classify the mutations as either stabilizing or
279 destabilizing, thus changing the Enrich2 cut-off may either increase false positives or false negatives.
280 To reduce possible noise, we analysed the data position-wise by calculating the ratio between the
281 number of destabilizing mutations and the number of total mutations for each position in the
282 sequence (N_{destab}/N_{total}) for both FoldX and the experimental sequencing data (Figure 5C top). The
283 ratio is high when most mutations lead to destabilization and small when most mutations are
284 neutral/stabilizing. From this analysis, we find a better correlation between the FoldX calculation and
285 our experimental data (Figure 5C top). Note that for the FoldX calculations we have performed full
286 saturation mutagenesis, which means that N_{total} is 19 for all positions, in contrast to the experimental
287 data where N_{total} varies. To remove the bias of selecting the $\Delta\Delta G$ cut-off for FoldX as well as
288 summarizing across whole amino acid positions, we also performed a Receiver Operating
289 Characteristic analysis (Figure 5D). Here, the sequencing data provides the mutation specific labels
290 (blue vs green in Figure 5C) and the $\Delta\Delta G$s predicted from FoldX represent the predicted scores. From

this analysis, we obtained an area under the curve of 0.61 suggestive of a reasonable but non-perfect correlation between the calculated stabilities and the experimental sequencing data.

Decoy detection in a protein structural ensemble is useful for protein structure prediction when using structural prediction tools such as Rosetta[38], which often produces a pool of candidate protein structures that might need additional filtering. Inspired by previous work that showed a correlation between the mutational tolerance of a site and how buried that site is in the protein structure[39], we examined whether the results from the folding sensor could be used in decoy discrimination. We used the mean of the Enrich2 scores for each position to individually score a pool of 20,000 structures generated by Rosetta. The assumption was that for a given residue in a native protein structural model, the residue depth should correlate with the mean Enrich2 score calculated from our experiments[39]. As an example, we depict a protein structure where each residue is coloured either red or blue depending on their individual mean Enrich2 scores (Figure 5E). Here, we find that low Enrich2 scores are likely attributed to residues in the core of the protein. Intuitively, one can imagine that the deeper a residue is embedded in the native protein structure, the more likely it is to destabilize the protein upon mutation due to packing issues. For each BRCT model we thus calculated the Spearman's correlation coefficient, ρ, to quantify the correlation between the residue depth and mutational tolerance (mean Enrich2 score). This correlation coefficient is considered as a structure specific score for which a higher coefficient is suggestive of a more native-like protein. To examine its usefulness in separating high quality structures from low quality structure, we plotted ρ as a function of the structural Global Distance Test – Total Score (GDT-TS) of the 20,000 generated structures with respect to the first conformer in the PDB structure (Figure 5F). GDT-TS range from zero to one where

one corresponds to a native or near native structure and zero is likely an extended protein. In Figure 5F, we find a clear correlation between the structural scores ρ, and the structural quality defined by GDT-TS, suggesting that our experimental data can indeed be used to identify likely structures in a pool of candidate structures.

Experimental validation of mutant variants identified through deep mutational scanning analysis

From the deep mutational scanning we chose 20 variants for further analysis, of which 14 mutations (G20V, G20W, A31T, I33N, G37E, G37R, G37W, G38R, C50Y, S52I, S52N, I72N, V74F, H97L) were suggested by the deep mutational scan to result in misfolding and 6 mutations (I33T, V74I, D78V, Q81R, A96P, A96V) that were found not to interfere with protein folding. The 20 variants were synthesized and introduced into the translation sensor plasmid and analysed using flow cytometry, and the protein concentration was quantified using Western blots (Figure 6). The translation levels detected by the mCherry signal and by Western blotting was comparable for 6 (A31T, I33T, D78V, Q81R, A96P, A96V) of the 20 variants (Figure 6A). Translation levels of the remaining 14 variants, of which 13 were predicted to misfold, were detected by a mCherry signal, but with either no translation level detected by Western blot, or with Western blot signals distinctly lower than what would be expected based on the mCherry signal. This suggests that these variants were translated correctly but possibly degraded by proteases in the cell before detection by Western blot. Moreover, lower GFP signals of variants with low or no translation level detected by Western blot, suggests that the proteins were degraded before chaperones were able to bind and protect the unfolded or partially unfolded protein (Figure 6B). The PARP1-BRCT I33N and G37E variants were predicted to misfold, but in contrast to the other variants, they were detectable by Western blotting, although still not at the

same level as the mCherry signal, suggesting that there was a significant difference in the degradation rate of the mutants. The corresponding high GFP signal shows that the stability of the PARP1-BRCT I33N and G37E variants was decreased as also predicted from the library sequencing data and the FoldX analysis. For the variants predicted not to interfere with protein stability, there was a correlation between the level of protein measured by mCherry fluorescence and the amount of protein quantified by Western blotting. This observation was corroborated by the low GFP signals demonstrating that the mutations do not decrease the stability. The percentage of soluble protein was further quantified, and visualized by Western blotting of the total protein fraction and the soluble protein fraction (Figure 6C). As expected, variants with high GFP signals have low fractions of soluble protein, whereas variants with low or no GFP signal have high fractions of soluble protein.

Identification of mutants with increased stability

Having demonstrated the potential of the dual reporter system for identifying residues important for protein folding, we expected that the reporter system can also be used to identify mutations that stabilize the folding of the protein. The PARP1-BRCT I33N variant was identified as a misfolded protein from the PARP1-BRCT library when screening for variants with decreased stability, and we asked which variants, if any, might suppress the effect of I33N. PARP1-BRCT I33N was thus used as a background for a new randomly generated mutant library with a mutation frequency of 1-3 mutations per protein, which was then co-expressed with the protein folding sensor pSEVA631(Sp)-IbpAp-GFP-ASV. Single cells that had high translation levels (Gate 1) as well as increased protein stability (Gate 2) were sorted by FACS (Figure 7A, left panel). As expected PARP1-BRCT-I33N and the PARP1-BRCT-I33N mutant library resulted in significantly higher GFP signals than PARP1-BRCT-WT.

After the first round of sorting, 64 single clones with high mCherry and low GFP signal (Gate 1 + Gate 2) were reanalysed by flow cytometry, and one of the clones (1.5%) had a GFP signal overlapping with the GFP signal for PARP1-BRCT-WT) (Sort 1, Figure 7B upper panel). After the second round of sorting, this population was further enriched to account for 12.5 % of cells (Sort 2, Figure 7B, lower panel)). A total of 64 clones were randomly selected from pool A and pool B and characterized by Sanger sequencing. Although the input library contains a wide range of mutations, all the selected clones were found to encode wild type PARP1-BRCT, except for one silent mutation, P10P, found after the second round of sorting. This mutation was caused by a codon change from CCA to CCT, neither of which are characterized as rare codons[40]. These results demonstrate that it is possible to select more stable and correctly folded variants by successive rounds of FACS. Our observations that only WT sequences were found after sorting for proteins with increased stability from a destabilized library is a likely result of the I33N mutation being a single nucleotide substitution making it likely to revert back to PARP1-BRCT WT. In addition, the most severely destabilizing single amino acid changes require multiple amino acid substitutions in order to recover or improve protein stability[12]. To increase the likelihood of finding more clones with increased stability, a library with a higher mutation frequency could be used. Still, our results show that it is possible to enrich the population of clones with low GFP clones and thus select for improved stability.

Discussion

The dual-reporter system presented here is a high-throughput screening method with fast and simultaneous monitoring of translation and protein folding at the single cell level. The setup has broad applicability and can be used as a screening tool to optimize expression conditions testing

different solubility and purification tags, as well as a tool in deep mutational scanning and directed evolution studies. The use of the small hexa-histidine tag for translational coupling to the reporter protein avoids a tag that would be likely to interfere with the three-dimensional structure of the protein and its function. Furthermore, the tag has the advantage that it can be used for down-stream quantification and purification steps.

Formation of inclusion bodies is often the bottleneck when expressing recombinant proteins in *E. coli*, and solubility tags or solubilisation of the inclusion bodies and subsequent refolding of the proteins are often necessary to recover folded and active proteins. In our reporter system, the GFP fluorescence is a result of a heat shock response initiated by the formation of aggregates and misfolded protein within the cell and is dependent on the presence of DnaK or DnaJ binding sites in the misfolded protein. The heat shock factor RpoH needed for binding to the lbpAp heat shock promoter on the protein folding sensor is released when DnaK binds the misfolded protein. The number of DnaK or DnaJ binding sites may influence the intensity of the GFP signal, as higher amounts of RpoH will be released with higher numbers of DnaK binding sites.

BRCA1-BRCT and E6 both expressed as insoluble proteins, but with different levels of GFP fluorescence. Using the Limbo DnaK binding site prediction tool[41], BRCA1-BRCT and E6 are predicted to contain four and two DnaK binding sites, respectively. Assuming that all DnaK binding sites are exposed in the unfolded protein, this may explain the difference in GFP intensity between the two proteins. This suggests that the GFP fluorescence may not be comparable when investigating unrelated proteins. However, for variants of the same protein the GFP output can be used as a direct measure of protein stability as we have demonstrated for CI2.

Mutant libraries are a main component of deep mutational scanning and directed evolution studies. The major drawback of randomly generated mutant libraries is the introduction of frame-shift mutations, stop-codons and indels that alters the amino acid sequence and results in non-functional proteins. The incorporation of the translation sensor aims to differentiate between nonsense and missense mutations due to the necessity of complete termination of translation for a reporter signal to emerge.

Generation of mutant libraries using an error-prone DNA polymerase is limited by the genetic code, thus full saturation mutant libraries are difficult to obtain. Depending on the purpose of the experiment, the mutant libraries should be designed accordingly. Since multiple point mutations are often necessary for obtaining protein variants with improved overall stability, a mutant library for selecting stabilized variants should have a higher mutation frequency than a library for selecting destabilised variants. On the other hand, global analysis of libraries with different number of amino acid changes may provide detailed insight into protein folding and function[42–44].

We have shown that the mutational profile can be used to provide insight into the structure of a protein through decoy detection. Very recently it has been shown that more extensive deep mutational scan can be used to determine accurate three-dimensional structures [45,46] and we envision that when the reporter system is used to select for stable protein variants it can be used in such structure-determination protocols.

Through screening for protein variants with improved stability from the destabilized PARP1-BRCT-I33N mutant, we successfully identified revertants to the wild type sequence, while a single silent mutation, PARP1-BRCT P10, was also identified. Where the amino acid sequence determines

the three-dimensional fold of a protein the nucleotide sequence may affect the translational rate and thus co-translational folding of the proteins. Silent mutations may therefore still improve protein solubility and stability.

All destabilized variants found in the PARP1-BRCT library are situated in the core of the protein fold, which is consistent with the core being more sensitive to mutation, which often also results in loss of function[47]. When mutating enzymes for improved translation levels and protein folding it involves a risk of altering the activity of the enzyme. It is known that mutations within or close to the catalytic sites of enzymes may result in an improved stability of the protein, but with a corresponding decrease in enzyme activity[48,49]. As the dual-reporter system can be used to obtain protein variants with high translation levels and high or moderate solubility, a downstream activity assay is needed to ensure an active enzyme. Activity assays, however, are generally protein specific and are difficult to incorporate in a generalized screening method.

Conclusion

The presented dual-reporter biosensor system assesses *in vivo* protein translation and solubility with a reliable output. The use of a translation coupling cassette in the protein translation sensor makes it possible to avoid fusion of the reporter protein to the test protein, which reduces the risk of altering the protein folding properties. The dual-reporter system has been demonstrated to be generally applicable, since the protein folding sensor is based on the cellular heat shock response system, where measurable protein functions are not a prerequisite. The dual-reporter system can be used as a screening assay in directed evolution and deep mutational scanning studies to identify protein variants with high expression levels and improved protein stability in a high-throughput setup. By

438 applying next generation sequencing on mutant libraries, we demonstrate a good correlation
439 between experiments and computational protein stability predictions. The dual-reporter system was
440 capable of identifying mutations that were not correctly predicted by computational tools, and we
441 therefore envision that the experimental data that can be generated using the system may be
442 valuable for further improving computational stability predictions.
443
444 **Materials and Methods**
445
446 *Chemicals and enzymes*
447 Standard chemicals were purchase from Sigma Aldrich and sodium acetate was purchased from
448 Scharlau, imidazole was purchased from PanReac AppliChem and IPTG was purchased from Fischer
449 Bioreagents. Enzymes for standard cloning procedures were purchased from Thermo Fisher Scientific
450 and New England Biolabs, respectively.
451
452 *Construction of a fluorescence-based protein folding reporter*
453 For construction of a protein folding sensor that reports on the formation of inclusion bodies (IB), the
454 IbpA promoter (Genbank: LQ302077.1) from *E. coli* MG1655 was fused to either a stable (GFP-mut3;
455 GenBank: LQ302079.1[10]) or a destabilized version of GFP (GFP-ASV; GenBank: LQ302078.1 [10]). The
456 GFP-ASV and GFP-mut3 were amplified by PCR using primer pairs and templates as indicated in Table
457 S1 and Table S2. PCR products were cloned into pSEVA441 (GenBank: JX560339.1) using the *Xba*I and
458 *Spe*I restriction sites, resulting in either pSEVA441-GFP-ASV or pSEVA441-GFP-mut3. The *E. coli* lbpA

promoter was amplified by PCR (Table S1) and cloned via the *Pac*I and *Xba*I restriction sites into pSEVA441-lbpAp-GFP-ASV and pSEVA441-lbpAp-GFP-mut3, respectively. To generate pSEVA631(Sp)-lbpAp-GFP-ASV or pSEVA631(Sp)-lbpAp-GFP-mut3, the lbpAp-GFP reporter gene was subcloned via *Pac*I and *Spe*I into the pSEVA631 (GenBank: JX560348.1). Finally, the gentamycin cassette of pSEVA631 was replaced by the spectinomycin cassette of pSEVA441 using the *Spe*I and *PshA*I restriction sites. All constructs were verified by Sanger sequencing.

Fusion of proteins with a fluorescent translation-sensor

A set of proteins were fused to the translation coupling cassette[24] (GenBank: LQ302080.1) followed by mCherry (GenBank: LQ302081.1). The BRCT-domain of human Poly [ADP-ribose] polymerase 1 (PARP1-BRCT, GenBank: LQ302082.1), a truncated version of BRCT-domain of human breast cancer 1, early onset (BRCA1-BRCT, GenBank: LQ302085.1 the human cyclin-dependent kinase 4 inhibitor D (p19, GenBank: LQ302086.1), and protein E6 from human *papillomavirus type 16* (GenBank: LQ302087.1) were amplified by PCR using the primers and templates as indicated in Table S1. Additionally, mCherry was amplified by PCR according to Table S1. Each protein encoding DNA fragment was assembled with the mCherry-PCR fragment and *Nde*I and *Hin*dIII digested pET22b vector (Novagen), using a Gibson assembly reaction (New England Biolabs). The resulting expression vectors pET22b-XXX-trans-mCherry (XXX stands for the respective protein; see also Table S2) comprise the coding sequence of the different proteins being linked via a C-terminal translation coupling cassette[24] to the open reading frame (ORF) of mCherry. All cloned constructs were confirmed by Sanger sequencing.

Cloning of NusA and SUMO fusion proteins

For analysing the impact of NusA and SUMO protein-tags on expression and translation levels of either PARP1-BRCT, BRCA1-BRCT, p19, or E6, proteins were N-terminally fused to NusA (GenBank: LQ302088.1) and SUMO (GenBank: LQ302089.1), respectively[50,51]. Thereby, NusA and SUMO were amplified by PCR using the primers indicated in Table S1 and inserted into pET22-XXX-trans-mCherry via the *Nde*I restriction site. The final protein expression reporter plasmids named pET22b-NusA-XXX-trans-mCherry and pET22b-SUMO-XXX-trans-mCherry (XXX stands for the respective protein; see also Table S2), respectively, were all verified by sequencing.

Impact of plasmid copy number and GFP stability on protein folding reporter assay sensitivity

The impact of the vector copy number and intracellular turnover rate of GFP, respectively, on the protein folding reporter system was analysed to optimize the readout sensitivity of the assay. Therefore, pSEVA631(Sp)-lbpAp-GFP-ASV and pSEVA631(Sp)-lbpAp-GFP-mut3 (pBBR1 origin), as well as pSEVA441-lbpAp-GFP-ASV and pSEVA441-lbpAp-GFP-ASV (ColE1 origin) (constructed as described above), were co-transformed with pET22b in *E. coli* Rosetta2™(DE3)pLysS (Novagen®). Transformants were selected on LB plates containing 25 µg/mL chloramphenicol, 50 µg/mL spectinomycin, and 100 µg/mL ampicillin. Single clones were inoculated in LB medium supplemented with the corresponding antibiotics and grown at 37°C and 300 rpm to an OD_{600} of 0.5. IB formation in *E. coli* was induced by performing a heat-shock for 10 min at 42°C. After heat shock, cells were grown for an additional 2.5 hours at 37°C and 300 rpm. Induction of the lbpAp promoter by IBs in single cells was monitored over

time by changes of the GFP signal using flow cytometry (Instrument: BD FACS-Aria™SORP cell sorter; Laser 1: 488 nm: >50 mW, Filter: 505LP, 530/30-nm FITC, Laser 2: 561 nm: >50 mW; Filter: 600LP, 610/20-nm PE-Texas Red®). As control, the GFP signal in un-induced cells was monitored for each time point. The GFP (FITC-A, X-mean) values at each time point analysed using the FlowJo V10 software were normalized to the corresponding background GFP signal.

To further investigate the impact of GFP stability on the sensitivity of the lbpAp-GFP reporter gene assay, pSEVA631(Sp)-lbpAp-GFP-ASV and pSEVA631(Sp)-lbpAp-GFP-mut3, respectively, were co-transformed with either pET22b, pET22-PARP1-BRCT-trans-mCherry or pET22-BRCA1-BRCT-trans-mCherry into *E. coli* Rosetta2^{TM}(DE3)pLysS (Novagen®). Transformants were selected on LB plates containing 25 µg/mL chloramphenicol, 50 µg/mL spectinomycin and 100 µg/mL ampicillin. Single clones were grown at 37°C and 300 rpm in LB medium supplemented with the corresponding antibiotics. At OD_{600} of 0.5-0.7 the expression of the human proteins was induced by addition of 0.5 mM IPTG. Directly after induction, the growth temperature was changed to 30°C. Induction of the lbpAp-GFP variants by misfolded proteins was analysed 1 hour after induction using flow cytometry as mentioned above. For data analysis the GFP-signal (FITC-A, X-mean) was normalized to the respective GFP-signal of the vector control.

Determination of protein localization by fractionated cell disruption

Intracellular localization of proteins was further analysed by fractionated cell disruption. Here, cells (from 1 mL culture) were harvested either 1 hour (for immunoblot analysis) or 3 hours (for InstantBlue staining) after induction of protein expression. The cell pellet was resuspended in 50 µL

resuspension buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl; 10 mM EDTA, 1 x HP-protease inhibitor mix (Serva)) and cells were broken by repeated cycles of freeze and thaw. Afterwards, cells were adjusted to a final OD_{600} of 5 in resuspension buffer supplemented with benzonase (≥ 500 units; Sigma Aldrich). After 20 min incubation on ice, cells were spun-down for 1 min at 500 x g to remove cell debris. The supernatant containing all soluble and insoluble proteins was transferred to a fresh reaction tube. An aliquot of the supernatant was taken, representing the total protein fraction (total). The remaining cell lysate was spun-down for 15 min at 20,000 x g and the supernatant containing all soluble proteins was transferred into a new reaction tube (sol). The isolated fractions were separated on SDS-PAGE (RunBlue 4-20 %, Expedeon; NuPAGE®Bis-Tris gel 4-12%, Invitrogen) and analysed by InstantBlue staining (Expedeon) and quantitative immunoblotting using an anti-His antibody (Novagen).

Dual reporter system for simultaneous monitoring of protein translation and folding in single E. coli *cells*

To analyse the combined reporter system, pSEVA631(Sp)-lbpAp-GFP-ASV and the protein expression reporter plasmids (pET22b-XXX-trans-mCherry, pET22b-NusA-XXX-trans-mCherry, pET22b-SUMO-XXX-trans-mCherry) were co-transformed into chemically competent *E. coli* Rosetta2™(DE3)pLysS (Novagen®). Transformants were selected on LB plates containing 25 µg/mL chloramphenicol, 50 µg/mL spectinomycin, and 100 µg/mL ampicillin. Single clones were grown in LB medium (supplemented with the corresponding antibiotics) at 37°C and 300 rpm to an OD_{600} of 0.5-0.7 and expression of proteins was induced by addition of 0.5 mM IPTG. Directly after induction, the growth

temperature was changed to 30°C. Protein expression and folding was analysed 1 hour after induction using flow cytometry as mentioned above. For data analysis, GFP (FITC-A, X—mean) signal was normalized to the corresponding PARP1-BRCT signal.

To confirm signal of the translation reporter, protein expression levels were further analysed by instant blue staining and quantitative immunoblotting using an anti-His-Antibody. Cell-disruption was performed by freeze and thaw cycles as described before and the total protein fractions as well as intracellular localization of the proteins were analysed. Western Blot signal was quantified using the Image J software[52].

Identification of PARP1-BRCT mutants with altered folding properties using FACS

To generate a PARP1-BRCT mutant library the PARP1-BRCT domain was randomly mutated, aiming at a mutation rate of 1 to 3 mutations per construct, using the GeneMorph II random mutagenesis kit (Agilent) according to manufacturer´s instructions. Primers and templates used for the reactions are indicated in Table S1. A megawhop reaction was performed with the random mutated PCR product as megaprimer and pET22-PARP1-BRCT-trans-mCherry as template. The resulting linear DNA fragments were transformed into MegaX DH10B™ T1R Electrocomp™ cells (Invitrogen) and transformants were selected on LB plates supplemented with 100 µg/mL ampicillin. The colonies (library size >100,000) were pooled and the plasmids were directly purified without further growth.

The vectors pET22b, pET22-PARP1-BRCT-trans-mCherry, and the created pET22-PARP1-BRCT-trans-mCherry mutant library were transformed into electro-competent Rosetta2(DE3)pLysS cells harbouring the protein folding sensor (pSEVA631(Sp)-IbpAp-GFP-ASV). After recovery, transformants

were directly inoculated into 2 mL LB medium containing 20 µg/mL chloramphenicol, 50µg/mL spectinomycin, 100 µg/mL ampicillin, and grown overnight at 37°C and 300 rpm. Cells were transferred into fresh medium and grown at 37°C and 300 rpm to an OD_{600} of 0.5 – 0.7. Expression of proteins was induced by addition of 0.5 mM IPTG and the growth temperature of the culture was shifted to 30°C. 1 hour after induction, cells were analysed by flow cytometry as mentioned above. 150.000 cells expressing a PARP1-BRCT mutant protein at wildtype level based on the translation sensor signal (Figure 5A, gate 1), and which had an increased GFP signal (Figure 5A, Gate 2) were sorted into 1 mL LB medium supplemented with antibiotics and grown overnight at 37°C and 300 rpm. To further enrich the *E. coli* fraction harboring proteins with altered folding properties, another round of protein expression and sorting (150.000 events) was carried out as described above.

The following day, the sorted cell population was again analysed 1 hour after induction of protein expression by flow cytometry. Subcellular localization of proteins in the sorted *E. coli* fraction was analysed by Immunoblotting using an anti-His antibody as described above.

For next generation sequencing, plasmids were isolated from the sorted *E. coli* population. As control, plasmids were isolated from the PARP1-BRCT mutant library, which was used as starting material for sorting. Two 300 bp DNA fragments were amplified from the PARP1-BRCT library using a high fidelity polymerase (primers as indicated in Table S1). The amplified fragments were purified using AMPure XP beads (Beckman Coulter) to remove free primers and primer-dimer species. Both PCR-products were mixed in a one-to-one ratio.

Next, a PCR reaction was performed to attach Illumina sequencing adapters (Nextera XT Index Kit, Illumina) to the DNA fragments. For the reaction a KAPA HiFi HotStart Polymerase (Kapa Biosystems)

was used. The resulting PCR products were purified with AMPure XP beads. The product size of the PCR reaction was verified on a Bioanalyzer DNA 1000 chip and the DNA was quantified using a Qubit® 2.0 Fluorometer. DNA fragments were normalized to 10nM in 10mM Tris pH8.5, 0.1% Tween 20. In order to reduce the background signal, the sample was spiked with 5% Phi-X control DNA (Illumina). The DNA was loaded onto the flow cell provided in the MiSeq Reagent kit v2, subjected to 300 cycles (Illumina), and sequenced on a MiSeq sequencing system (Illumina).

Enrichment analysis

The analysis was carried out using Enrich2 software[36]. However, due to issues running Enrich2 directly from raw fastq files, we converted the fastq files into Enrich2 compatible variant counts using python scripts. The scripts for doing this as well as an Enrich2 analysis config file are available at https://doi.org/10.11583/DTU.10265420. The script does the following: Reads were merged using FLASH v.1.2.11[53] and mapped to the reference sequence using bowtie2 v.2.3.4.1[54]. The SAM files that bowtie2 outputs are then parsed to create Enrich2 compatible variant count files.

Folding properties of PARP1-BRCT single mutants

To generate PARP1-BRCT single mutants, a 2-fragment Gibson assembly reaction was performed. For each single mutant two overlapping DNA fragments were amplified by PCR using pET22b-PARP1-BRCT-trans-mCherry as template. Primer pairs are listed in Table S3. Finally, the two DNA fragments were joined using Gibson Assembly® Cloning Kit (New England Biolabs) according to manufacturer's

instructions. The sequence of each single mutant was confirmed by sequencing. Resulting mutant constructs are listed in Table S3.

To examine the translation levels and protein stability of PARP1-BRCT single mutants, each mutant construct (Table S3) was co-transformed with pSEVA631(Sp)-IbpAp-GFP-ASV into chemically competent *E. coli* Rosetta2™(DE3)pLysS (Novagen®). Protein expression was induced by addition of IPTG and protein translation and folding were analysed by flow cytometry and quantitative immunoblotting as described before. To determine the percentage of soluble protein, the western blot signal was quantified using the Image J software.

Isolation of PARP1-BRCT-I33N single mutants with rescued folding properties using the dual reporter system

A PARP1-BRCT-I33N library was generated as described before, using pET22b-PARP1-BRCT-I33N-trans-mCherry as template. The plasmids pET22b, pET22-PARP1-BRCT-I33N-trans-mCherry and the created pET22-PARP1-BRCT-I33N-trans-mCherry mutant library were transformed into electro-competent Rosetta2(DE3)pLysS cells harbouring the protein folding sensor (pSEVA631(Sp)-IbpAp-GFP-ASV). Protein expression was induced with IPTG and flow cytometry was performed as described above. 64 single clones that show protein expression (Figure 7A, Pool 1, Gate 1) in combination with a decreased GFP signal (Figure 7A, Pool 1; Gate 2) were sorted in 200 µl LB medium supplemented with antibiotics and grown to stationary phase at 37°C and 300 rpm. To further enrich the *E. coli* fraction harbouring proteins with rescued folding properties, a pool of 150,000 cells was sorted (identical gating as single clones) into 1 mL LB medium supplemented with antibiotics and grown again

626 overnight at 37°C and 300 rpm. Subsequently, a second round of IPTG induction and sorting was
627 performed to gain another 64 single clones (Figure 7A, Pool 2; Gate 1 and 2). To verify GFP signal, all
628 single clones (Pool 1 and Pool 2) were inoculated into fresh medium, protein expression was induced,
629 and GFP expression was analysed using a BD LSRFortessa™ cell analyser in the HTS mode (Laser 1: 488
630 nm: >50 mW, Filter: 505LP, 530/30-nm FITC). Finally, plasmids were isolated from single clone
631 cultures, which showed no GFP signal after induction, and analysed by Sanger sequencing.
632
633 *FACS-based CI2 stability assay*
634 To generate five CI2 mutants with varying stabilities, Site-Directed II Lightning mutagenesis kit (Agilent
635 Technologies) was used with CI2 WT as template. Each mutant was amplified by PCR using primer
636 pairs as indicted in Table S1. PCR products were cloned into pET22b-mCherry vector using the *NdeI*
637 and *SpeI* restriction sites and joined using Gibson Assembly® Cloning Kit (New England Biolabs)
638 according to manufacturer's instructions.
639 The CI2 variants were co-transformed with pSEVA631(Sp)-IbpAp-GFP-ASV into Rosetta2 (DE3) pLysS
640 chemically competent cells and expressed in 50 ml LB media supplemented with 100 µg/µl ampicillin,
641 25 µg/ml chloramphenicol, and 50 µg/ml spectinomycin at 30°C, 250 rpm to an OD_{600} of 0.8. Protein
642 expression was induced with 0.5 mM IPTG. Cells were extracted before and 1 hour after induction and
643 kept on ice until FACS analysis. The mCherry and GFP fluorescence was analysed on a BD FACS-
644 ARIA^{TM}SORP cell sorter as mentioned above.
645

CI2 expression and purification for stability measurements

CI2 variants transformed into Rosetta2 (DE3) pLysS competent cells and expressed in 1 L LB in the presence of 100 µg/µL ampicillin and 25 µg/ml chloramphenicol at 37°C. Protein expression was induced at OD_{600}~0.5-0.7 with 0.5 mM IPTG and cells were further grown at 30°C for 4-5 hours. Cells were harvested by centrifugation at 5,000 x g for 20 min. Cell pellets were resuspended in 20 mL buffer A (20 mM sodium acetate pH 5.3) and frozen at -20°C. Cell lysis was performed by 2 rounds of sonication (1 min, 80% amplitude, 0.5 cycles, (Hielscher UP200S)) followed by 30 min incubation on ice in presence of 1 mg DNase. Cell debris and protein aggregates were removed by centrifugation at 20,000 x g, 4°C for 30 min. The supernatants were loaded onto a 1 mL HisTrap HP column (GE Healthcare) equilibrated with buffer A, and eluted with a gradient of buffer B (20 mM sodium acetate pH 5.3, 1 M imidazole) from 0-100 %. Fractions containing CI2 determined from SDS-PAGE analysis were concentrated and loaded onto a superdex75 10/300 GL column (GE Healthcare) equilibrated with 20 mM sodium phosphate pH 7.4, 150 mM NaCl. For buffer exchange the samples were concentrated and loaded onto a superdex75 10/300 GL column equilibrated with 50 mM MES pH 6.25. The purity of the proteins was assessed by SDS-PAGE and the protein concentration was determined using a spectrometer (PerkinElmer lambda40) with an extinction coefficient of 6990 $M^{-1}cm^{-1}$.

The CI2 variants were diluted to 10 µM in MES pH 6.25 with or without 6 M guanidium chloride. Using both solutions a dilution series of guanidium chloride ranging from 0 – 6 M guanidium chloride was prepared. Intrinsic tryptophan and tyrosine fluorescence of the CI2 variants was measured in triplicates using nanoDSF technologies on a Prometheus NT.48 instrument (nanoTemper

667 technologies) with a temperature range from 15-95°C with 1°C/min increments. Global fitting of the
668 temperature and denaturant unfolding was performed using the 330 and 360 nm fluorescence and
669 ΔG_U and was obtained.
670
671 *Generating and scoring decoy structures.*
672 We generated 20,000 decoy structures using Rosetta's threading protocol[38] with PDBID: 2COK
673 (Solution structure of BRCT domain of poly(ADP-ribose) polymerase-1) as a template. As a means to
674 score a given decoy structure, we calculated the spearman's correlation coefficient ρ between the
675 residue depths of the decoy structure and the mean Enrich2 positional score for the corresponding
676 positions in the sequence.
677
678 **Acknowledgements**
679 This work was supported by the Novo Nordisk Foundation through a grant to DTU Biosustain
680 (NNF10CC1016517) as well as through grants for Protein OPtimization (POP) (NNF15OC0016360), a
681 Hallas-Møller stipend (R173-A14446) and a project grant from the Lundbeck Foundation (R126-2012-
682 12589). The pSEVA plasmids were a kind gift of Professor de Lorenzo, Centro Nacional de
683 Biotecnologia-CSIC, Spain.
684 **Author contributions**
685 AZ, LH, MK, KT, KLL, ATN designed the experimental work, AZ, LH, AK, MK performed the experiments,
686 AZ, EP, LH, LEP, MK, KT, KLL, ATN analysed the data, LH, AZ wrote the manuscript, LH, AZ, MK, KT, KLL,
687 ATN edited and reviewed the manuscript, and all authors read and approved the final manuscript.

688 **Additional information**

689 Supplementary Information accompanies this paper can be found at…

690

691 **Competing interests**: AZ and ATN have filed a provisional application on this work.

692

693 **References**

694 1. Costa, S., Almeida, A., Castro, A. & Domingues, L. Fusion tags for protein solubility, purification,
695 and immunogenicity in Escherichia coli: the novel Fh8 system. *Front. Microbiol.* **5**, 1–20 (2014).
696 2. Marblestone, J. G. *et al.* Comparison of SUMO fusion technology with traditional gene fusion
697 systems: Enhanced expression and solubility with SUMO. *Protein Sci.* **15**, 182–189 (2006).
698 3. Carson, M., Johnson, D. H., McDonald, H., Brouillette, C. & DeLucas, L. J. His-tag impact on
699 structure. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **63**, 295–301 (2007).
700 4. Angov, E. Codon usage : Nature ' s roadmap to expression and folding of proteins. *Biotechnol. J.*
701 **6**, 650–659 (2011).
702 5. Komar, A. A., Lesnik, T. & Reiss, C. Synonimous codon substitution affects ribosome traffic and
703 protein folding during in vitro translation. *FEBS Lett.* **462**, 387–391 (1999).
704 6. Yu, C. H. *et al.* Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-
705 translational Protein Folding. *Mol. Cell* **59**, 744–754 (2015).
706 7. Fedorov, a. N. Cotranslational Protein Folding - Minireview. *J. Biol. Chem.* **272**, 32715–32718
707 (1997).
708 8. Sørensen, H. P. & Mortensen, K. K. Advanced genetic strategies for recombinant protein

expression in Escherichia coli. *J. Biotechnol.* **115**, 113–128 (2005).

9. Humbard, M. A., Surkov, S., De Donatis, G. M., Jenkins, L. M. & Maurizi, M. R. The N-degradome of Escherichia coli: Limited proteolysis in vivo generates a large pool of proteins bearing N-degrons. *J. Biol. Chem.* **288**, 28913–28924 (2013).

10. Anders, J. B. *et al.* New Unstable Variants of Green Fluorescent Protein for Studies of Transient Gene Expression in Bacteria. *Appl. Environ. Microbiol.* **64**, 2240–2246 (1998).

11. Soto, C. *et al.* Solubility As Funtion of Proteins Structure and Solvent Components. *Nat. Med.* **4**, 822–826 (1998).

12. Goldenzweig, A. *et al.* Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **63**, 337–346 (2016).

13. Shih, Y. *et al.* High-throughput screening of soluble recombinant proteins. *Protein Sci.* **11**, 1714–1719 (2002).

14. Vincentelli, R., Canaan, S., Julien, O. V, Cambillau, C. & Bignon, C. Automated expression and solubility screening of His-tagged proteins in 96-well format. *Anal. Biochem.* **346**, 77–84 (2005).

15. Wang, Z. *et al.* Coupled selection of protein solubility in E. coli using uroporphyrinogen III methyltransferase as red fluorescent reporter. *J. Biotechnol.* **186**, 169–174 (2014).

16. Waldo, G. S., Standish, B. M., Berendzen, J. & Terwilliger, T. C. Rapid protein-folding assay using green fluorescent protein. *Nature* **17**, 691–695 (1999).

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

18. Klesmith, J. R., Bacik, J.-P., Wrenbeck, E. E., Michalczyk, R. & Whitehead, T. A. Trade-offs

between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc. Natl. Acad. Sci.* **114**, 2265–2270 (2017).

19. Foit, L. *et al.* Optimizing Protein Stability In Vivo. *Mol. Cell* **36**, 861–871 (2009).

20. Araya, C. L. *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci.* **109**, 16858–16863 (2012).

21. Lesley, S. A., Graziano, J., Cho, C. Y., Knuth, M. W. & Klock, H. E. Gene expression response to misfolded protein as a screen for soluble recombinant protein. *Protein Eng. Des. Sel.* **15**, 153–160 (2002).

22. Kraft, M. *et al.* An online monitoring system based on a synthetic sigma32-dependent tandem promoter for visualization of insoluble proteins in the cytoplasm of Escherichia coli. *Appl. Microbiol. Biotechnol.* **75**, 397–406 (2007).

23. Schultz, T., Martinez, L. & de Marco, A. The evaluation of the factors that cause aggregation during recombinant expression in E. coli is simplified by the employment of an aggregation-sensitive reporter. *Microb. Cell Fact.* **5**, 1–9 (2006).

24. Mendez-Perez, D., Gunasekaran, S., Orler, V. J. & Pfleger, B. F. A translation-coupling DNA cassette for monitoring protein translation in Escherichia coli. *Metab. Eng.* **14**, 298–305 (2012).

25. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–7 (2014).

26. Allen, S. P., Polazzi, J. O., Gierse, J. K. & Easton, A. M. Two novel heat shock genes encoding proteins produced in response to heterologous protein expression in Escherichia coli. *J.*

Bacteriol. **174**, 6938–6947 (1992).

27. Chuang, S. E., Burland, V., Plunkett, G., Daniels, D. L. & Blattner, F. R. Sequence analysis of four new heat-shock genes constituting the hslTS/ibpAB and hslVU operons in Escherichia coli. *Gene* **134**, 1–6 (1993).

28. Silva-Rocha, R. *et al.* The Standard European Vector Architecture (SEVA): A coherent platform for the analysis and deployment of complex prokaryotic phenotypes. *Nucleic Acids Res.* **41**, 666–675 (2013).

29. Grossman, A. D., Straus, D. B. & Walter, W. A. Simga 32 synthesis can regulate the synthesis of heat shock proteins in Escherichia coli. *Genes Dev.* **1**, 179–184 (1987).

30. Langelier, M., Servent, K. M., Rogers, E. E. & Pascal, J. M. A Third Zinc-binding Domain of Human Poly (ADP-ribose) Polymerase-1 Coordinates DNA-dependent Enzyme Activation. *J. Biol. Chem.* **283**, 4105–4114 (2008).

31. Rowling, P. J. E., Cook, R. & Itzhaki, L. S. Toward classification of BRCA1 missense variants using a biophysical approach. *J. Biol. Chem.* **285**, 20080–20087 (2010).

32. Luh, F. Y. *et al.* Structure of the cyclin- dependent kinase. *Lett. to Nat.* **389**, 999–1003 (1997).

33. Zanier, K. *et al.* Formation of well-defined soluble aggregates upon fusion to MBP is a generic property of E6 proteins from various human papillomavirus species. *Protein Expr. Purif.* **51**, 59–70 (2007).

34. Itzhaki, L. S., Otzen, D. E. & Fersht, a R. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260–288 (1995).

772 35. Jackson, S. E. & Fersht, A. R. Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state
773 transition. *Biochemistry* **30**, 10428–10435 (1991).
774 36. Rubin, A. F. *et al.* A statistical framework for analyzing deep mutational scanning data. *Genome*
775 *Biol.* **18**, 1–15 (2017).
776 37. Guerois, R., Nielsen, J. E. & Serrano, L. Predicting changes in the stability of proteins and
777 protein complexes: A study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387 (2002).
778 38. Raman, S. *et al.* Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins*
779 **77**, 89–99 (2009).
780 39. Adkar, B. V *et al.* Protein model discrimination using mutational sensitivity derived from deep
781 sequencing. *Structure* **20**, 371–381 (2012).
782 40. Zhang, S., Zubay, G. & Goldman, E. Low-usage codons in Escherichia, yeast, fruit fly and
783 primates. *Gene* **105**, 61–72 (1991).
784 41. Van Durme, J. *et al.* Accurate prediction of DnaK-peptide binding via homology modelling and
785 experimental data. *PLoS Comput. Biol.* **5**, (2009).
786 42. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–
787 401 (2016).
788 43. Otwinowski, J. Biophysical inference of epistasis and the effects of mutations on protein
789 stability and function. *Mol. Biol. Evol.* **35**, 2345–2354 (2018).
790 44. Nisthal, N. *et al.* Protein stability engineering insights revealed by domain-wide comprehensive
791 mutagenesis. *bioRxiv Biochem.* 1–48 (2018). doi:10.1101/484949
792 45. Schmiedel, J. M. & Lehner, B. Determining protein structures using deep mutagenesis. *Nat.*

Genet. **51**, (2019).

46. Rollins, N. J. *et al.* Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* **51**, (2019).

47. Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L. & Tawfik, D. S. The Stability Effects of Protein Mutations Appear to be Universally Distributed. *J. Mol. Biol.* **369**, 1318–1332 (2007).

48. Beadle, B. M. & Shoichet, B. K. Structural bases of stability-function tradeoffs in enzymes. *J. Mol. Biol.* **321**, 285–296 (2002).

49. K.Shoichet, B., A.Baase, W., Kuroki, R. & W. Matthews, B. A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 452–456 (1995).

50. Butt, T. R., Edavettal, S. C., Hall, J. P. & Mattern, M. R. SUMO fusion technology for difficult-to-express proteins. *Protein Expr. Purif.* **43**, 1–9 (2005).

51. Davis, G. D., Elisee, C., Newham, D. M. & Harrison, R. G. New Fusion Protein Systems Designed to Give Soluble Expression in Escherichia coli. *Biotechnol. Bioeng.* **65**, 382–388 (1999).

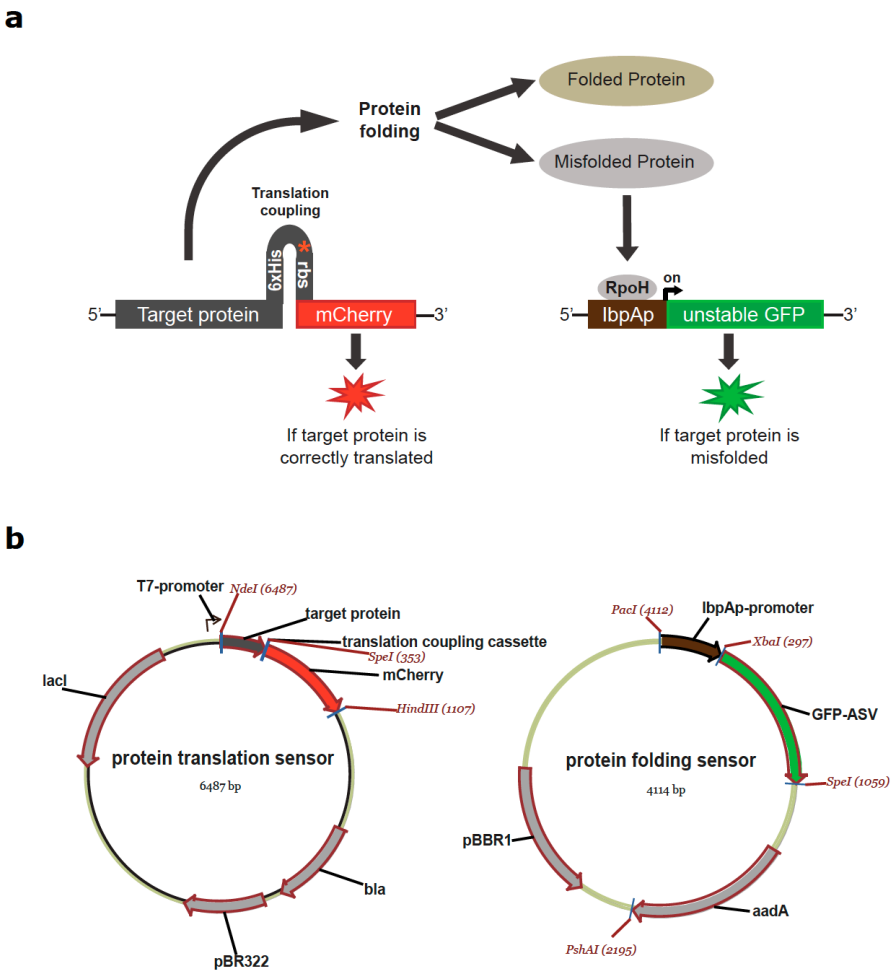
52. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ : 25 years of Image Analysis. *Nat. Methods* **9**, 671–675 (2012).

53. Magoc, T. & Salzberg, S. L. FLASH : fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).

54. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–360 (2012).

814

815



816

817

818 **Figure 1:** Schematic overview of a dual-reporter system with simultaneous monitoring of protein

819 translation and protein folding at the single cell level. (a) The translation sensor is comprised of the

820 gene of interest translationally coupled to the reporter protein mCherry. When the target gene is

821 correctly translated, the RNA polymerase unfolds the secondary structure and the mCherry gene is

822 transcribed resulting in a red fluorescent signal. The synthesized polypeptide chain then either folds

into a soluble protein conformation or it fails to fold, thereby typically forming protein aggregates that accumulate as inclusion bodies. Formation of inclusion bodies increases the cellular level of RpoH (heat shock sigma-factor σ^{32}). RpoH binds to the lbpA promoter in the protein folding sensor, initiating the expression of an unstable GFP variant, GFP-ASV yielding a green fluorescent signal. (b) Overview of the plasmids used for the protein translation and protein folding sensors.

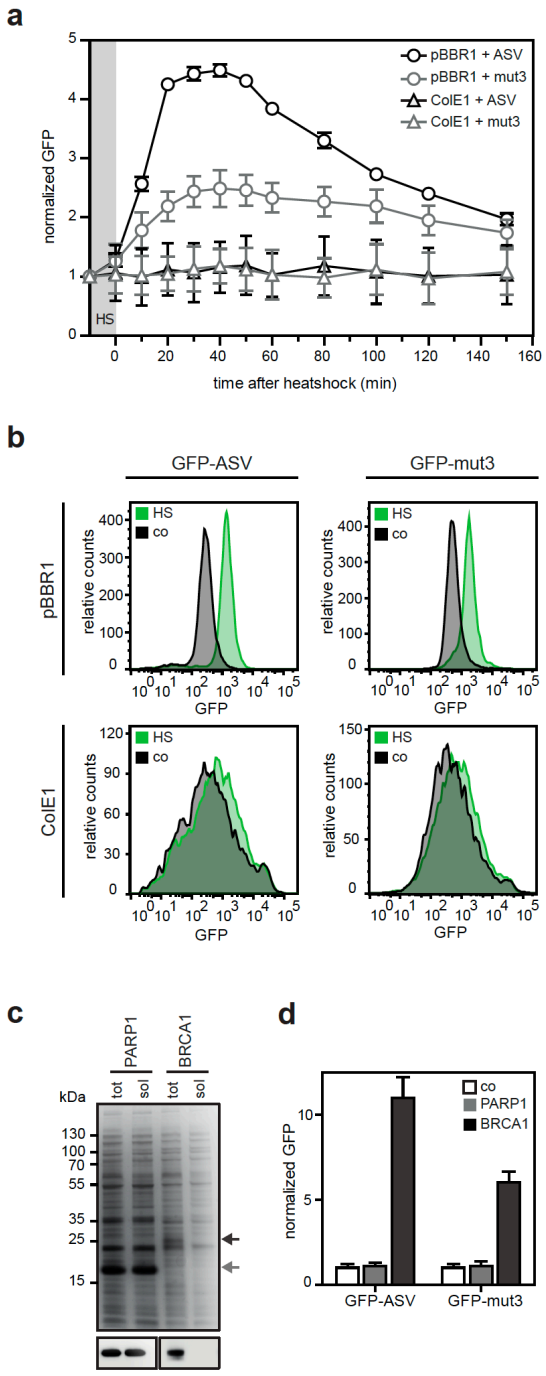


Figure 2: Optimization of the protein folding sensor plasmid to improve differentiation of heat shock response signals. (a) Monitoring of the heat shock response signals of protein folding sensors

(pSEVA441-IbpAp and pSEVA631(Sp)-IbpAp) with different origin of replications (ColE1 and pBBR1, respectively) and GFP variants (GFP-mut3 and GFP-ASV) after induction of the IbpA promoter. Changes in the GFP signal after induced heat shock (HS) are monitored using flow cytometry and the GFP signals in triplicates (average ± SD) are normalized to the respective background signal at each time-point. (b) FACS profiles for the GFP signals 60 min after induced heat shock for GFP-mut3 and GFP-ASV in plasmids with different origin of replications. Relative counts of GFP fluorescence intensities are shown from the analysis of 10,000 single cells. The heat shock induced (HS) GFP variants expressed from pBBR1 (pSEVA631(Sp)-IbpAp) shows well-defined and distinct peaks, which are easy to distinguish from the un-induced control plasmids (co). The GFP variants expressed from ColE1 (pSEVA441-IbpAp) resulted in very broad and not well-defined peaks making it difficult to distinguish between the heat shock induced plasmids and the control. (c) SDS-PAGE and immunoblot analysis of total (tot) protein yield and soluble protein (sol) after fractionated cell disruption of two human proteins, PARP1-BRCT and a truncated version of BRCA1-BRCT, shows high expression of a soluble PARP1-BRCT protein, and an insoluble BRCA1-BRCT protein. (d) Flow cytometry analysis 60 min after protein induction of the co-expression of PARP1-BRCT and BRCA1-BRCT with the pSEVA631(Sp)-IbpAp-GFP-ASV and pSEVA631(Sp)-IbpAp-GFP-mut3 plasmids. The soluble PARP1-BRCT does not initiate a heat shock response and results in a low green fluorescent signal, whereas the insoluble BRCA1-BRCT protein triggers the heat shock response causing a high green fluorescent signal. The pSEVA631(Sp)-IbpAp-GFP-ASV plasmid has an improved signal-to-noise ratio and is preferred over the pSEVA631(Sp)-IbpAp-GFP-mut3 plasmid. All measurements were performed with n=3 and n≥3, respectively.

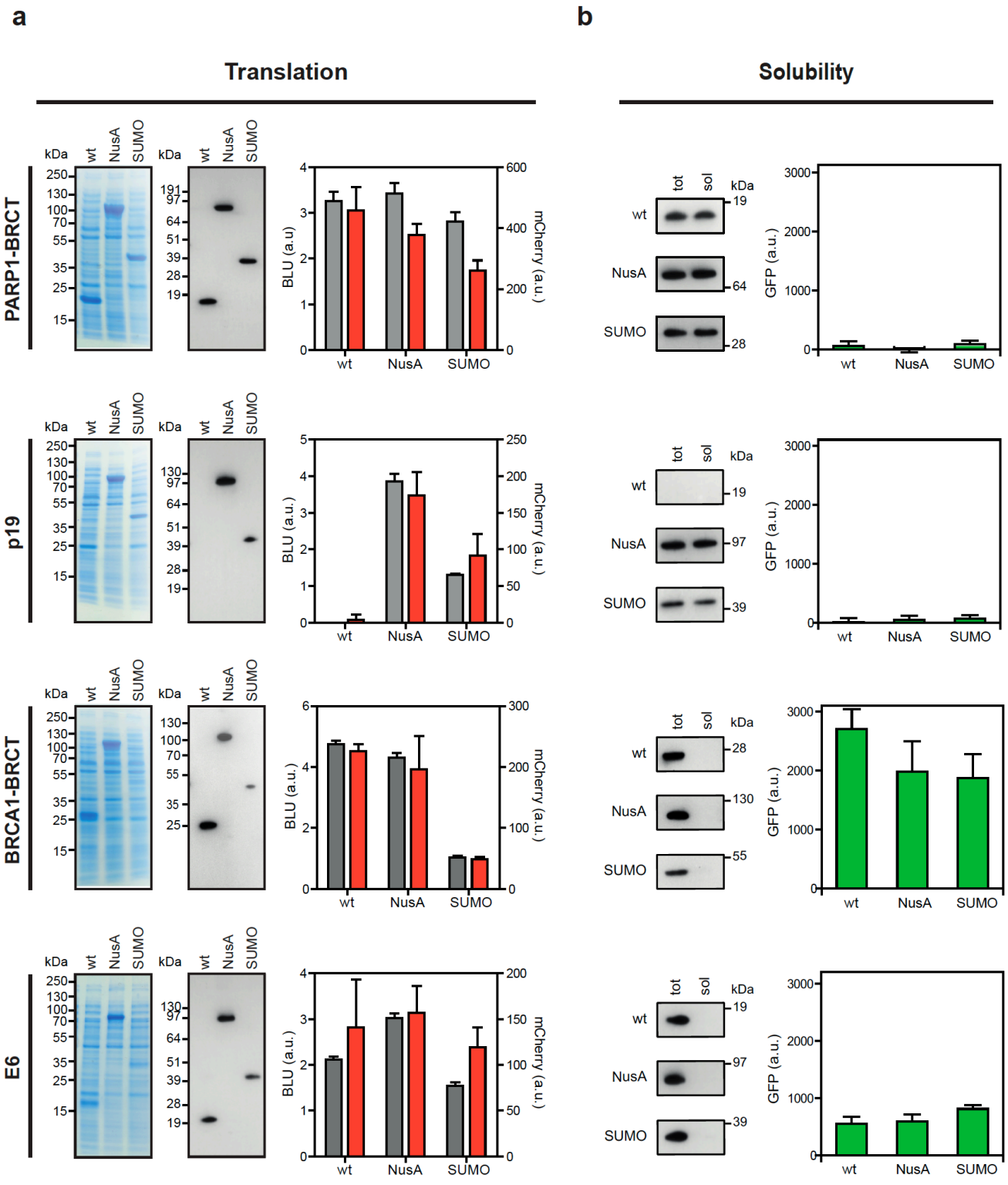
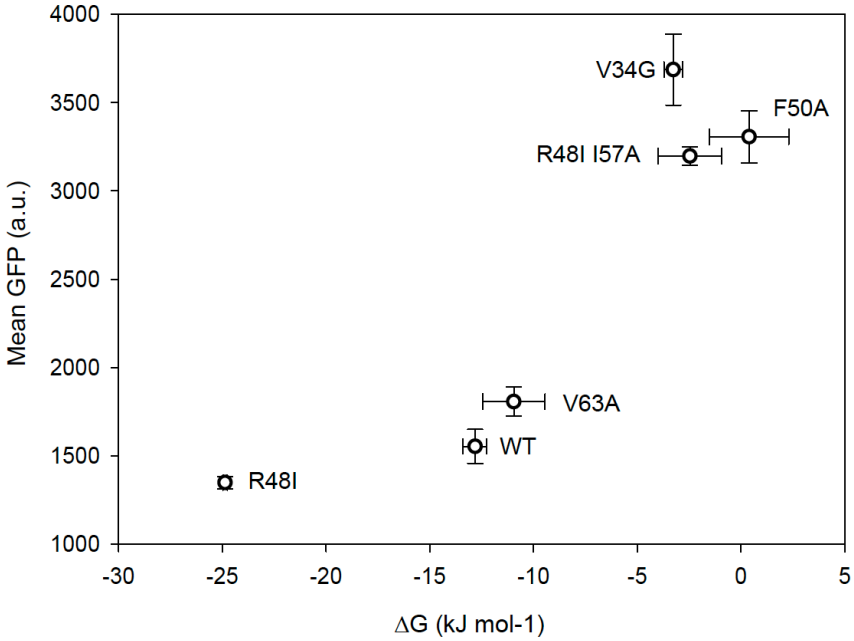


Figure 3: Validation of the dual protein translation and misfolding biosensor. The solubility tags NusA and SUMO were fused to four proteins; PARP1-BRCT, p19, a truncated BRCA1-BRCT, and E6, with known propensities for misfolding. (a) The proteins were translationally coupled to the fluorescent protein mCherry to monitor the translation using FACS. A minimum of five independent samples was analysed for each plasmid combination. Protein expression was analysed by SDS-PAGE analysis and quantified from Western blots (grey) (BLU = biochemical luminescence unit) and correlated to the mean mCherry fluorescence signal from the analysis of 10,000 cells (red). (b) Western blot analysis of total protein yield (tot) and soluble protein (sol) after fractionated cell disruption in association with the quantified GFP response signal for insoluble protein.



870

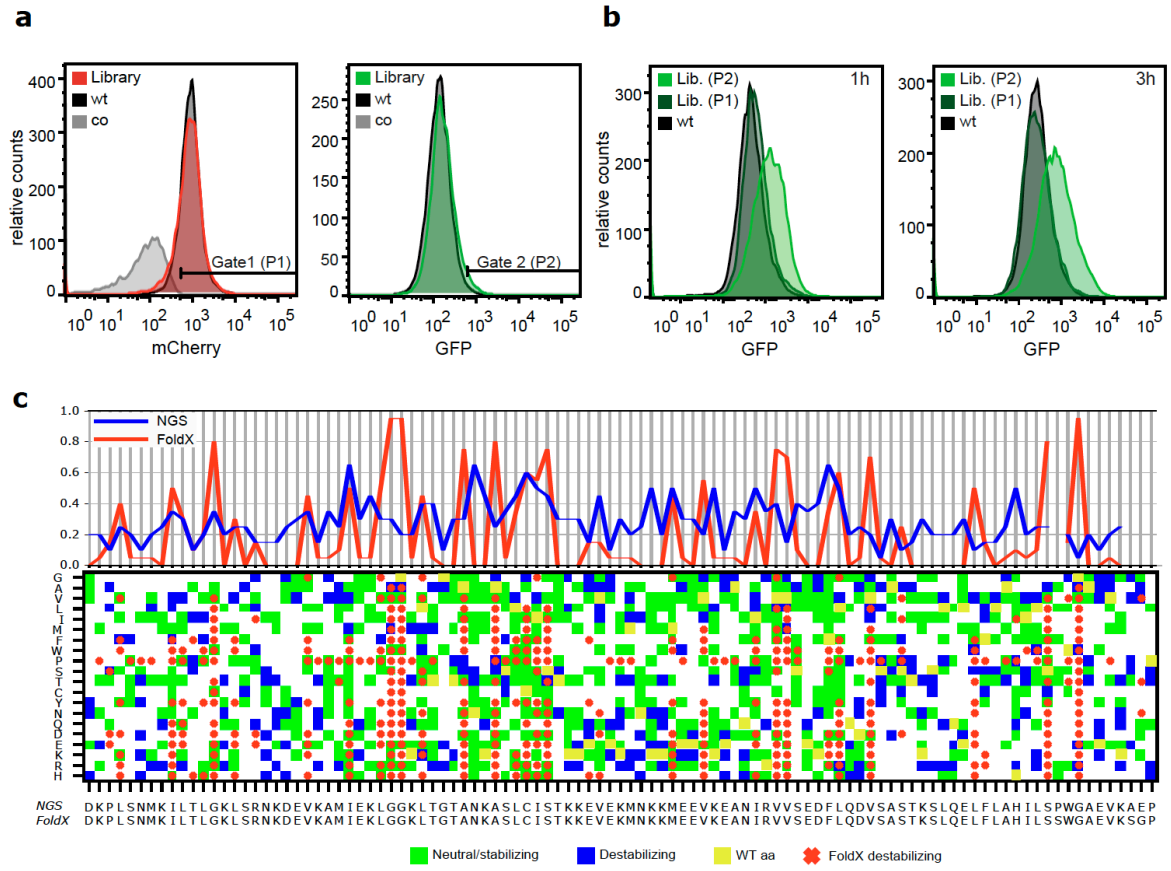
871 **Figure 4:** Correlation between GFP fluorescence and protein stability. Six CI2 variants were co-

872 expressed with the protein folding sensor (pSEVA631(Sp)-lbpAp-GFP-ASV), and the GFP fluorescence

873 was analysed by FACS. The average mean GFP fluorescence was compared to the Gibbs free energy of

874 unfolding (ΔG_{unf}) determined from global fits of thermal and chemical unfolding of each protein. All

875 measurements were determined in triplicates with standard deviations.



876

Figure 5: FACS sorting and deep mutational scanning to identify variants of PARP1-BRTC with decreased protein folding. (a) FACS sorting of PARP1-BRCT mutant library (red and green), PARP1-BRCT WT (black), and the translation sensor plasmid without a gene inserted (grey). Cells were sorted for high translation levels (Gate 1) and degree of protein misfolding (Gate 2). (b) The sorted cells were grown overnight and analysed by flow cytometry one and three hours after protein expression was induced. (c) Top: Ratio between the number of destabilizing mutations and the number of total mutations for each amino acid residue for both FoldX (red) and experimental data (blue). Bottom: Matrix plot indicating if an amino acid change (y-axis) of the sequence (x-axis) was destabilizing according to the high-throughput sequencing data as well as for FoldX calculations. For the experimental data, green and blue squares indicate neutral/stabilizing and destabilizing mutations, respectively. Yellow marks the wildtype to wildtype mutants, and white marks mutations with no experimental readout. Red x's indicate destabilizing mutations according to FoldX, with a cut-off of 3 kcal/mol. All squares without red x's are predicted to be neutral or stable mutants. (d) Receiver Operating Characteristic analysis of sequencing data and predicted FoldX $\Delta\Delta$Gs. The sequencing data provides the mutation specific labels (blue vs green in Figure 5C) and the $\Delta\Delta$Gs predicted from FoldX are the mutation specific scores. (e) Structural visualization of stable vs destablilizing sequence positions of the PARP1-BRCT structure based on the experimental data. Blue residues that destabilize the protein have a $N_{destable}/N_{total} \geq 0.2$, while the remaining are colored red. (f) Scoring of 20.000 structural decoys based on the experimental data. The plot shows the Spearman's correlation coefficient, ρ, that quantifies the correlation between residue depth and mutational tolerance based on the experimental data, as well as a structural quality measure defined by the structural Global

898 Distance Test – Total Score (GDT-TS) score, where one corresponds to a native or near native
899 structure. Here, the mean ρ is plotted for structures binned to the closest 0.1 GDT-TS bins. The error
900 bars represent standard deviations for the individual bins.
901
902
903
904

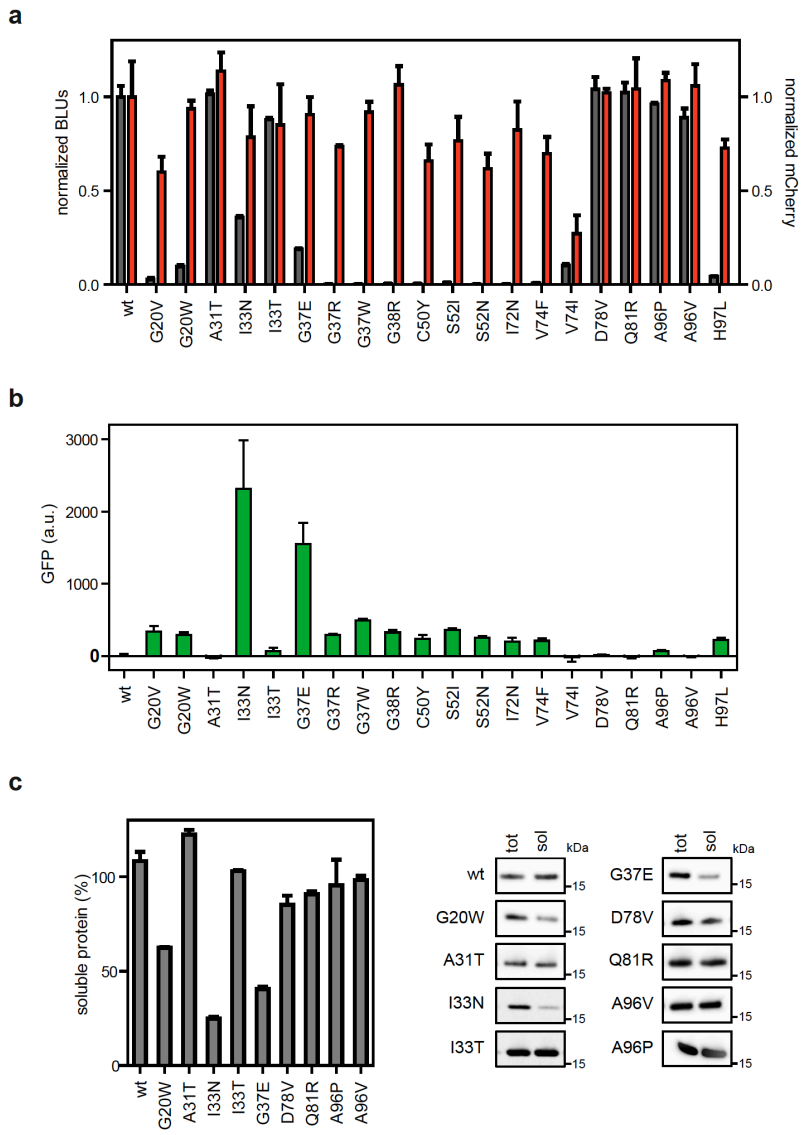


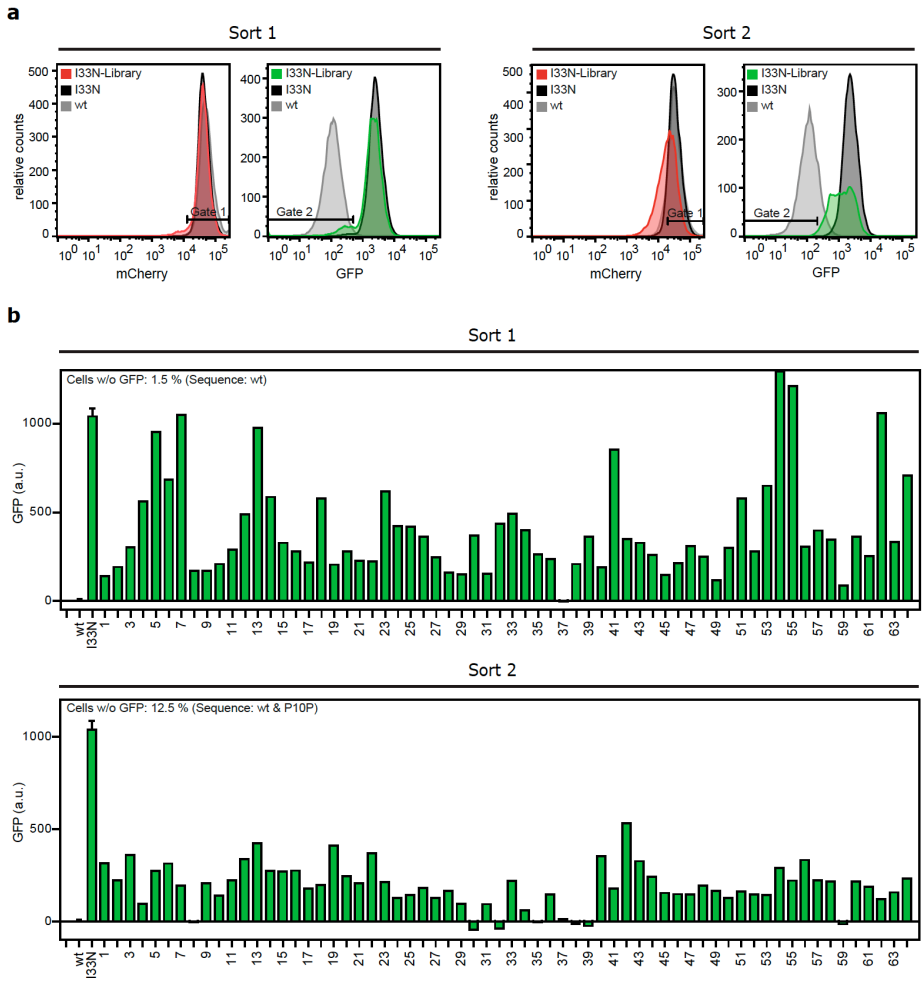
Figure 6:

PARP1-BRCT mutants with changed folding properties identified from a randomly generated mutant library using the dual-reporter system. (a) Correlation between translation levels of 20 PARP1-BRCT mutants quantified from Western blots (grey) and flow cytometry analysis of mean mCherry fluorescence (red), each normalized to the WT signal. (b) GFP levels analysed using flow cytometry as a measure for protein solubility and folding properties. (c) Percentage of soluble protein determined

912 by Western blot for the 9 PARP1-BRCT mutants with a detectable GFP response signal. Western blot

913 analysis of total protein yield (tot) and soluble protein (sol) after fractionated cell disruption.

914



915

916 **Figure 7:**

917 Identification of protein variants with improved folding properties from a PARP1-BRCT-I33N mutant

918 library using the dual-reporter system. A random PARP1-BRCT-I33N mutant library was co-expressed

919 with the protein folding sensor. The cell populations were analysed using FACS one hour after IPTG

920 induced protein expression. (a) FACS analysis of PARP1-BRCT WT, PARP1-BRCT-I33N, and the PARP1-

921 BRCT-I33N mutant library, where the mCherry signal correlates with the translation level of PARP1-
922 BRCT, while the GFP fluorescence is a measure of folding properties. Two gates are defined for sorting
923 populations with high translation and low GFP fluorescence, thus with improved folding properties
924 compared to PARP1-BRCT-I33N. A shift is observed in GFP signal distribution and intensities between
925 the two rounds of sorting, showing that it is possible to enrich the population with low GFP clones
926 after multiple rounds of sorting. (b) Single clones sorted after each round of sorting with 1.5 % or 12.5
927 % overlapping with the PARP1-BRCT WT GFP signal. The single clones were sequenced and all were
928 reverted back to PARP1-BRCT WT, except one silent mutation, P10P, found after the second sorting.