## ARTICLE

Check for updates

# DnaK response to expression of protein mutants is dependent on translation rate and stability

Signe Christensen [1✉], Sebastian Rämisch[2] & Ingemar André [1✉]

Chaperones play a central part in the quality control system in cells by clearing misfolded and aggregated proteins. The chaperone DnaK acts as a sensor for molecular stress by recognising short hydrophobic stretches of misfolded proteins. As the level of unfolded protein is a function of protein stability, we hypothesised that the level of DnaK response upon over-expression of recombinant proteins would be correlated to stability. Using a set of mutants of the λ-repressor with varying thermal stabilities and a fluorescent reporter system, the effect of stability on DnaK response and protein abundance was investigated. Our results demonstrate that the initial DnaK response is largely dependent on protein synthesis rate but as the recombinantly expressed protein accumulates and homeostasis is approached the response correlates strongly with stability. Furthermore, we observe a large degree of cell-cell variation in protein abundance and DnaK response in more stable proteins.

[1] Department of Biochemistry and Structural Biology, Lund University, Lund, Sweden. [2] Max Planck Unit for the Science of Pathogens, Berlin, Germany.
✉email: signechristensen89@gmail.com; ingemar.andre@biochemistry.lu.se

The production of functional proteins typically requires folding into well-defined three-dimensional structures. However, protein synthesis can also lead to the accumulation of unfolded, misfolded, or aggregated protein due to their marginal stability. As aggregated and misfolded proteins can have toxic effects on the cell there is a substantial fitness cost associated with them. To reduce the harm caused by aberrant proteins all living organisms have developed a sophisticated protein quality control (PQC) network. The PQC is a collection of chaperones and proteases that assists with refolding of misfolded proteins, protects from irreversible aggregation by the formation of inclusion bodies or, eventually, direct misfolded or aggregated proteins for degradation. Due to the ubiquitous nature of the PQC the involved proteins are required to be highly promiscuous. E.g., hsp70 (known as DnaK in *E. coli*) binds to hydrophobic stretches approximately five residues long, that are exposed upon misfolding[1]. These short hsp70/DnaK-binding motifs are present in most proteins and allow for recognition of unfolded or misfolded proteins, in a matter largely independent of sequence and structure. Recognition of unfolded and misfolded proteins by DnaK is one of the major regulators of the transcription factor $\sigma^{32}$ and hence PQC network in *E. coli*, although $\sigma^{32}$ is also regulated by, e.g., GroEL/S[2]. Under low-stress conditions DnaK binds to $\sigma^{32}$ targeting it for degradation by the metallo-protease FtsH[3]. In response to stress, DnaK binds preferentially to misfolded protein instead of $\sigma^{32}$ allowing $\sigma^{32}$ to act as a transcription factor for the expression of DnaK and other PQC proteins. Other proteins in the PQC network that are regulated by $\sigma^{32}$ includes chaperones involved in protein folding and disaggregation (ClpB, DnaJ, GroEL and GroES), cytosolic proteases that degrade proteins (ClpP, ClpX and Lon), as well as the membrane-bound FtsH involved in the regulation of $\sigma^{32}$ levels. Misfolded proteins can also aggregate and be directed into inclusion bodies. DnaK interacts with aggregated proteins at the surface of inclusion bodies and appear to have a critical role in reducing cytotoxicity and delivering misfolded proteins to ClpB[4,5].

Whereas it is well-established that misfolded proteins trigger a response from the PQC the molecular properties of the client proteins that impact its regulation are not fully understood. One open question is whether this response is primarily determined by the concentration of misfolded/unfolded protein, or if there is a sequence-dependent component to the response beyond chaperone recognition sequences. The lower the stability of a protein translated within the cell, the less will be found in the folded native state. Consequently, if the PQC is primarily dependent on the concentration of unfolded or misfolded protein a correlation between protein stability and PQC response should be expected. Whether such a simple relationship exists is not clear, although upregulation of GroEL, c62.5 (hsp90 homolog) and DnaK[6], has been observed upon introduction of mutations that destabilize wild-type sequences. A combination of computational models and in vitro studies performed by Sekhar et al. suggest that DnaK binds preferentially to unstable proteins at folding equilibrium, but to slow-folding proteins while folding[7]. As DnaK can interact both co-translationally with the unfolded nascent peptide chain and with misfolded protein[8] a similar process could occur in vivo. However, no in vivo studies of the correlation between folding rates and DnaK response have been presented.

Another central question that is not fully understood is how the chaperone response is regulated by protein production. Overexpression of proteins in *E. coli* results in increased levels of chaperones[9,10], but it is not clear to what extent this response is controlled by the level of protein translation or simply the abundance of misfolded/unfolded proteins in the cytosol. There is a negative correlation between the level of expression of a protein in *E. coli* and its dependency on DnaK for folding[11]. Thus,

proteins that are difficult to fold might express at lower levels. This indicates that the cell has a limited capacity to process misfolded/unfolded proteins and perhaps also a limited range to regulate the chaperone response. Mutations do not only affect the folding stability and kinetics of proteins, but also the level of expression. Therefore, to understand the regulation of the chaperone response also protein expression and abundance must be considered.

The presence of misfolded/unfolded proteins carries a fitness cost to the cell[12–15]. By modifying the abundance of misfolded proteins, chaperones can modify the rate of protein sequence evolution[11,16–22]. Hence, chaperones, like DnaK, can enable proteins to evolve faster. However, there is still an associated fitness cost as triggering the PQC response consumes resources; both directly by consumption of ATP and by occupying the synthesis machinery for production of chaperones and proteases forming the PQC network. In theory, the interaction of proteins with the PQC network can also impact which sequences are fixated during evolution. Protein variants triggering a large PQC response would be expected to have lower fitness and thus be selected against. In light of this, information about how molecular variation impacts DnaK interaction and the PQC can provide valuable insight into the role of the PQC on protein evolution.

As summarised above the components of the PQC system and the regulation network are well described, but it has yet to be studied in detail how the biophysical properties of the client protein impact the chaperone response. Thus, in this study, we investigated the effect of molecular variation on the PQC. More specifically, we focused on the following questions: (i) What is the relationship between protein stability and DnaK response? (ii) Is the DnaK response controlled by the level of protein translation, the amount of unfolded protein or both? (iii) How much does the PQC response vary with small variation in biophysical properties of protein variants? (iv) How much does the PQC response vary between individual cells? We used superfolder Green Fluorescent Proteins (sfGFP) expressed from the stress-inducible DnaK promoter to evaluate the effect on the PQC response upon introducing stability-altering point mutations to the λ repressor construct, N102LT. We found that induction of the DnaK promoter is dependent on protein translation rates early after onset of recombinant protein synthesis, but dependent on protein stability when homeostasis is reached. We further found that protein stability has a significant impact on the accumulation of protein in the cell due to increased removal of proteins from the cytosol with low stability and that the response varies substantially between cells. The investigation also indicates how the fluorescence system can be used as a high-throughput stability assay.

## Results

**Monitoring protein accumulation and DnaK response with fluorescent reporters**. To study how the biophysical properties of proteins impact the stress response in *E. coli*, we constructed a reporter plasmid system to simultaneously monitor the abundance of recombinantly expressed protein and chaperone response upon recombinant expression (Fig. 1a). It contains two fluorescent reporters: (i) The red fluorescent protein tagRFP is fused to the N-terminal of a protein of interest (POI), and expression of the POI-tagRFP fusion protein is controlled by the IPTG inducible *lac* promoter. The fusion protein allows for monitoring the abundance of recombinant POI in the cell. In addition, the folding state of the POI could impact the folding of tagRFP and therefore acts as an additional stability reporter, as previously shown for GFP-tagged proteins[23–25]. (ii) Expression of Superfolder Green Fluorescent Protein (sfGFP) is controlled by the DnaK promoter to enable monitoring of the stress response in
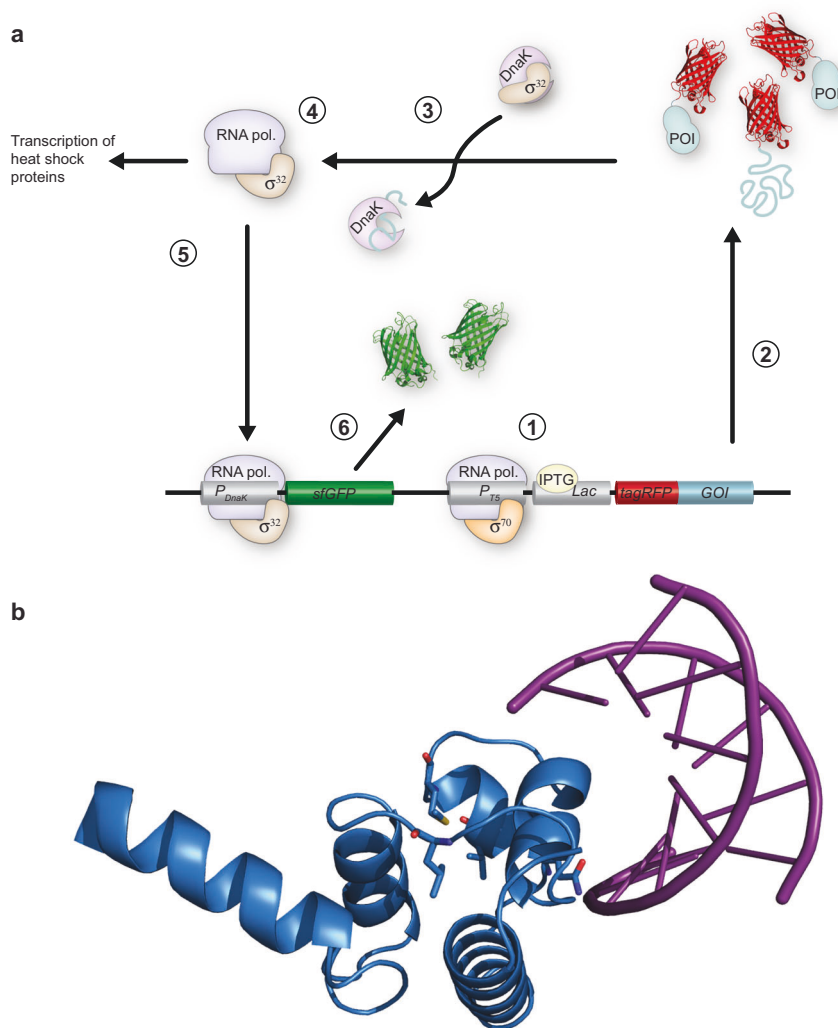
**Fig. 1 A fluorescent-based reporter system for protein abundance and DnaK response. a** Overview of reporter system: (1) The addition of IPTG allows for transcription of the GOI (gene of interest)-tagRFP by RNA polymerase in a complex with the transcription factor $\sigma^{70}$. (2) Following transcription of GOI-tagRFP the protein of interest, POI-tagRFP is synthesised. (3) DnaK releases the transcription factor $\sigma^{32}$ and binds to unfolded/misfolded POI-tagRFP. (4) $\sigma^{32}$ binds to RNA polymerase. (5) Binding of $\sigma^{32}$ to RNA polymerase allows for transcription of endogenous PQC proteins, such as DnaK. The RNA polymerase—$\sigma^{32}$ complex will also bind to the DnaK promoter on the reporter plasmid and transcribe the *sfGFP* gene. (6) Following transcription of the *sfGFP* gene, the protein sfGFP will be synthesised. **b** The N-terminal domain of the λ repressor (blue) bound to DNA (purple). Mutated residues are shown as sticks (blue). PDB:1lmb[56].

individual cells. The PQC network consists of many different components, thus the GFP signal gives only a partial picture of the overall state. But since DnaK is a central component in the response to folding stress, we consider it a meaningful proxy for the PQC network. This is further emphasised by the observation of a strong correlation between the expression of DnaK and the green fluorescence signal (Fig. S1). However, it is worth noticing that synthesis and degradation of PQC proteins are tightly regulated in the cell[26–28]. As the degradation of GFP is decoupled from the degradation of DnaK the GFP signal does not report on the current level of DnaK in the cell, but rather the accumulated amount of DnaK produced until that point. However, by evaluating the rate of GFP production—which is correlated to the rate of DnaK synthesis—we can measure when the DnaK response is triggered. The protein abundance reporter (tagRFP) and the DnaK reporter (sfGFP) are located on the same plasmid. This enables normalization of the chaperone signal with protein abundance so that cell–cell variation and variation in plasmid copy numbers can be accounted for. It is important to note that

even though a DnaK promoter is utilised for the expression of GFP, the endogenous DnaK promoter is still in place ensuring a functional PQC network in the cells. The assay is carried out in the *E. Coli* strain Bl21 (DE3) pLysS, which lacks the *lon* protease that is responsible for ~50% of the protease activity in *E. Coli*[29,30]. Consequently, more folded and unfolded/misfolded protein is expected to be present in the cytosol in the protease-deficient strain and the effect of misfolding on DnaK response will be amplified.

As a model system for investigating the correlation between DnaK response and protein stability, we use several variants of the λ repressor construct, N102LT, with different thermodynamic stability. N102LT is a 128 amino acid long protein consisting of the N-terminal domain of the λ repressor extended by the C-terminal domain of the protein arc that protects the fusion protein against proteolysis[31]. It has previously been used to study the heat shock response in *E. coli*[32–36]. To separate effects caused by changes in stability from effects caused by site-specific structural factors, it is beneficial to also evaluate variants with
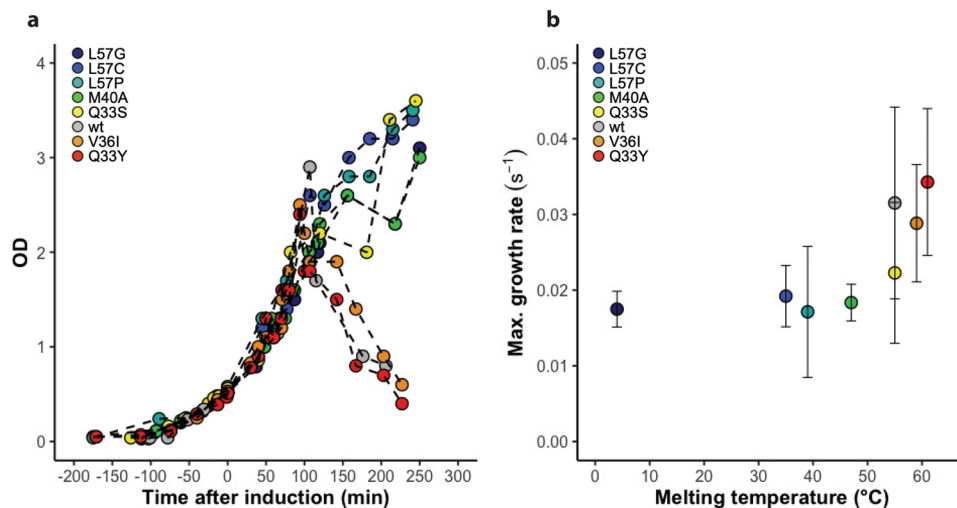
**Fig. 2 Growth of cells expressing N102LT variants. a** $OD_{600}$ as a function of time after induction of protein synthesis for cells expressing the variants: L57G (dark blue), L57C (blue), L57P (cyan), M40A (green), Q33S (yellow), V36I (orange), Q33Y (red) and wt (grey). One of three replicates is shown for each variant. **b** Maximum cell growth rates are estimated based on fitting logistic growth models and extracting the maximum rate. Errors bars represent the uncertainty due to the fitting of a growth model to the data.

mutations at the same position. In this study, we therefore included seven variants with single point mutations in four different positions: Q33, V36, M40 and L57 with in vitro melting temperatures from below 10 to 61 °C in addition to wild-type N102LT (Tm = 55 °C). Positions L57, M40 and V36 are part of the hydrophobic core, whereas Q33 is surface exposed and is part of the cognate DNA recognition (Fig. 1b). Mutations in position L57 and M40 result in four variants (L57C, L57G, L57P and M40A) with reduced thermostability compared to the wt sequence, while V36I results in increased stability. Mutating the surface accessible Q33 to either serine (Q33S) or tyrosine (Q33Y) retains or increases thermostability, respectively.

Changes in the ability of the N102LT variants to bind DnaK could potentially impact the observed stress response. To check whether the introduced mutations cause significant changes to the DnaK binding properties we use the ChaperISM algorithm to predict DnaK binding properties for the eight variants[37] (Fig. S2). N102LT wt is predicted to have three DnaK binding sites. The introduction of the Q33Y mutation increases the DnaK binding propensity slightly around position 33, but without introducing a new DnaK binding site. Mutating position L57 reduces the DnaK binding propensity, but without disrupting a DnaK binding site.

**Point mutations in N102LT result in differences in accumulation of protein, DnaK response and cell morphology**. To monitor changes in protein abundance and DnaK response the N102LT variants were expressed from the reporter plasmid in *E. coli* and grown in liquid media. Synthesis of the N102LT variants was induced with IPTG during exponential growth. Flow cytometry was employed to simultaneously monitor the abundance of protein (red fluorescence), DnaK transcription (green fluorescence), as well as morphological changes (forward and side scatter) on a single-cell level. In addition to analysis by flow cytometry, cell growth was monitored by optic density at 600 nm ($OD_{600}$) (Fig. 2a, Fig. S3). Cells harbouring all variants grow the first 120 min after induction after which cells expressing the wt as well as Q33Y and V36I show a drop in $OD_{600}$ (Fig. 2a, Fig. S3) indicating that these cells might be dying. To study the effect of the thermostability of the protein variants on growth rate we fitted a logistic growth function to the individual growth curves and calculated the maximum growth rate during the time course (Fig. 2b and Supplementary text). The maximal growth rate

correlates well with melting temperature, indicating that the expression of less stable protein variants may reduce maximum cell growth. However, due to uncertainty in the maximum growth rate due to the experimental uncertainty in the cell density measurements at later timepoints, more precise growth curve experiments will have to be carried out to validate this result.

Proteins can be cleared from the cytosol by protease activity or the formation of insoluble deposits like inclusion bodies. Upon aggregation into inclusion bodies, the fluorescence signal of the fusion protein will become "dark", resulting in a reduction in fluorescence signal[38]. We characterized the amount of soluble and insoluble protein expression of protein variants at various time points. While there is a significant amount of insoluble protein (Fig. S4), the proportion of insoluble to soluble protein is similar for all variants independent of thermostability. Hence, aggregation into inclusion bodies is expected to have a similar impact on all mutants and is thus unlikely to complicate the conclusions on variant-dependent effects drawn in this study.

**Accumulation of protein is stability-dependent**. To monitor the abundance of protein in the cell we followed the development of red fluorescence over time. The first 30 min after induction with IPTG the red fluorescence intensity is close to constant followed by a continuous increase throughout the four hours for which protein production is monitored (Fig. 3a, Fig. S5). Due to the long maturation time of tagRFP ($t_{50}$ ~ 42 min[39]) the low red fluorescence levels in the first 30 min do not necessarily indicate a lack of protein synthesis. Throughout the entire experiment, the red fluorescence correlates well with variant melting temperature, with a Spearman rank correlation of 0.81 after 120 min (Fig. 3b, c). At that point, variants with melting temperatures higher than the wt have red fluorescence intensities double to the variants with lower stabilities, with the exception of one replicate of Q33Y (Fig. 3c). Looking at different mutations at the same site we observe that L57G and L57C have similar fluorescence, while the more stable L57P variant has an increased value. This indicates that the decreased expression in the former variants is due to stability rather than a consequence of mutating a particular position. After 120 min the variation in red fluorescence increases both between variants, but also between replicates of individual variants. This is especially prominent among the Q33S and the
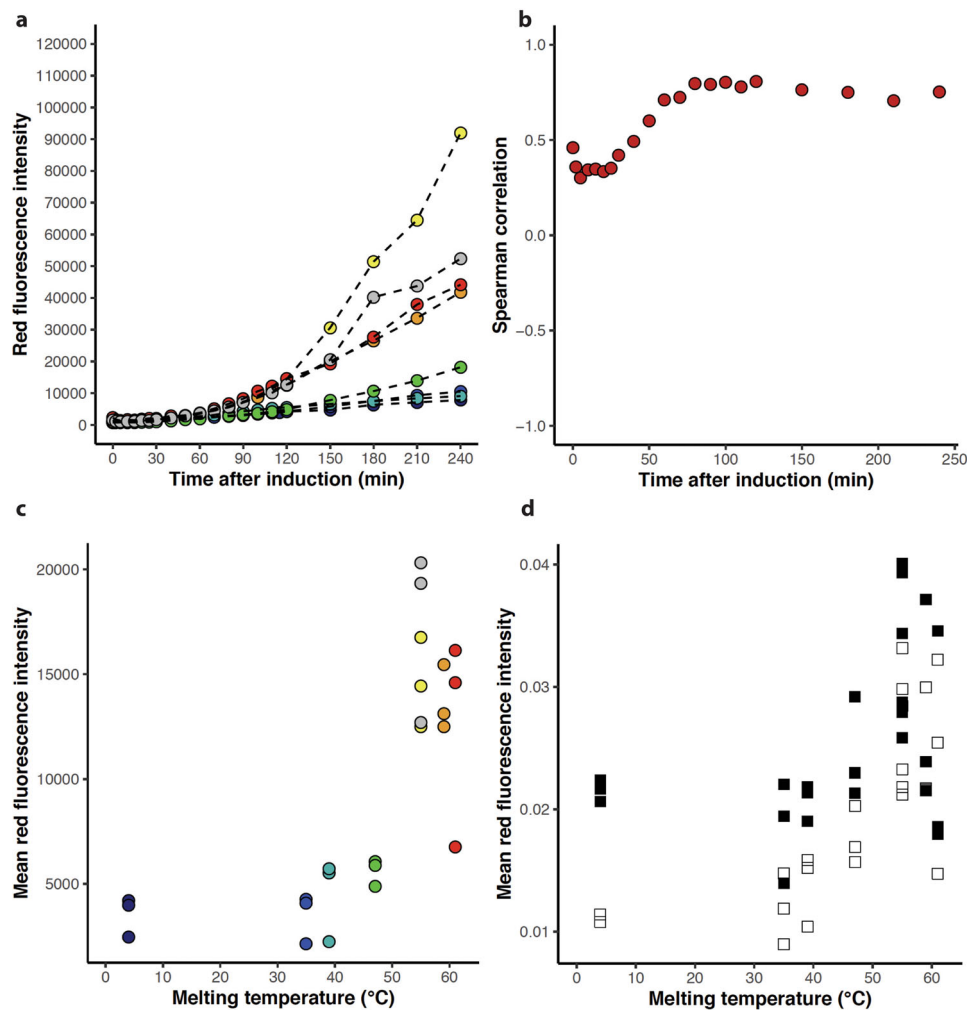
**Fig. 3 Abundance of tagRFP fusion proteins monitored by red fluorescence. a** Mean red fluorescence intensity of cells expressing variants: L57G (dark blue), L57C (blue), L57P (cyan), M40A (green), Q33S (yellow), V36I (orange), Q33Y (red) and wt (grey). One of three replicates are shown for each variant. **b** Spearman correlation between mean red fluorescence and variant melting temperature at each time point. Calculations are based on three replicates of each variant. **c** Mean red fluorescence intensity at the 120 min time point for the eight N102LT variants with different melting temperature. Spearman correlation = 0.81. Colour scheme as in (**a**). **d** Early (30–60 min, black) and late (60–90 min, white) relative rates of red fluorescence calculated for the eight N102LT variants with different melting temperature. Spearman correlation between early rate and variant melting temperature is 0.33. Spearman correlation between late rates and variant melting temperature is 0.76.

wt, where the change in red fluorescence after 240 min is ~1.3 and ~2 fold between the three replicates.

**Variation in protein accumulation is due to differences in rate of protein degradation rather than synthesis.** The amount of accumulated protein is a consequence of the opposing effects of synthesis and degradation. Variation in abundance between mutants can thus be explained by differences in protein synthesis rate, protein degradation rate or both. We estimated the relative rate of red fluorescence change (change in fluorescence intensity divided by the intensity) between 30 and 60 min ("early" rates), and between 60 and 90 min ("late" rates) after induction. Both late and early rates vary with variant (Fig. 3d). Interestingly, for variants less stable than wt the late rates are significantly reduced compared to the early rates. Assuming degradation is minimal early after the onset of protein synthesis, this indicates that eventually, the rate of degradation surpasses the rate of synthesis for these variants. For the more stable variants (wt, Q33S, Q33Y and V36I) the relative rate is not reduced or only minimally reduced after 60 min, which indicates that less protein is degraded in the cells harbouring

these variants. The late rates correlate with thermostability with a spearman rank correlation of 0.76 (Fig. 3d). Taken together, our results suggest that the difference in relative rate is most likely due to stability-dependent degradation of variants rather than a difference in the rate of translation.

The measured synthesis rate can be impacted by cell growth and division, which results in dilution of the protein concentration[40,41]. We simulated the impact of dilution on early and late rates by using the experimental growth curves (see Supplementary text). The predicted impact is small for the early rate (standard deviation of 2.7%). The effect is larger for the late rates due to a larger variation in growth rates (standard deviation of 12%) but still small. The effect of dilution on late rates is more difficult to estimate due to the increased importance of degradation in this time window.

The fluorescence signal is also reduced by aggregation into the "dark" state of inclusion bodies. The aggregation rate is expected to be similar for the different variants based on the solubility data (Fig. S4) and the small sequence differences between mutants. Hence, we attribute differences in relative fluorescence rates between variants largely to degradation.
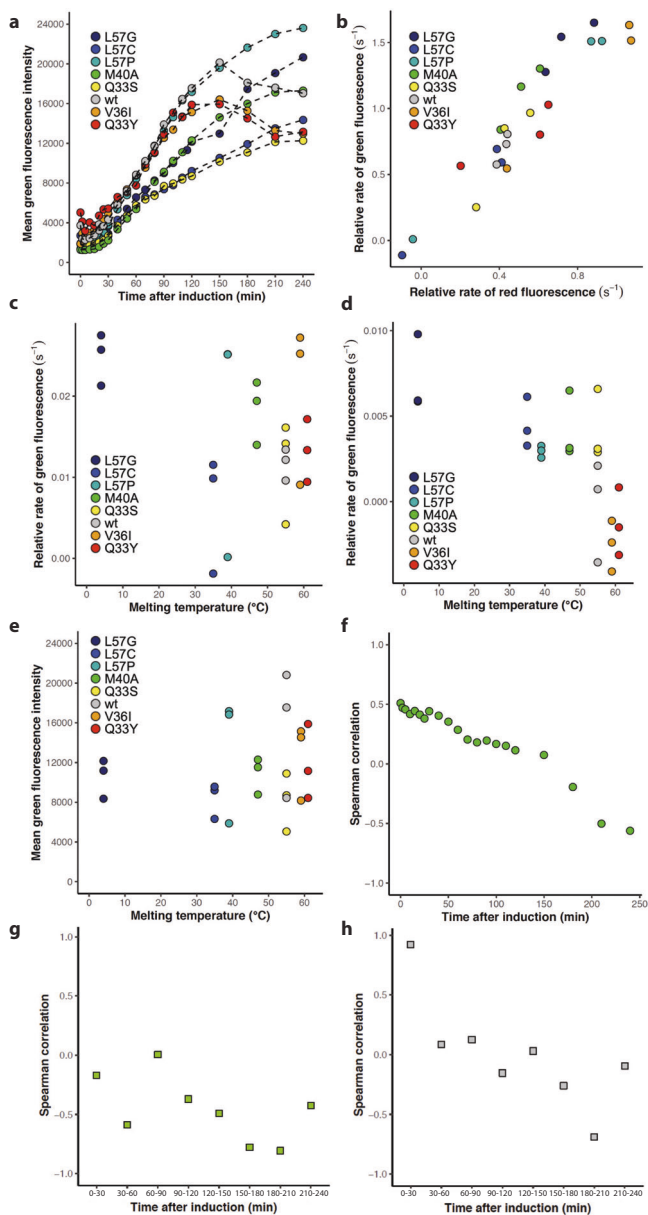
**Fig. 4 Transcription from the DnaK promoter monitored by green fluorescence. a** Mean green fluorescence intensity of cells expressing variants: L57G (dark blue), L57C (blue), L57P (cyan), M40A (green), Q33S (yellow), V36I (orange), Q33Y (red) and wt (grey). One of three replicates is shown for each variant. **b** The initial (0–30 min) relative rate of green fluorescence as a function of the initial rate of red fluorescence. Spearman correlation = 0.92. Colour scheme as in (**a**). **c** The initial (0–30 min) relative rate of green fluorescence was calculated for the eight N102LT variants with different melting temperatures. Spearman correlation between early rate and variant melting temperature is 0.29. Colour scheme as in (**a**). **d** The relative rate (150–180 min) of green fluorescence was calculated for the eight N102LT variants with different melting temperatures. Spearman correlation between rate and variant melting temperature is 0.71. Colour scheme as in (**a**). **e** Mean green fluorescence intensity at the 120 min time point for the eight N102LT variants with different melting temperatures. Spearman correlation between green fluorescence intensity and variant melting temperature is 0.11. Colour scheme as in (**a**). **f** Spearman correlation between accumulated mean green fluorescence and variant melting temperature at each time point. Correlations are calculated based on three replicates of each variant. **g** Spearman correlation between rate of green fluorescence in the indicated time interval and variant melting temperature. Correlations are calculated based on three replicates of each variant. **h** Spearman correlation between rate of green fluorescence and rate of red fluorescence in the indicated time interval. Correlations are calculated based on three replicates of each variant.

green and/or red fluorescence instead of total fluorescence we eliminate the effect of different maturation times.

**Protein translation alone triggers DnaK transcription**. The initial rate (0–30 min after IPTG induction) of green fluorescence correlates strongly with the initial rate of red fluorescence (spearman rank correlation of 0.92) (Fig. 4b), but poorly with the melting temperature of variants (Fig. 4c). This observation suggests that at the onset of protein synthesis the DnaK promoter is upregulated purely in response to increased protein translation—independent of the stability of the expressed protein. Already after 30 min the rate of red fluorescence no longer corresponds to the rate of translation due to the increased influence of degradation, and the correlation between the relative rate of red and relative rate of green fluorescence diminishes (Fig. 4h).

**Expression from the DnaK promoter is impacted by protein translation rates and thermostability**. Protein folding stress triggers the PQC network and we wished to investigate whether protein translation alone can trigger a response and to what extent the biophysical properties, such as thermostability, of a protein, impact the stress response.

Assuming that the degradation of GFP is independent of the variant expressed, the total green fluorescence signal in Fig. 4a and Fig. S6 reports on the total synthesis of GFP initiated at the stress-controlled DnaK promoter. Therefore, it can serve as a proxy for the accumulated DnaK response triggered by the expression of the eight variants. Immediately upon induction with IPTG an increase in green fluorescence is observed (Fig. 4a, Fig. S6) which shows that the sfGFP signal emerges even before protein synthesis can be detected as an increase in red fluorescence. However, this can be explained by a shorter maturation time for sfGFP than tagRFP (14 min vs 42 min[39]). Due to the much slower maturation time of tagRFP compared to sfGFP, total fluorescent intensities are not synchronized like the molecular events in the cell. However, by monitoring the rate of

**Variants with low stability show increased DnaK transcription**. In contrast to what is observed at early rates, there is a moderate correlation between the rate of GFP and melting temperature following the first 60 min of protein synthesis. The correlation appears to increase at later timepoints (Fig. 4d, g). While individual correlation values fluctuate and are associated with statistical uncertainty, there is a clear indication that the DnaK response is affected by the thermal stability of the variants (Fig. 4g). Notably, after 120 min, cells expressing the most stable variants (wt, Q33Y and V36I) experience a significant drop in green fluorescence (Fig. 4a, d). This could suggest that the stable variants do not engage DnaK, likely due to a lower concentration of unfolded/misfolded proteins. However, concurrently with a drop off in green fluorescence cells expressing these three variants also experience a decrease in OD. Consequently, the decrease in expression from the DnaK promoter could be a trait associated with cell death. The green fluorescence in cells expressing the remaining five variants continues to increase, although with a reduced rate (Fig. 4a, d) and with a higher rate of green fluorescence for the less stable variants. A correlation between the rate
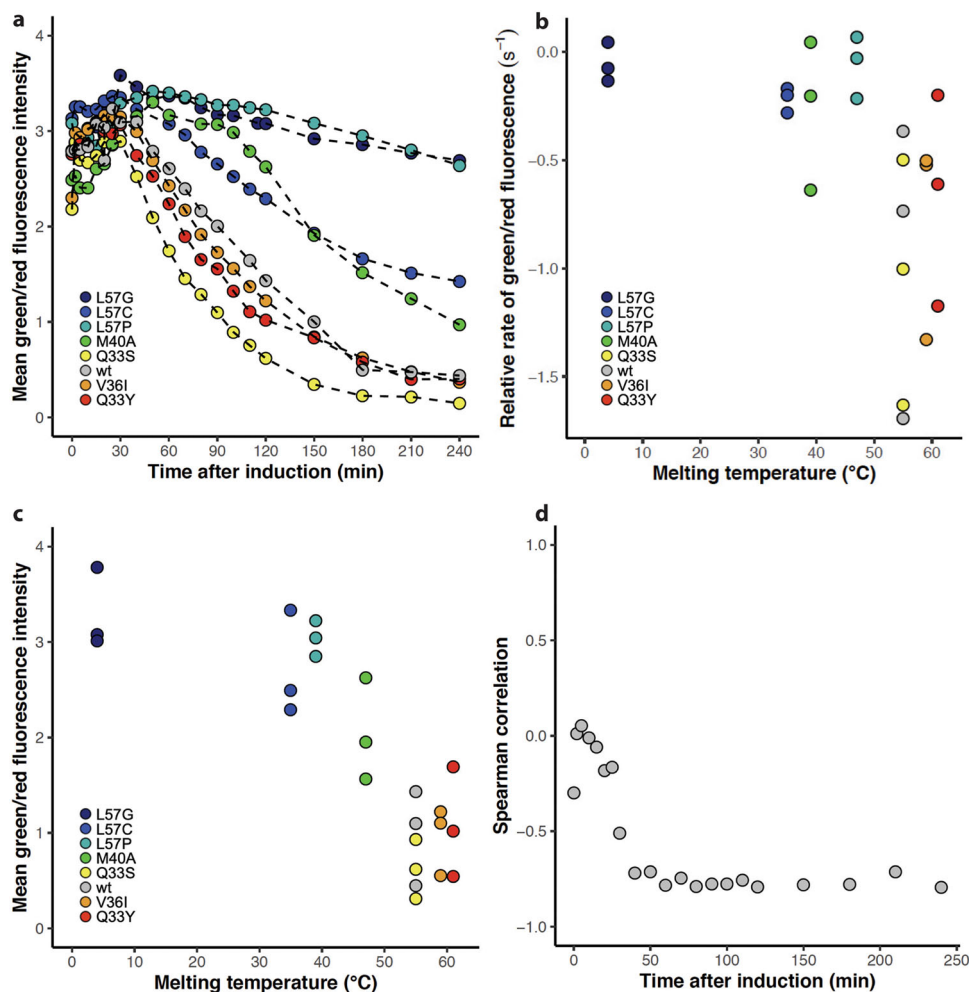
**Fig. 5 Accumulated transcription from the DnaK promoter normalised to protein abundance. a** Mean green fluorescence intensity divided by red fluorescence intensity of cells expressing variants: L57G (dark blue), L57C (blue), L57P (cyan), M40A (green), Q33S (yellow), V36I (orange), Q33Y (red) and wt (grey). One of three replicates is shown for each variant. **b** Relative rate of green/red fluorescence (30–60 min) as a function of melting temperature. Spearman correlation = −0.63). Colour scheme as in (**a**). **c** Mean green fluorescence intensity divided by red fluorescence intensity at the 120 min time point for the eight N102LT variants with different melting temperatures. Spearman correlation between green/red fluorescence intensity and variant melting temperature is −0.79. Colour scheme as in (**a**). **d** Spearman correlation between green/red fluorescence intensity and variant melting temperature at each time point. Calculations are based on three replicates of each variant.

of green fluorescence and melting temperature is observed even when considering only the subset of five with positive GFP rates, but only after 150 min (spearman rank correlation −0.57 for rates calculated in the range 150–180 min and −0.59 for the 180–210 interval). So, as could be expected, our assay indicates an increased requirement for chaperone assistance for the less stable variants.

In the first 50 min post-induction there is a moderate correlation between melting temperature and accumulated green fluorescence. However, onwards from 60 min the correlation decreases steadily (Fig. 4c, f). This most likely reflects the variation in protein abundance as the effect of degradation of the N102LT variants becomes increasingly pronounced. To study how much the expression and stress response varies between individual cells, we must consider differences in plasmid copy number in cells. Since both RFP and GFP are expressed from the same plasmid we can account for this by evaluating the ratio between GFP and RFP signal. This ratio also provides a metric of the chaperone response relative to protein abundance. For all variants, we see an increase in green to red fluorescence ratio up until around 30 min after induction (Fig. 5a, Fig. S7). After that

point, the normalised DnaK response decreases in a manner highly dependent on the expressed variant (Fig. 5a, b). In addition, the accumulated and normalised DnaK response correlates strongly with the melting temperature of the variant following the initial peak at 30 min and throughout the entire 240 min (Fig. 5d). This is exemplified by the 120-min time point, where we see a clear distinction between the chaperone response triggered for the eight N102LT variants, with a strong negative correlation between the green to red fluorescence ratio and the melting temperature of the expressed variants (Fig. 5c). Less stable variants induce a stronger DnaK response, and that pattern persists over time. This observation is a strong indication that the less stable variants require continuous assistance from PQC chaperones to maintain their native fold, whereas more stable variants do not.

**Expression of the wt, Q33Y and V36I variants leads to the occurrence of subpopulations with altered fluorescence phenotypes.** The above analysis shows that evaluation of protein abundance, as well as DnaK response, gives a good insight into
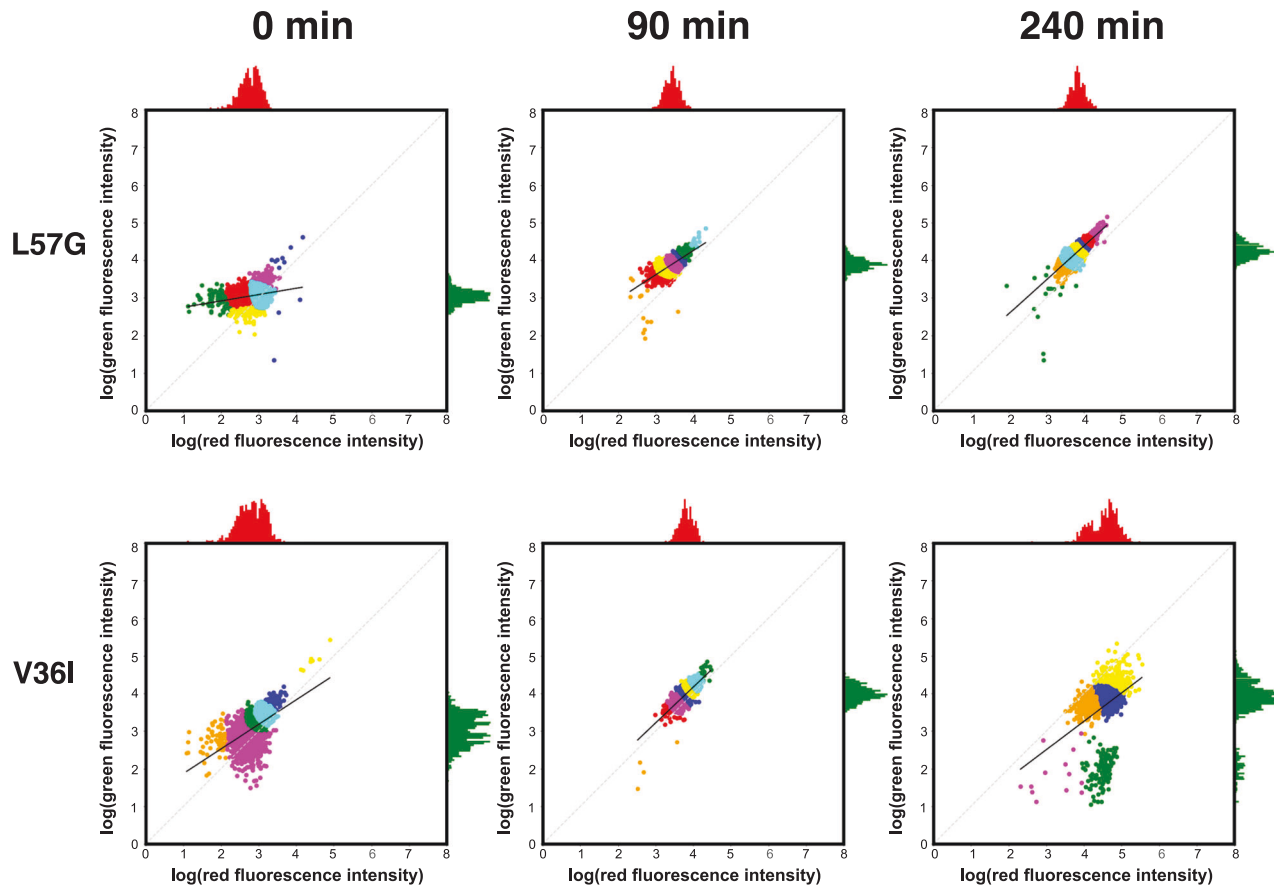
**Fig. 6 Red and green fluorescence of single cells expressing L57G and V36I.** The red and green fluorescence for each individual cell in a population expressing L57G (top) and V36I (bottom) at the time of induction (left) and 90 min (middle) and 240 min (right) after induction with IPTG. Colouring in scatter plot is based on a Bayesian Gaussian mixture model for cluster assignment. The x-axis histogram (red) shows the distribution of red fluorescence, and the y-axis histogram (green) shows the distribution of green fluorescence within the population.

the effect of protein biophysical properties on whole populations of cells expressing the same variant. However, with flow cytometry, we can go beyond the study of cells at the population level and investigate the cell-to-cell variation in protein expression and DnaK response. To fully explore the phenotypic variation within a population, we cluster the cells based on red and green fluorescence using a Bayesian Gaussian mixture model. Although not all identified clusters may represent homogenous phenotypes, automated clustering allows for unbiased assignments. Figure 6 shows the clusters identified for variant L57G and V36I at times 0, 90, and 240 min in a double logarithmic plot. If the cells form a straight line with a slope of one in the double logarithmic plot the green to red fluorescence ratio is constant in all cells and is equivalent to the intersection with the y-axis. A slope below one indicates that the variation in red fluorescence is larger than the variation in green fluorescence, whereas a slope above one indicates that the variation in green fluorescence is bigger than the variation in red fluorescence.

For all variants, we identify one main population consisting of 3–4 different clusters (Fig. 6, Figs. S8, S9). The distribution of red and green fluorescence within the main population differs between variants and with time. At the time of induction, both green, and especially red fluorescence, show significant cell-to-cell variation for all variants. This is evident by the dispersed distribution around a line with a slope below one and is observed regardless of variant. However, for some variants, there is an indication of a diagonal line pre-induction. This is likely due to leaky expression. Once recombinant protein synthesis is induced,

the distribution of red and green fluorescence in the main population becomes increasingly homogenous and approaches a near-constant ratio between green and red fluorescence for all cells in a population (Figs. 6, 7, Figs. S8, S9). Interestingly, when expressing variants with a melting temperature below the wt, reaching the near-constant ratio is significantly delayed (Fig. 7, Fig. S10).

For the least stable variants (L57G, L57C, L57P and M40A) and Q33S, with stability close to wt, there appears to be only a single population (Fig. 6, Figs. S8, S9). For wt and the two most stable variants (Q33Y and V36I), however, individual cells with less protein abundance (lower red fluorescence intensity) emerge after 120 min, and the fraction of these cells increase over time. After 240 min, distinct populations with phenotypic variation in DnaK response can be observed for cells expressing wt, Q33Y and V36I. Cells with low green fluorescence have either low or high red fluorescence, forming two distinct populations. We note that the emergence of subpopulations with altered fluorescence phenotypes is highly variant-specific and is only observed for cells expressing the three most stable variants and concur with a reduction in OD for cells expressing these variants. Neither formation of subpopulations nor reduction in OD is observed for cells expressing Q33S which has thermostability equivalent to the wt.

**Morphological changes are observed for _E. coli_ expressing N102LT variants.** Stressed cells often show morphological changes. We monitored cell morphology by forward (FSC) and

side scatter (SSC) to identify potential variant-dependent changes. FSC provides information about the size of the cell, and for most variants, there are no changes in FSC over the four-hour time course indicating that the size of the cells remains constant (Fig. 8a, Fig. S10). However, for three variants (wt, Q33Y and V36I) the FSC increases after 120 min, indicating an increase in cell size. Furthermore, in two of three replicates these three variants are distinctly different from the other five variants throughout the timeseries, indicating that cells expressing wt, Q33Y and V36I might be bigger even prior to induction. However, this pattern is not observed in the third replicate. For

eukaryotic cells, SSC reports on the complexity of the cell but it remains unclear how to interpret SSC from bacterial cells, although it has been suggested that SSC is associated with cell wall density[42]. Increasing side-scattering width is also associated with the formation of filamentous phenotype [Burke]. We observe a variant-dependent decrease in SSC during the first two hours after induction (Fig. 8b, Fig. S11) with a moderate negative correlation to the melting temperature of the expressed variant. Although the implications of SSC are not understood it is evident from the patterns of FSC and SSC variation that cells undergo morphological changes, and that the changes depend on the expressed variant.
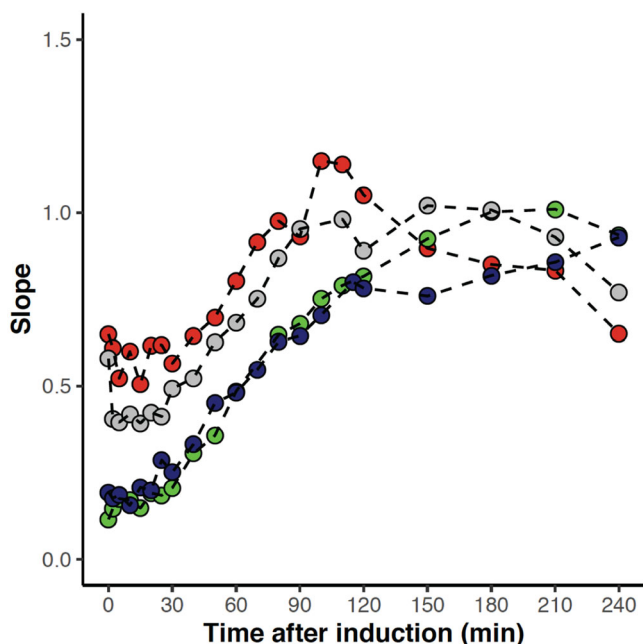


**Fig. 7 Slope of fluorescent population over time.** For each time point, the slope of log-transformed red fluorescence vs log-transformed green fluorescence is calculated for L57G (dark blue) M40A (green), Q33Y (red) and wt (grey). One of three replicates are shown for each variant.

## Discussion

We have demonstrated that DnaK expression is activated immediately once recombinant protein expression starts. It then shifts from being dependent on protein translation rates to being primarily governed by protein stability at later time points. The initial, stability-independent activation is most likely due to the increase of unfolded protein as the nascent peptide chain exits the ribosome. The observed correlation between the onset of protein synthesis and initial DnaK response can be caused by either co-translational binding of DnaK to the emerging peptide chain or binding of DnaK to newly synthesised, unfolded, protein after it is released from the ribosome (Fig. 9). However, the presence of significant amounts of unfolded λ-repressor is unlikely as several λ-repressor constructs have been shown to fold in submillisecond time scales[43–45]. The binding of DnaK to the unfolded protein releases DnaK-bound $\sigma^{32}$ which acts as a transcription factor for the synthesis of DnaK as well as other PQC proteins. The consequence is an increase in transcription of PQC proteins that is unrelated to protein thermostability but tightly linked to translation rate. Besides aiding in nascent chain folding, this linkage adjusts the amount of PQC network components to the overall protein expression level and thus ensures that enough chaperones are present to deal with potential protein misfolding. The rate of GFP synthesis as well as accumulated GFP/RFP signal correlates with variant melting temperature after the initial 30 min of protein synthesis. This correlation suggests that the DnaK response is impacted by the proportion of protein in the unfolded/misfolded state at folding equilibrium (Fig. 9). Hence, variants with lower
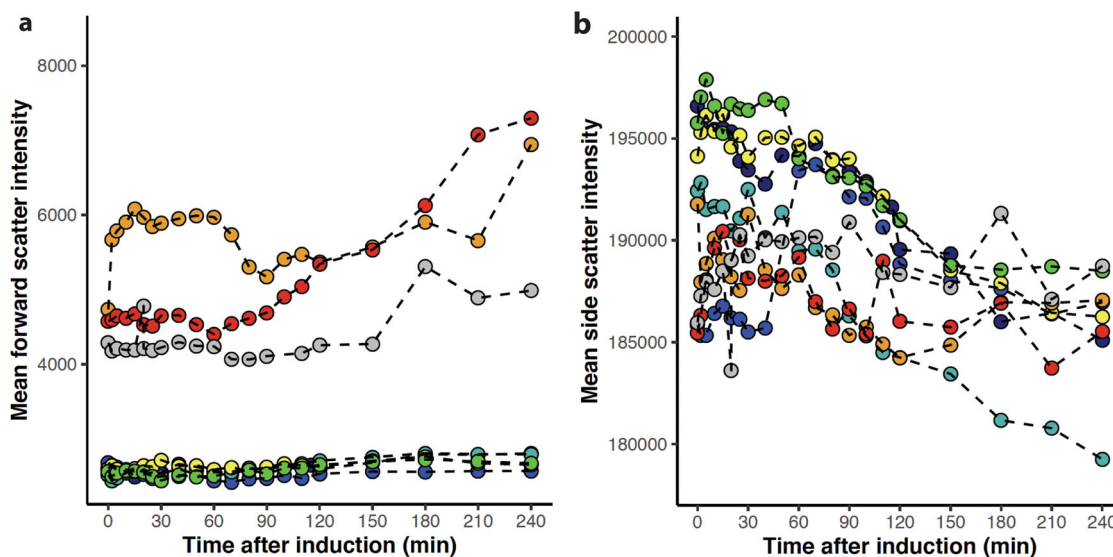


**Fig. 8 Cell morphology of whole populations expressing N102LT variants.** Forward (**a**) and side (**b**) scatter as a function of time for variants: L57G (dark blue), L57C (blue), L57P (cyan), M40A (green), Q33S (yellow), V36I (orange), Q33Y (red) and wt (grey). One of three replicates are shown for each variant.
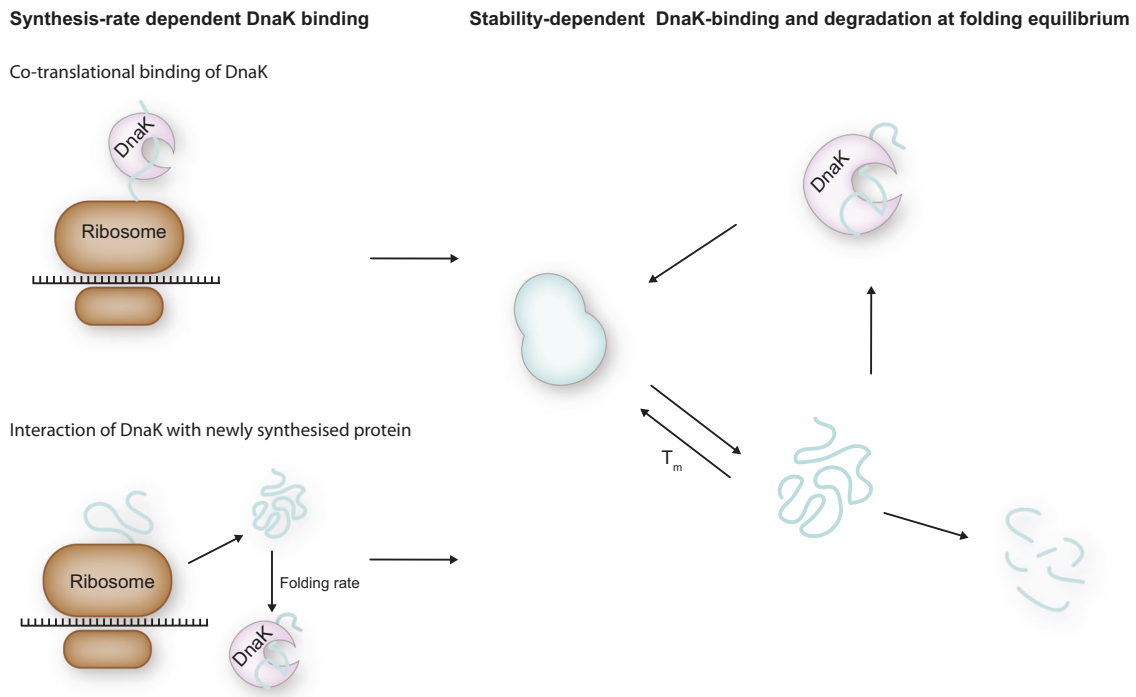
**Synthesis-rate dependent DnaK binding**          **Stability-dependent DnaK-binding and degradation at folding equilibrium**

Co-translational binding of DnaK

Interaction of DnaK with newly synthesised protein



**Fig. 9 A model for DnaK interaction with client proteins.** Two different factors impact the interaction of DnaK with client proteins: synthesis rate and protein stability. DnaK can act co-translationally with the nascent peptide chain as it exits the ribosome exit tunnel, in which case the DnaK response is expected to be proportional to the translational rate. In addition, DnaK can bind to unfolded newly synthesised protein that has been released from the ribosome. For slow-folding protein, the interaction with DnaK will then also depend on folding rate. Fast folding protein will interact only minimally with DnaK. Fully synthesised proteins are in an equilibrium between folded and misfolded determined by the stability of the protein. Thus, DnaK binding to misfolded protein at folding equilibrium is dependent on protein stability.

stability require increased DnaK activity compared to more stable variants. Increased interaction with DnaK increases the degree of degradation by proteases, as DnaK, given its role as a central hub in the PQC network, is involved in targeting misfolded proteins for degradation[46]. As a consequence, in our study, most protein is accumulated in cells expressing the four variants that are at least as stable as the wt (wt, Q33S, Q33Y and V36I), and thus do not require continuous assistance from chaperones.

Analysis of the accumulated DnaK response and N102LT variant accumulation in single cells revealed that cell–cell variation, especially in protein accumulation, is bigger early after the onset of protein synthesis and decreases with time. There is a prolonged period with a wide distribution in protein accumulation observed for variants less stable than wt. This lag can potentially be explained by increased degradation: The variation in protein abundance can be caused by a variance in both protein synthesis and degradation. The variation for the more stable variants, however, is primarily due to variation in synthesis. Once folding and synthesis/degradation equilibrium are reached, a single population of cells is seen, with a similar chaperone response relative to protein abundance. This suggests that, as the cells reach some sort of homeostasis, the variation in DnaK response can be largely explained by variation in accumulated protein.

However, for the wt, Q33Y and V36I variants subpopulations with reduced DnaK response emerge after ~120 min. In association with the emergence of subpopulations in these three stable variants, there is a decrease in OD$_{600}$ and DnaK as well as an increase in cell size (a phenotype often associated with stressed cells[47]) is observed. The transformation into these alternative phenotypes is potentially related to the burden associated with a high protein load in the cell. Despite that Q33S reaches an

abundance of protein levels similar to and above the abundance of the wt, Q33Y and V36I, this variant does not share the described phenotype indicating a more complex relationship between protein stability and cellular response.

Although we find a good overall correlation between the variant melting temperature and the DnaK response after expression in *E. coli*, not all the variation can be explained by the altered thermostability. Consequently, there might be other sequence-dependent effects that we have not investigated further. The folding equilibrium might be shifted in vivo compared to the conditions in vitro. The unfolded state, for instance, could be stabilized by inadvertent interactions with other macromolecules. Similarly, the folded state could be further stabilised by non-specific interactions, e.g., with DNA. Altered affinity for DNA could also affect a variant's toxicity and potentially explain the altered phenotypes observed exclusively for the wt, Q33Y and V36I which cannot be explained by thermostability alone. The DnaK-dependent signal is further potentially skewed by a variant's aggregation propensity as aggregation itself has an impact on PQC network activity[15].

Several studies have investigated the coupling between biophysical properties and protein evolution. A strong negative correlation between rates of protein-coding sequence evolution and gene expression levels have been observed[12,48]. This has been coupled to protein stability by two different models. In the first, unstable proteins have a lower concentration of active protein resulting in lower fitness. In the second, the misfolded proteins are toxic to the cell with an associated fitness cost. Mathematically, these models become identical for proteins with stabilities beyond 5 kcal/mol[49]. Geiler-Samerotte et al.[15] directly tested the misfolding hypothesis by evaluating growth rates in yeast expressing misfolded proteins. They found a significant fitness

cost due to misfolded proteins and associated upregulation of HSP70 (DnaK). The increased evolutionary rate of proteins with stronger binding to DnaK suggests that chaperones can buffer deleterious mutations[11]. Nonetheless, triggering the PQC network comes with a cost and it is possible that this burden on the cell can impact the selection of protein variants during evolution. The observation that small amino acid changes can have a significant impact on the chaperone response thus suggests that the degree of *dnaK*-activation by a protein may itself act as a significant selection factor. The exact relation between the magnitude of the PQC and fitness, however, is not clear. In our experiments, we observe a correlation between the maximum growth rate and protein stability. However, more accurate and detailed studies of the relative growth of protein variants need to be carried out to link fitness and protein stability and to prove that differential growth is caused by the cost of the quality control system. It should be noted that we carry out our experiments under conditions of strong protein overexpression. Hence, it is not clear if our results are relevant to the evolution of protein sequences under more standard conditions.

Stability is often a prerequisite for the application of proteins in research, medicine and biotechnology. The development of accurate and high-throughput assays that can screen for stability independent of protein function is thus desired. Our data demonstrate strong correlations between in vitro protein stability and fluorescence values recorded from single cells expressing the N102LT variants. This highlights the utility of the fluorescent reporter plasmid in combination with flow cytometry analysis as an assay for screening for protein stability in vivo. We propose two potential approaches for using the reporter assay for screening: (i) using the GFP/RFP ratios of tagRFP fusion proteins as a high throughput screen for variant libraries, e.g., for directed evolution. The strong correlation between GFP/RFP ratio and thermostability provides optimal conditions for screening for variants with increased stability since the correlation varies minimally with time. (ii) Using the rate of GFP for high-throughput stability screening of protein variants. This approach is potentially less robust than the GFP/RFP ratio method but has the advantage of not requiring the proteins of interest to be fused to tagRFP. This can for example be beneficial in cases where the studied protein is involved in multimerization reactions where the large tagRFP molecule can inhibit proper assembly.

## Methods

**Strains and plasmids**. A modified pQlinkN plasmid[50] was created by inserting additional restriction sites for PstI, SacII and NdeI between the Link1-region and the multiple cloning site (MCS) of pQlinkN. The genes encoding sfGFP and the dnaK-promoter ($P_{DnaK}$) were amplified by PCR from *E.coli* BL21(De3)pLysS genomic DNA, respectively. Both *sfgfp* and $P_{dnak}$ were inserted into the modified pQlinkN plasmid using restriction enzyme cloning.

The gene for N102LT N-terminally fused with tagRFP (gi: 336287738) was synthesised (GenScript) and inserted into an unmodified pQlinkN plasmid using restriction enzyme cloning. Point mutations in the N102LT gene were made by site-directed mutagenesis with Phusion HotStart II DNA polymerase to create seven N102LT variants: Q33S, Q33Y, V36I, M40A, L57C, L57G and L57P. The full reporter plasmid was created by ligation independent cloning with pQLinkN harbouring *N102LT-tagRFP* variants as recipient vector and modified pQlinkN harbouring $P_{DnaK}$ and *sf-gfp* as insert vector[50].

**Protein expression**. *E. coli* BL21(DE3) pLysS was transformed with reporter plasmid harbouring the N102LT variants for expression. Cells were grown at 37 °C in LB medium with shaking (150 rpm) and were induced with 1 mM isopropyl-β-d-thiogalactopyranoside (IPTG) when cells reached an $OD_{600}$ between 0.4 and 0.6 and continued to grow for four hours. Variants were expressed on separate days, but two replicates of each variant were expressed individually on the same day.

**Flow cytometry**. Culture samples were taken just before induction with IPTG (time point 0 min), after 2 min, every five min the first 30 min, every 10 min from 30 min to 120 min and every half hour until 240 min after induction. Immediately after taking the sample it was diluted 1:1000 in 10 mM $Na_2HPO_4$, 1.8 mM $KH_2PO_4$,

137 mM NaCl, 2.7 mM KCl pH = 7.4 and analysed on s3e cell sorter (Biorad). 3000 events were collected with an event rate of ~500 events/s. A laser with 488 nm was used for detection. Cells expressing N102LT variants were gated on forward and side scatters to exclude doublets. Fluorescence was detected using a 525/30 filter (green fluorescence) and a 615/25 filter (red fluorescence). Flow cytometry data were analysed using the flowCore[51], flowViz[52] and flowWorkspace[53] packages within the Bioconductor frame work.

**Western blot analysis**. Cell samples were harvested by centrifugation for 1 min at $6000 \times g$ and lysed with 1 x bugbuster (Novagen) according to the manufacturer's recommendation. The cell lysate was separated on 4–20% mini protean tgx precast protein gels (Biorad) and transferred to a nitrocellulose membrane. Primary antibody against *E. coli* DnaK raised in mouse was purchased from Enzo life sciences and used with the WesternBreeze™ Chromogenic anti-mouse kit. A 1:10,000 dilution was used for the anti-bodies.

**Aggregation study**. Cell samples were collected 30 and 60 min after induction of protein synthesis and lysed with 1 x bugbuster (Novagen) according to the manufacturer's recommendation. The soluble and insoluble fraction was separated by centrifugation at $16,000 \times g$ for 10 min. The soluble and insoluble fraction was analysed on 4–20% mini protean tgx precast protein gels (Biorad).

**Clustering**. Flow cytometry data were clustered using the Bayesian Gaussian mixture model, implemented in scikit-learn[54]. Up to seven different components were allowed and priors for weights were described by a Dirichlet distribution with a weight concentration prior to 1e−6. A diagonal covariance matrix was used for each component.

**Statistics and reproducibility**. Mean red and green fluorescence as well as mean forward and side scatter intensities were calculated from the absolute intensity for each event for a given variant at a given time. Mean green/red fluorescence intensity was calculated by taking the mean over the green/red fluorescence intensity for each variant for each timepoint. Rates were estimated by fitting a linear trendline to the mean red, green or green/red fluorescence intensity within the indicated time interval. Relative rates were estimated by normalizing the rate of fluorescence with the mean fluorescence within the same time interval. Mean red, green and green/red fluorescence, as well as relative rates, were calculated individually for each of the three replicates. Spearman correlations between red, green or green/red and variant melting temperature were calculated based on three replicates of each variant. A linear trend line was fitted to the log10 transformed red and green fluorescence for each event for a given variant at any given time point. A trend line was fitted individually for each replicate.

**Growth rate analysis**. The individual growth curves were fit to a logistic growth function using the curveball python package[55], which uses the lmfit python package {Newville, #68} for non-linear curve fitting to identify the maximum likelihood solution. The lmfit package was then used together with the emcee python package to carry out Markov Chain Monte Carlo with uniform priors in order to calculate posterior probabilities of parameters. Samples from the Monte Carlo chain were then used to evaluate growth rate curves, from which maximum growth rates were numerically determined. The variation in maximum growth rates were summarized in a standard deviation plotted in Fig. 2b.

**Reporting summary**. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Data associated with this manuscript have been archived in GitHub (https://github.com/Andre-lab/DnaK_paper). All additional relevant data are available by request from the authors.

## References

1. Rudiger, S., Buchberger, A. & Bukau, B. Interaction of Hsp70 chaperones with substrates. *Nat. Struct. Biol.* **4**, 342–349 (1997).
2. Guisbert, E., Herman, C., Lu, C. Z. & Gross, C. A. A chaperone network controls the heat shock response in E. coli. *Genes Dev.* **18**, 2812–2821 (2004).
3. Blaszczak, A., Georgopoulos, C. & Liberek, K. On the mechanism of FtsH-dependent degradation of the sigma 32 transcriptional regulator of *Escherichia*

*coli* and the role of the Dnak chaperone machine. *Mol. Microbiol.* **31**, 157–166 (1999).

4. Jurgen, B. et al. Quality control of inclusion bodies in *Escherichia coli*. *Micro. Cell Fact.* **9**, 41 (2010).

5. Gonzalez-Montalban, N., Carrio, M. M., Cuatrecasas, S., Aris, A. & Villaverde, A. Bacterial inclusion bodies are cytotoxic in vivo in absence of functional chaperones DnaK or GroEL. *J. Biotechnol.* **118**, 406–412 (2005).

6. Plata, G., Gottesman, M. E. & Vitkup, D. The rate of the molecular clock and the cost of gratuitous protein synthesis. *Genome Biol.* **11**, R98 (2010).

7. Sekhar, A., Lam, H. N. & Cavagnero, S. Protein folding rates and thermodynamic stability are key determinants for interaction with the Hsp70 chaperone system. *Protein Sci.* **21**, 1489–1502 (2012).

8. Calloni, G. et al. DnaK functions as a central hub in the *E. coli* chaperone network. *Cell Rep.* **1**, 251–264 (2012).

9. Hoffmann, F. & Rinas, U. Kinetics of heat-shock response and inclusion body formation during temperature-induced production of basic fibroblast growth factor in high-cell-density cultures of recombinant *Escherichia coli*. *Biotechnol. Prog.* **16**, 1000–1007 (2000).

10. Thomas, J. G. & Baneyx, F. Protein misfolding and inclusion body formation in recombinant *Escherichia coli* cells overexpressing Heat-shock proteins. *J. Biol. Chem.* **271**, 11141–11147 (1996).

11. Kadibalban, A. S., Bogumil, D., Landan, G. & Dagan, T. DnaK-dependent accelerated evolutionary rate in prokaryotes. *Genome Biol. Evol.* **8**, 1590–1599 (2016).

12. Drummond, D. A., Raval, A. & Wilke, C. O. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23**, 327–337 (2006).

13. Drummond, D. A. & Wilke, C. O. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352 (2008).

14. Dong, H., Nilsson, L. & Kurland, C. G. Gratuitous overexpression of genes in *Escherichia coli* leads to growth inhibition and ribosome destruction. *J. Bacteriol.* **177**, 1497–1504 (1995).

15. Geiler-Samerotte, K. A. et al. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc. Natl Acad. Sci. USA* **108**, 680–685 (2011).

16. Rutherford, S. L. & Lindquist, S. Hsp90 as a capacitor for morphological evolution. *Nature* **396**, 336–342 (1998).

17. Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. How protein stability and new functions trade off. *PLoS Comput. Biol.* **4**, e1000002 (2008).

18. Bogumil, D., Landan, G., Ilhan, J. & Dagan, T. Chaperones divide yeast proteins into classes of expression level and evolutionary rate. *Genome Biol. Evol.* **4**, 618–625 (2012).

19. Wyganowski, K. T., Kaltenbach, M. & Tokuriki, N. GroEL/ES buffering and compensatory mutations promote protein evolution by stabilizing folding intermediates. *J. Mol. Biol.* **425**, 3403–3414 (2013).

20. Agozzino, L. & Dill, K. A. Protein evolution speed depends on its stability and abundance and on chaperone concentrations. *Proc. Natl Acad. Sci. USA* **115**, 9092–9097 (2018).

21. Williams, T. A. & Fares, M. A. The effect of chaperonin buffering on protein evolution. *Genome Biol. Evol.* **2**, 609–619 (2010).

22. Lachowiec, J., Lemus, T., Borenstein, E. & Queitsch, C. Hsp90 promotes kinase evolution. *Mol. Biol. Evol.* **32**, 91–99 (2015).

23. Moreau, M. J., Morin, I. & Schaeffer, P. M. Quantitative determination of protein stability and ligand binding using a green fluorescent protein reporter system. *Mol. Biosyst.* **6**, 1285–1292 (2010).

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

25. Matreyek, K. A. et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).

26. Arsene, F., Tomoyasu, T. & Bukau, B. The heat shock response of *Escherichia coli*. *Int J. Food Microbiol.* **55**, 3–9 (2000).

27. Cluss, R. G., Goel, A. S., Rehm, H. L., Schoenecker, J. G. & Boothby, J. T. Coordinate synthesis and turnover of heat shock proteins in Borrelia burgdorferi: degradation of DnaK during recovery from heat shock. *Infect. Immun.* **64**, 1736–1743 (1996).

28. Mitchell, H. K., Petersen, N. S. & Buzin, C. H. Self-degradation of heat shock proteins. *Proc. Natl Acad. Sci. USA* **82**, 4969–4973 (1985).

29. Chung, C. H. & Goldberg, A. L. The product of the lon (capR) gene in *Escherichia coli* is the ATP-dependent protease, protease La. *Proc. Natl Acad. Sci. USA* **78**, 4931–4935 (1981).

30. Mahmoud, S. A. & Chien, P. Regulated proteolysis in bacteria. *Annu Rev. Biochem.* **87**, 677–696 (2018).

31. Parsell, D. A. & Sauer, R. T. The structural stability of a protein is an important determinant of its proteolytic susceptibility in *Escherichia coli*. *J. Biol. Chem.* **264**, 7590–7595 (1989).

32. Lim, W. A. & Sauer, R. T. The role of internal packing interactions in determining the structure and stability of a protein. *J. Mol. Biol.* **219**, 359–376 (1991).

33. Parsell, D. A. & Sauer, R. T. Induction of a heat shock-like response by unfolded protein in *Escherichia coli*: dependence on protein level not protein degradation. *Genes Dev.* **3**, 1226–1232 (1989).

34. Lim, W. A. & Sauer, R. T. Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature* **339**, 31–36 (1989).

35. Lim, W. A., Hodel, A., Sauer, R. T. & Richards, F. M. The crystal structure of a mutant protein with altered but improved hydrophobic core packing. *Proc. Natl Acad. Sci. USA* **91**, 423–427 (1994).

36. Hecht, M. H., Sturtevant, J. M. & Sauer, R. T. Effect of single amino acid replacements on the thermal stability of the NH2-terminal domain of phage lambda repressor. *Proc. Natl Acad. Sci. USA* **81**, 5685–5689 (1984).

37. Gutierres, M. B. B., Bonorino, C. B. C. & Rigo, M. M. ChaperISM: improved chaperone binding prediction using position-independent scoring matrices. *Bioinformatics* **36**, 735–741 (2020).

38. Iafolla, M. A. et al. Dark proteins: effect of inclusion body formation on quantification of protein expression. *Proteins* **72**, 1233–1242 (2008).

39. Balleza, E., Kim, J. M. & Cluzel, P. Systematic characterization of maturation time of fluorescent proteins in living cells. *Nat. Methods* **15**, 47–51 (2018).

40. Hintsche, M. & Klumpp, S. Dilution and the theoretical description of growth-rate dependent gene expression. *J. Biol. Eng.* **7**, 22 (2013).

41. Nordholt, N., van Heerden, J., Kort, R. & Bruggeman, F. J. Effects of growth rate and promoter activity on single-cell protein expression. *Sci. Rep.* **7**, 6299 (2017).

42. Diaz, M., Herrero, M., Garcia, L. A. & Quiros, C. Application of flow cytometry to industrial microbial bioprocesses. *Biochem. Eng. J.* **48**, 385–407 (2010).

43. Huang, G. S. & Oas, T. G. Submillisecond folding of monomeric lambda repressor. *Proc. Natl Acad. Sci. USA* **92**, 6878–6882 (1995).

44. Yang, W. Y. & Gruebele, M. Folding lambda-repressor at its speed limit. *Biophys. J.* **87**, 596–608 (2004).

45. Burton, R. E., Huang, G. S., Daugherty, M. A., Calderone, T. L. & Oas, T. G. The energy landscape of a fast-folding protein mapped by Ala—>Gly substitutions. *Nat. Struct. Biol.* **4**, 305–310 (1997).

46. Sherman, M. & Goldberg, A. L. Involvement of the chaperonin dnaK in the rapid degradation of a mutant protein in *Escherichia coli*. *EMBO J.* **11**, 71–77 (1992).

47. Jeong, K. J. & Lee, S. Y. Enhanced production of recombinant proteins in *Escherichia coli* by filamentation suppression. *Appl. Environ. Microbiol.* **69**, 1295–1298 (2003).

48. Pal, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927–931 (2001).

49. Norn, C., Andre, I. & Theobald, D. L. A thermodynamic model of protein structure evolution explains empirical amino acid substitution matrices. *Protein Sci.* https://doi.org/10.1002/pro.4155 (2021).

50. Scheich, C., Kummel, D., Soumailakakis, D., Heinemann, U. & Bussow, K. Vectors for co-expression of an unrestricted number of proteins. *Nucleic Acids Res.* **35**, e43, https://doi.org/10.1093/nar/gkm067 (2007).

51. flowCore: flowCore: Basic structures for flow cytometry data. R package version 1.38.2 (2016).

52. flowViz: Visualization for flow cytometry. R package version 1.56.0 (2021).

53. flowWorkspace: Infrastructure for representing and interacting with gated and ungated cytometry data sets. R package version 4.4.0 (2021).

54. Pedregosa, F. V. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

55. Ram, Y. et al. Predicting microbial growth in a mixed culture from growth curve data. *Proc. Natl Acad. Sci. USA* **116**, 14698–14707 (2019).

56. Beamer, L. J. & Pabo, C. O. Refined 1.8 A crystal structure of the lambda repressor-operator complex. *J. Mol. Biol.* **227**, 177–196 (1992).

## Author contributions

S.R.: Methodology, writing—review and editing. S.C.: Conceptualization, Methodology, formal analysis, investigation, writing—original draft, visualization. I.A.: Conceptualization, methodology, software, resources, supervision, writing the original draft.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s42003-022-03542-2.

**Correspondence** and requests for materials should be addressed to Signe Christensen or Ingemar André.

**Peer review information** *Communications Biology* thanks Per Widlund and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Gene Chong.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.