The FEBS Journal

FEBS PRESS science publishing by scientists

Increasing recombinant protein production in *E. coli* via FACS-based selection of N-terminal coding DNA libraries

Štěpán Herynek[1], Jakub Svoboda[1], Maroš Huličiak[1], Yoav Peleg[2], Ľubica Škultétyová[1], Pavel Mikulecký[1] and Bohdan Schneider[1]

1 Institute of Biotechnology, Czech Academy of Sciences, BIOCEV, Prague, Czech Republic
2 Structural Proteomics Unit (SPU), Department of Life Sciences Core Facilities (LSCF), Weizmann Institute of Science, Rehovot, Israel

Keywords
directed evolution; DNA libraries; fluorescence-activated cell sorting (FACS); N-terminal sequences; protein expression optimization

Correspondence
B. Schneider, Institute of Biotechnology, Czech Academy of Sciences, BIOCEV, Průmyslová 595, Prague West, Prague 25250, Czech Republic
Tel: +420 728 303 566
E-mail: bohdan.schneider@ibt.cas.cz

(Received 18 July 2024, revised 28 August 2024, accepted 25 November 2024)

doi:10.1111/febs.17376

Here, we present a previously undescribed approach to modify N-terminal sequences of recombinant proteins to increase their production yield in *Escherichia coli*. Prior research has demonstrated that the nucleotides immediately following the start codon can significantly influence protein expression. However, the impact of these sequences is construct-specific and is not universally applicable to all proteins. Most of the previous research has been limited to selecting from a few rationally designed sequences. In contrast, we used a directed evolution-based methodology, screening large numbers of diversified sequences derived from DNA libraries coding for the N-termini of investigated proteins. To facilitate the identification of cells with increased expression of the target construct, we cloned a GFP gene at the C-terminus of the expressed genes and used fluorescent activated cell sorting (FACS) to separate cells based on their fluorescence. By following this systematic workflow, we successfully elevated the yield of soluble recombinant proteins of multiple constructs up to over 30-fold.

Introduction

Bacteria *Escherichia coli* (*E. coli*) is frequently utilized for producing recombinant proteins. This expression system is preferred due to well-established workflow, simple cultivation and transformation processes, rapid growth rate, and overall cost-effectiveness [1–4]. Nevertheless, recombinant protein production in *E. coli* remains a challenge and requires numerous optimization strategies. Factors such as vector type, promoter, *E. coli* strain choice, inducer concentration, temperature, and media composition significantly influence yield [5].

It has been reported multiple times that the composition of the first few codons following a start codon plays a critical role in translation regulation [6–15]. However, it is not clear whether the control mechanism takes place on the level of nucleic acid sequence-induced structure, codon usage, or folding of nascent peptide.

Translation rate can be influenced by N-terminal amino acids [16] and it has been identified that the crucial molecular component is the second amino acid [6]. The authors have utilized a pET vector using different *E. coli* strains and mutated the second amino acid position of the extracellular domain of Igα protein to all 20 natural amino acid residues. The data indicated that the amino acid at the second position had a 10-fold impact on the yield of recombinant protein expression [6]. The importance of the second amino acid position has been found crucial before,

Abbreviations
CN, natural codons; CO, optimized codons; ERFE, erythroferrone; FACS, fluorescence-activated cell sorting; GFP, green fluorescent protein; GST, glutathione S-transferase; IFNb, interferon beta; IL9, interleukin 9; MCS, multiple cloning site; RF, Restriction-Free (cloning); SOL, soluble fraction; TrxA, thioredoxin A.

though authors attributed the effect to the nucleotide triplet rather than the amino acid [17].

The protein production has been successfully increased by adding a small number of amino acids common to the N termini of highly expressed proteins. The N-terminal sequence MSKIK was designed by this approach and proven to enhance the protein yield although the mechanism underlying this phenomenon has not been identified [9]. It is worth noting that this sequence contained a relatively rare ATA codon (coding for isoleucine) at its N-terminal part. Changing the nucleotide sequence without altering the amino acid sequence did not affect the recombinant protein yield and authors also reported that the mRNA levels coding the protein tagged with the MSKIK sequence were 4–5 times lower than those of equivalent untagged proteins, suggesting that the mRNA level is not the key factor [9]. Six years later the same authors proposed that their MSKIK N-terminal peptide prevents or releases ribosomal stalling resulting in a higher protein yield, while also ruling out the influence of codon usage [18].

The speed of translation initiation is also influenced by the mRNA composition and structure. Structural organization of the ribosome binding site is especially important and important is also beginning of the open reading frame [14,19–22]. Indeed, bioinformatic analysis revealed a trend across various eubacteria wherein the content of mRNA secondary structures decreased towards the 5′ end of analyzed genes [23]. The stability of mRNA structures is usually estimated using nearest-neighbor minimum-free energy models. These calculations performed on a sequence near the Shine-Dalgarno sequence have revealed that a more stable mRNA secondary structure decreases the protein expression [24,25]. Subsequently, a bioinformatics tool called TISIGNER has been developed to minimize the likelihood of mRNA secondary structure formation. It designs optimal initial codons of the gene while keeping the same amino acid sequence [26]. An alternative methodology has been employed modifying the amino acid sequence to achieve the optimal codon utilization [17]. To systematically investigate this, all 64 codons were introduced after the start codon of modified *lacZ* gene, and their impact on recombinant protein yield was analyzed, showing a 15-fold disparity in yield between the lowest and the highest amount of the produced recombinant protein [17]. As expected, the outcome of several studies suggests a connection between translation efficiency and mRNA disorder near the start codon in *E. coli* [20], but also in other organisms including other bacteria, archaea, fungi, plants, insects, and fishes [27] and even in dsDNA of viruses [28].

Codon usage bias is another factor influencing the translation process [29,30]. This phenomenon has also been addressed while researching the N-terminal sequences. Some findings indicate that the point of interest should be the relative frequency of codons. This is consistent with the previous results showing that rare codons such as AGG, the least used codon in *E. coli*, appear more frequently among the first 25 codons [11,31], even though the overall codon usage usually correlates with the abundance of tRNAs [32,33]. The ability of the rare codon placed at the beginning of the gene to increase the yield has been experimentally proven using different proteins such as lacZ and streptokinase, both of which have relatively high amounts of rare codons in their sequences (about 30%). The expression of AGG triplet codon encoding Arginine at various positions within a gene has been evaluated. Placing the codon at the second position resulted in increased expression. However, when it was placed farther along the gene, the yield decreased, and expression was stalled [11,34]. This again emphasizes the importance of the position of the second amino acid in sequences [6]. In contrast, more frequently used Arginine codons at the second and at the third positions lowered the expression [34]. This is in concordance with the previous studies, which show that the placement of rare codons (such as AGG or AGA codons) has a crucial effect on a yield of a produced protein [31,35].

The varying levels of expression yields might be influenced by the number of mRNA copies, suggesting that optimizing protein production may be linked to the level of DNA transcription. However, this is not a universal mechanism. For example, in the case of ovine growth hormone, higher production levels do not correlate with increased mRNA levels compared to the wild-type variant [36]. In another instance, the mRNA levels for an anti-*E. coli* O157 fragment of antigen binding fused with a leucine zipper were reported to be even lower [9].

The above cited and other studies demonstrate that in specific cases the N-terminal sequence of to-be-produced proteins can be optimized for yield improvement. However, they also demonstrate that to formulate the ultimate and generally applicable rules how to optimize the yield or explain why the optimization has been fulfilled is not easy [37]. From the point of the final goal, i.e., increasing the protein production, it is not critical if the increased yield has been achieved by optimized codon composition at the level

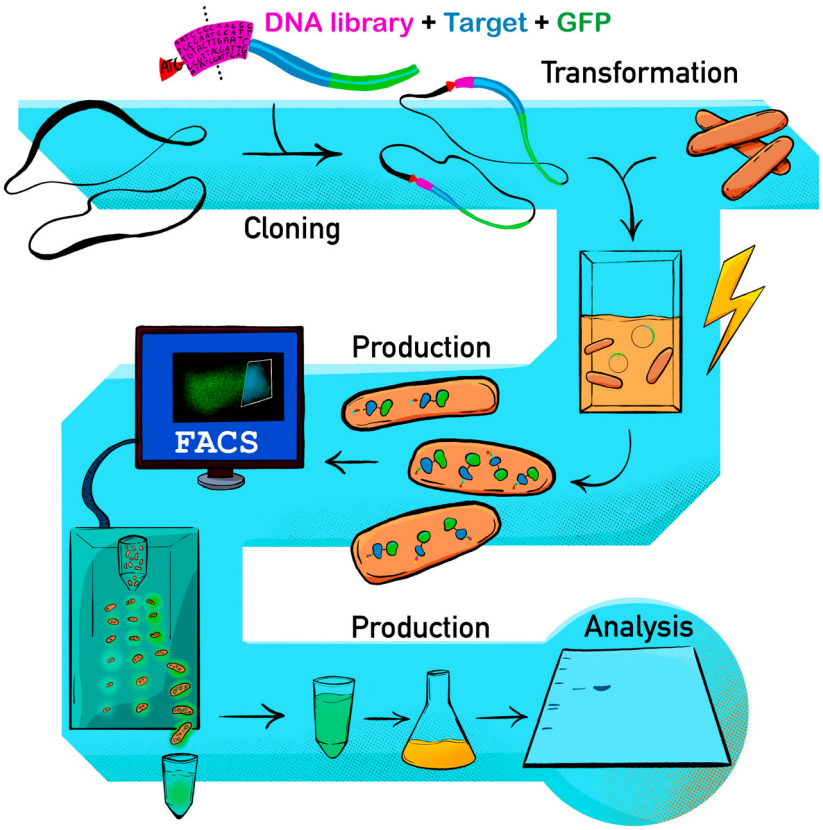


Fig. 1. Schema of the protocol for enhancing protein yields in *E. coli*. The main steps of the protocol are: *Cloning* of the production vector, which includes inserting the target gene (in blue) into the plasmid along with a DNA library coding the randomized N-terminus (in purple) and the GFP gene (in green) fused to the C-terminus of the construct. *Transformation* into electrocompetent cells and protein *production* followed by *FACS selection* of the clones overproducing target protein monitored by higher fluorescence levels. Sorted cells are used for protein *production* with the library-derived sequences. The protein production is *analyzed* by SDS/PAGE.

of the transcription, initial phases of the translation, or behavior of a group of few N-terminal amino acids of the nascent protein. To directly address the issue of protein production we decided to systematically manipulate three to five amino acid residues at the N termini of several proteins. We utilized randomized DNA libraries, which we previously used with success to increase stability and binding of a small protein binders [38,39]. To separate cells showing higher protein production, we used fluorescent activated cell sorting (FACS) and summarized our findings into an easy-to-follow protocol.

Results and discussion

Workflow overview

We present a new method to enhance the production of various recombinant proteins in *E. coli* using genetic engineering techniques. To accomplish this, we constructed DNA libraries that randomly modified the N-terminal sequences of the target proteins. These libraries were fused with the genes encoding the target proteins and a green fluorescent protein (GFP) marker at the C terminus. We used cell sorting to assess the effectiveness of these modified constructs. Based on our successful results, we established a streamlined protocol for this process, as depicted in Fig. 1 and described in detail below.

Experimental design and gating strategy

Selection of the reporter protein molecule and vector for FACS

We used the cycle 3 GFP as a reporter molecule for all our constructs because it includes several mutations resulting in increased fluorescence compared to the wild-type GFP. These mutations also facilitate proper maturation of the protein at 37 °C and the molecule shows lower tendency to aggregate [40,41]. We commonly refer to this molecule as "GFP" in our study.

N-terminal sequence optimization by FACS selection

To verify that changing the N-terminal nascent protein sequence affects the protein yield, we used *E. coli* BL21-Gold (DE3) cells transformed either with pET22b vector without GFP (negative control), pET22b with

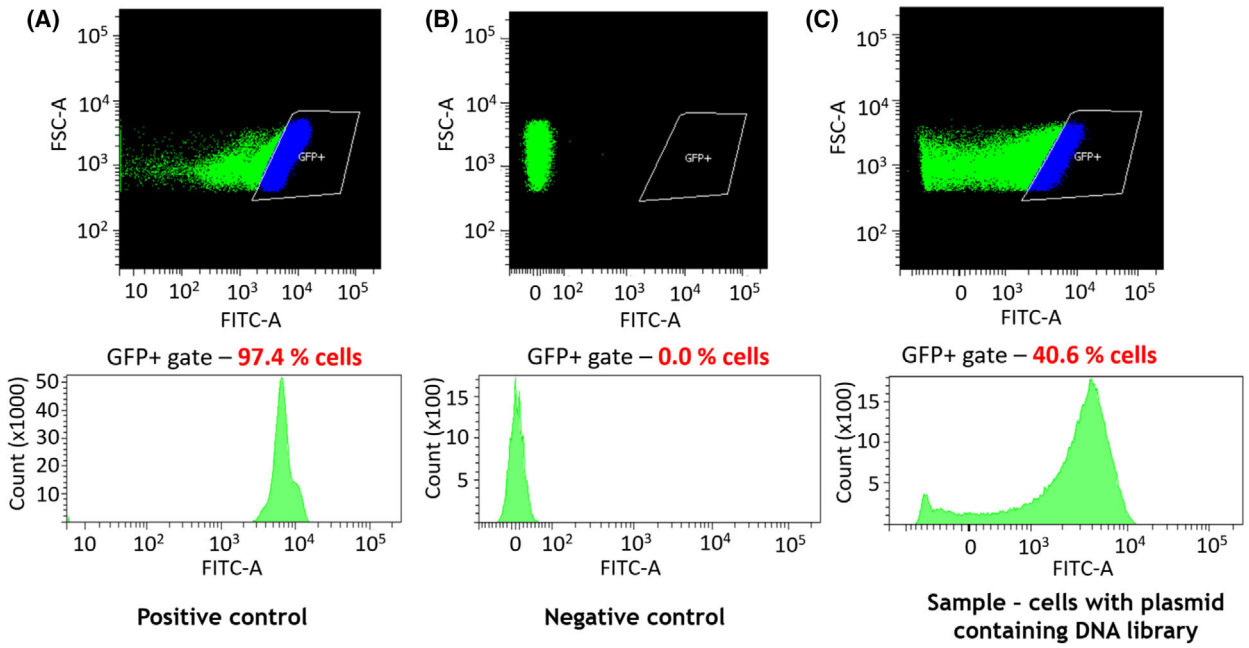


Fig. 2. Initial experimental setup. (A) Positive control showing 97.4% of the singlet cells in the GFP+ gate, which was formed based on this sample. (B) Negative control (pET22b) without GFP, showing no fluorescence and 0.0% of the shown singlet cells in the GFP+ gate. (C) Sample containing the library showing 40.6% of the singlet cells in the GFP+ gate.

GFP gene (positive control) or N-terminal library fused to GFP molecule in the pET22b vector library containing GFP sample (library containing GFP sample). Productions were carried out overnight at 18 °C. Results are shown in Fig. 2.

Based on the positive control, we designed the GFP+ gate, which captured the major population containing 97.4% of the shown singlet cells. We kept the same gate for other samples, showing that the negative control (pET22b sample) contained 0.0% of the shown singlet cells, which formed a separate population exhibiting no fluorescence. The library containing the sample did not form a uniform population regarding the fluorescence level. In this case, the GFP+ gate contained 40.6% of the shown singlet cells, as expected.

This initial experiment had shown that adding the library to the N terminus of constructs affected the yield, since we had observed a much wider distribution of the FITC-A median values across the spectrum, compared to samples without any library.

Enhancing expression yield by selecting the best N-terminal sequences by FACS

In the following paragraphs, we will show results of enhancing yield of two proteins, glutathione S-transferase (GST) and interferon beta (IFNb), all results are summarized in Fig. 7. *Escherichia coli* cells were transformed by pQIK plasmid containing GST gene and used for protein production. Cells were harvested and sorted based on their GFP fluorescence as shown in Fig. 3. Collected fractions of cells (labeled as P2 and P3 during the sorting) were analyzed by SDS/PAGE shown in Fig. 4. This experiment was performed on four samples in total: GST combined with 2 versions of the DNA library (representing four or five amino acids). We utilized two *E. coli* strains: BL21-Gold (DE3), a B strain derivative known for its high transformation efficiency and deficiency in endonuclease I (*endA*), making it ideal for protein production, and XL1-Blue (XL-1), a K-12 strain traditionally used for cloning. While XL-1 also lacks endonuclease I, it is unsuitable for T7 promoter-based protein expression and has a slower growth rate compared to BL21-Gold (DE3). To illustrate the sorting procedure in Fig. 3, we used only one of the samples – GST with the five-amino acid library in *E. coli* BL21-Gold (DE3) cells. During the FACS experiments, we consistently sorted 50 000 cells for each sample to compare their production levels. This method allowed us to directly compare the yield between 50 000 sorted cells plated onto solid LB-agar medium and used for protein production and 50 000 cells carrying the control vector, which were treated the same way.

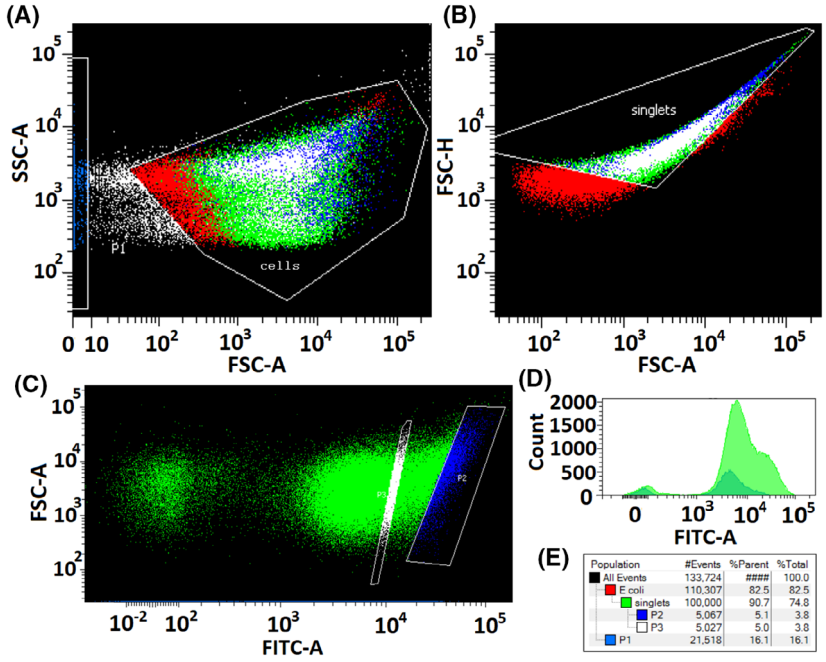
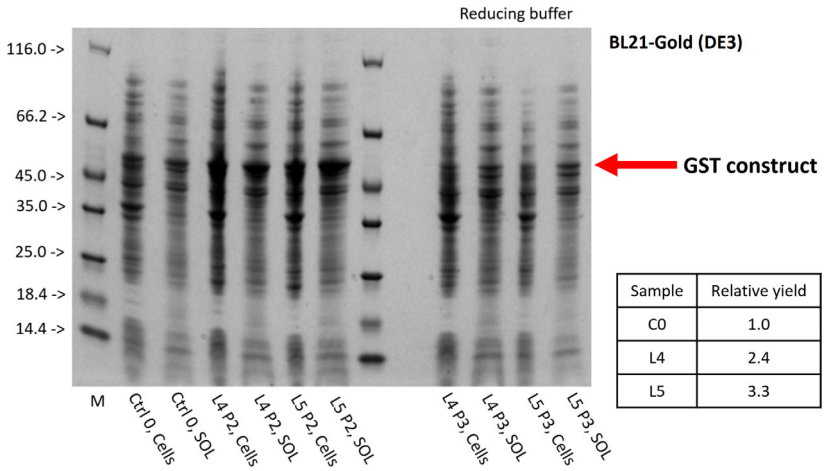


Fig. 3. Sorting procedure of the *E. coli* BL21-Gold (DE3) cells containing pQIK plasmid with GST gene with five-amino acid library. (A) Gating of live cells. We cut out the smallest particles running through the cytometer (hits in the left part of the dot plot). (B) Gating of the singlet cells. The rectangular gate filters out events unlikely to represent singlet cells. As bacteria are near the instrument's detection limit due to their size, this gate excludes extremes to improve singlet cell analysis accuracy. (C) Gating based on the fluorescence level of analyzed cells (gate P2 shown in dark blue and gate P3 shown in white). (D) Histogram of the fluorescence level values of the dot plot shown in (C). (E) Table showing population percentage in all gates. Reason why there are multiple gates for the cell sorting (P2 and P3) is explained in detail in 'Preventing the GFP molecule cleavage'.



Sample	Relative yield
C0	1.0
L4	2.4
L5	3.3

Fig. 4. SDS/PAGE analysis of the GST produced by sorted *E. coli* BL21-Gold (DE3). Expected size of the construct is 53.0 kDa (plus four or five amino acids derived from the N-terminal library). Control samples without any library are labeled Ctrl0, libraries containing samples are labeled L4 and L5 for four- and five-amino acid libraries, respectively. Different samples based on their gating strategy are labeled P2 or P3 as labeled in Fig. 3. Samples labeled Cells contain disintegrated cells after the protein production, samples labeled SOL contain soluble fraction only, which was used to calculate the yield of the protein. The protein yield is summarized in the table on the right showing that we were able to get 2.4 times higher yield out of the sorted cells in the P2 gate with the four-amino acid library and 3.3 times higher yield while working with the five-amino acid library. Molecular weight markers (M, kDa) are indicated on the left.

Sorted cells were plated onto solid agar medium containing carbenicillin and grown overnight. The plate was washed by LB medium, and the protein production procedure was performed. Protein yield including a solubility analysis was analyzed by SDS/PAGE, as shown in Fig. 4. The enrichment was more significant while using *E. coli* BL21-Gold (DE3) strain compared to XL-1 strain, which is why we show *E. coli* BL21-Gold (DE3) data only. Using the BL21-Gold (DE3) cell strain, we were able to increase the yield of the GST fusion 3.3-fold, while the XL-1 cell strain parallel experiment delivered only 1.8-fold enrichment.

Another example of successful protein enrichment through the N-terminal sequence manipulation beside

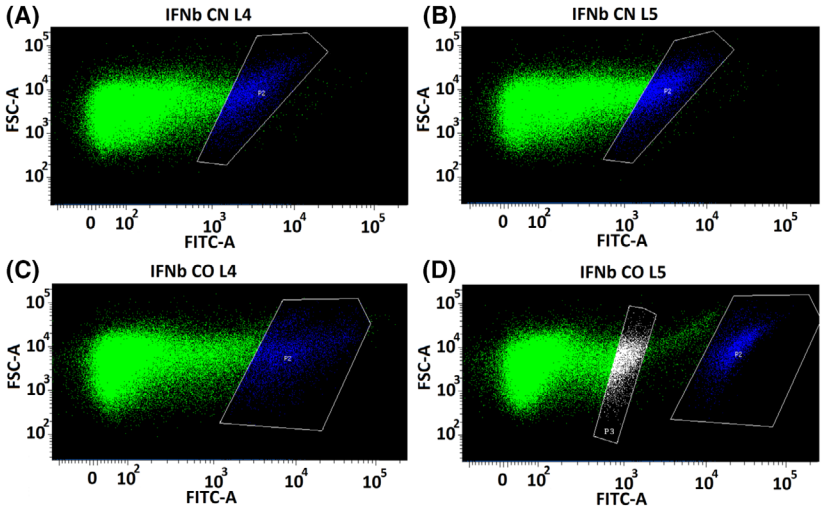
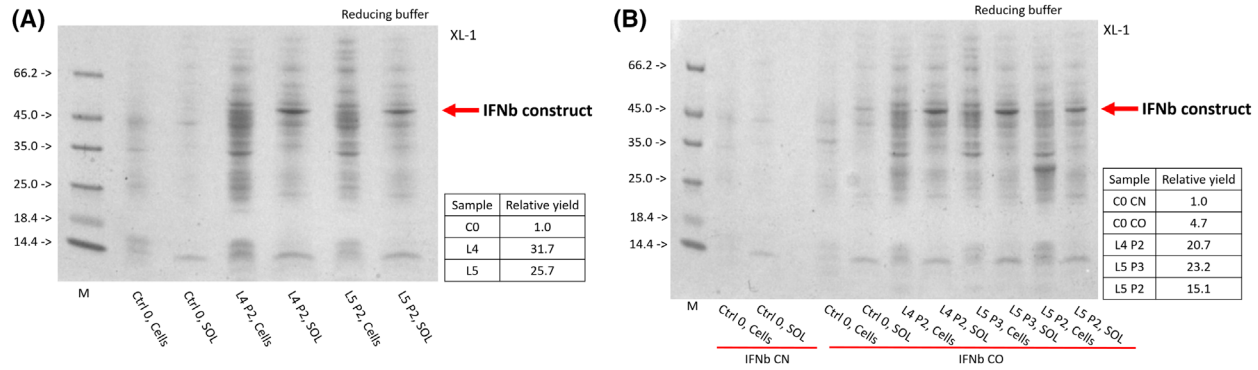


Fig. 5. Gate design during IFNb sorting. IFNb samples (top part (A) and (B)) with natural codons are labeled CN, optimized codon samples (lower part (C) and (D)) are labeled CO. Samples with four-amino acid library (on the left) are labeled L4 and five-amino acid library samples (on the right) are labeled L5. In the case of IFNb CO L5, we designed the P3 gate as well, because the population was not uniform (this strategy is described in more detail in the next chapter Preventing the GFP molecule cleavage). Reason why there are multiple gates for the cell sorting (P2 and P3) is explained in detail in chapter Preventing the GFP molecule cleavage.



Sample	Relative yield
C0	1.0
L4	31.7
L5	25.7

Sample	Relative yield
C0 CN	1.0
C0 CO	4.7
L4 P2	20.7
L5 P3	23.2
L5 P2	15.1

Fig. 6. SDS/PAGE analysis of IFNb yield enrichment. The expected size of the IFNb construct is 47.7 kDa (plus four or five amino acids delivered from the library). Control samples without any library are labeled Ctrl0, libraries containing samples are labeled L4 and L5 for four- and five-amino acid libraries, respectively. Different samples based on their gating strategy are labeled P2 or P3 as labeled in Fig. 5. Samples labeled Cells contain disintegrated cells after the protein production, samples labeled SOL are soluble fraction only, which was used to calculate the yield of the protein. (A) IFNb with natural codons (CN). (B) IFNb with optimized codons (CO) compared to the natural codon variant (first two lanes containing samples). Tables on the right side of each gel shows how many times the yield increased compared to a control with no library (C0). The yield increases are calculated from the soluble fractions. In this case the results are shown as relative multiples compared to a variant with natural codons; optimizing the codons already increased the product yield almost five-fold. Molecular weight markers (M, kDa) are indicated on the left.

the GST construct is IFNb. We used two versions of the gene: with natural codons (labeled CN) and with optimized codons (labeled CO), optimized using the GenScript codon optimization tool. Both versions were cloned into the pQIK vector and transformed into BL21-Gold (DE3) and XL-1 *E. coli* cell strains, produced, and the cells were sorted as described earlier. The entire experiment was conducted similarly to the previous one with the GST construct. The IFNb results are shown in Fig. 5 (cell sorting) and Fig. 6 (-SDS/PAGE analysis). Enrichment in BL21-Gold (DE3) was insignificant (data not shown). However, in the XL-1 cell strain, codon optimization alone increased IFNb yield over 4-fold, and our approach further enhanced production up to over 30-fold. Results of optimization for both IFNb and GST recombinant protein constructs are summarized in Fig. 7 along with other constructs that we tried to optimize using our N-terminal sequence manipulation.

Lanes in all gels, such as the lanes in Figs 6A,B and 10C, respectively, are produced by the same volume of the cell suspension grown from the 50 000 FACS sorted cells. The same number of cells, 50 000, were obtained for both cell cultures, those containing plasmids with variable N termini and cells carrying the control plasmid. Both cell samples were grown under

Target protein	Library size (amino acids)	*E. coli* strain used	relative yield
TrxA	4	XL-1	10.7
	5		15.8
	4	BL21-Gold (DE3)	1.7
	5		4.3
EL222	4	XL-1	6.6
	5		7.1
	4	BL21-Gold (DE3)	not increased
	5		not increased
GST	4	XL-1	1.7
	5		1.8
	4	BL21-Gold (DE3)	2.4
	5		3.4
IFNb (CN)	4	XL-1	31.7
	5		25.7
	4	BL21-Gold (DE3)	not increased
	5		not increased
IFNb (CO)	4	XL-1	4.4
	5		4.9
	4	BL21-Gold (DE3)	not increased
	5		not increased

Fig. 7. Summary of proteins that we enriched using the presented protocol. Each construct was linked to two versions of DNA libraries coding four or five amino acids (2nd column) and produced in two different *E. coli* strains, XL-1 or BL21-Gold (DE3) (3rd column). Relative yields compared to a construct without any library are in the last column. The proteins of interest (1st column) were thioredoxin A (TrxA) (UniProt: P0AA25), EL222 (UniProt Q2NB98 residues 17–225), GST (UniProt P08515 residues 1–217) and interferon beta in two variants (CN = codon natural; CO = codon optimized for production in *E. coli* – Uniprot P01574 residues 23–187).

the same conditions and the equal volumes were used for the SDS/PAGE analysis. Since the goal of this process is to increase the yield of the target protein, and we began with the starting cultures containing the same number of cells, the result suggests that in some cases the sorting process itself enhances the overall efficiency of the presented protocol by sorting cells showing higher viability as well.

Preventing the GFP molecule cleavage

In some cases, we experienced a spontaneous cleavage of the GFP molecule from the rest of the construct, which usually results in higher fluorescence of the cells during the cell sorting. Whenever the cleavage occurs, the cell population on the dot plot during sorting appears to lack uniformity, serving as a cautionary signal in the sorting process, as shown in Fig. 8. First time we noticed this problem was while trying to enhance production of human interleukin 9 (IL9).

To prevent selection of clones with cleaved GFP, which occurs even during production of the recombinant protein after sorting, we came up with a solution, which utilizes a different gating strategy. Precise gate design is crucial to separation of these cells. When fluorescence levels appear non-uniform (as illustrated in Fig. 8D), it is important to collect not only the highest fluorescence population but also the second highest. An illustrative example of gate design is presented in section B of Fig. 8. We demonstrate this phenomenon on thioredoxin A (TrxA) with a five-amino acid library. During this experiment the production was carried out at 37 °C and took 3 h. The production was repeated using sorted cells under the same conditions. Results are shown in Fig. 8.

In one of the vectors, we used (pKIK, described in more detail further in the text), there is an HRV3C protease cleavage site to cleave the GFP after the selection. To avoid the unwanted cleavage, we replaced the HRV3C cleavage site with a tetrapeptide GGSG linker and dodecamer (GGSG)3 linker. We also created one version without any linker between the IL9 gene and the GFP. These 3 new constructs were expressed, and their expression compared to the performance of the original vector. The ratio of the cleaved and not cleaved versions was compared by a western blot using anti-His antibody conjugated with a peroxidase, revealing improved coherence of all three new constructs (data not shown) so that the cleavage was prevented. However, the overall expression of IL9 was too low even after the cell sorting, likely due to inability of the bacterial expression system to accommodate the used IL9 constructs. We therefore decided not to further study this protein in this project.

Four- and five-amino acid library is sufficient

Based on the current state of knowledge, we designed three different library sizes – 9, 12, and 15 nucleotides coding for three, four, and five N-terminal amino

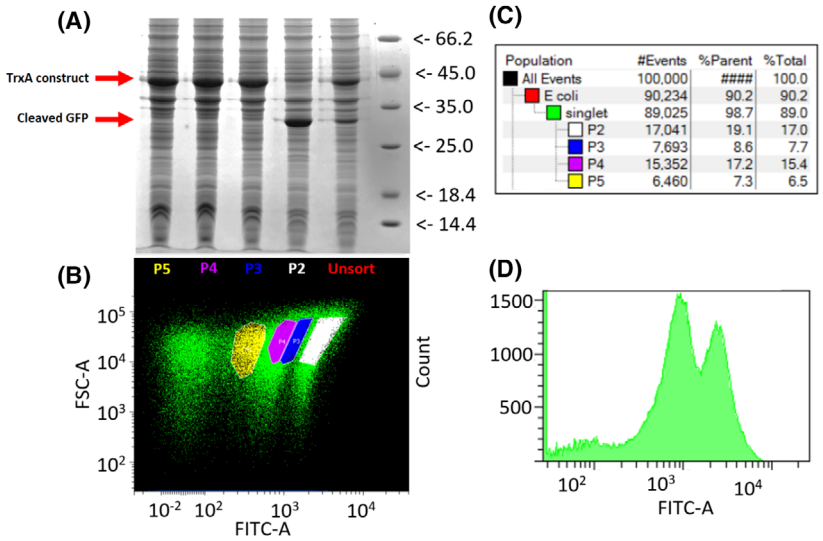


Fig. 8. Identification of GFP cleavage in cell-sorting experiments. (A) SDS/PAGE analysis of the produced constructs. Expected size of the whole construct (TrxA + GFP) is 40.6 kDa. Expected size of the cleaved GFP only is 27.5 kDa. In the P2 sample, the major band corresponds to cleaved GFP, while in sample P3, P4 and P5 the major band represents the whole TrxA-GFP construct. Molecular weight markers (kDa) are indicated on the right. (B) Flow cytometry showing the TrxA-GFP construct with the five-amino acid library. Four different gates (P2–P5) were designed to obtain cells from the whole fluorescent spectrum: P2 (white) for the highest fluorescent population, P3 (blue) for the top of the second highest fluorescent population, P4 (magenta) for the middle part of the same population, and P5 (yellow) for the lowest part. It turned out that the P2 population represented cells with cleaved GFP inside. (C) Table showing population percentage in all gates. (D) Histogram of the fluorescence level values of the dot plot shown in (B).

acids. We selected these sizes with the consideration that utilizing a shorter library would likely result in a better protein product for subsequent experiments. These library versions were cloned on the N terminus of the GFP gene produced in an "empty" pQIK vector. These vectors were transformed into both XL-1 and BL21-Gold (DE3) *E. coli* cell strains alongside a control vector lacking any library insertion.

Following the production, cell cultures underwent FACS-based sorting procedures. Sorted cells were harvested and used for a protein expression afterwards to analyze the difference in the construct expression in sorted cells and the control sample. SDS/PAGE analysis and differences in the yield calculated using ImageJ [42] are shown in Fig. 9.

Following the outcomes of this experiment, the decision was made to progress with the 12 and 15 nucleotide libraries because the smallest version containing 9 nucleotides yielded the least favorable results.

It is also worth noting that yield of proteins that are easily expressed is hard to optimize using this method since their original N terminus is more likely to be closer to the optimal composition. We stumbled over this limitation while failing to significantly increase the yield of easy-to-produce photoactive yellow protein and GFP by manipulating N-terminal parts of their genes.

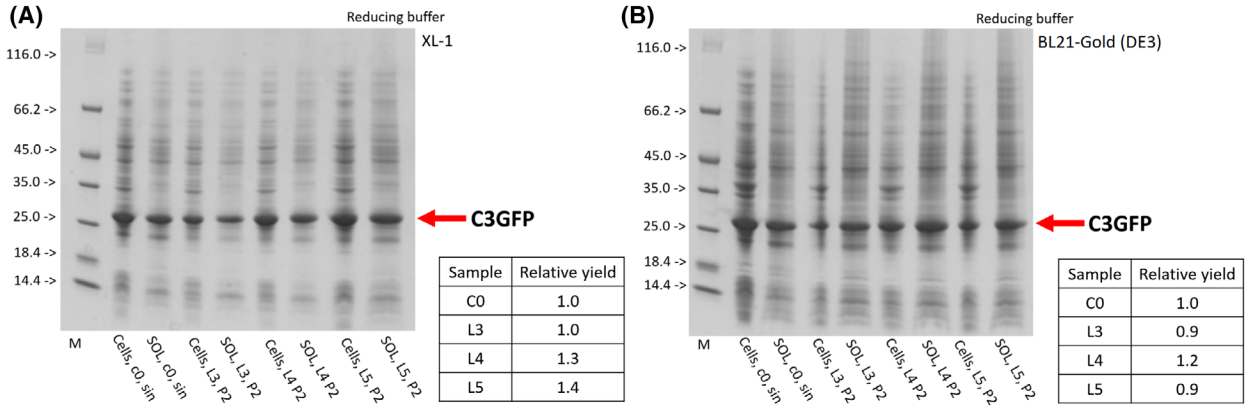
One round of cell sorting is enough

To quantify the impact of the number of rounds of cell sorting on the yield of recombinant protein production, we ran two rounds of the cell sorting with two different proteins (TrxA and EL222) in three different variants: no library, library coding four amino acids and library coding five amino acids.

Sorted cells were plated onto solid agar medium and used for protein production the next day. 0.5 mL of the suspension was taken from each sample and used for the second round of the cell sorting. The rest of the cell suspensions were harvested, and the solubility analysis was performed. Protein yields were analyzed by SDS/PAGE (Fig. 10). The same analytical procedure was then repeated with the cells sorted during the second round. The results, summarized in Fig. 10 (SDS/PAGE), and Fig. 11 (the yield analysis), showed that adding another round of the cell sorting was not improving the overall protein yield.

The P2 gate should aim for the top 5% of the cells with the highest fluorescence and 50 000 cells

Experiments we conducted to find out the correct size of the sorting gate showed that the gate should contain approximately 5% of the singlet cells. If the



Sample	Relative yield
C0	1.0
L3	1.0
L4	1.3
L5	1.4

Sample	Relative yield
C0	1.0
L3	0.9
L4	1.2
L5	0.9

Fig. 9. SDS/PAGE analysis of the produced GFP constructs. Digested cells (labeled "Cells") and soluble fractions (labeled "SOL") of desired constructs. Expected size of the GFP construct is 27.5 kDa (plus four or five amino acids derived from the N-terminal library). Molecular weight markers (kDa) are indicated on the right. Tables in A and B show the relative yield calculated from the soluble fractions using ImageJ. (A) The GFP construct produced in XL-1 cells. (B) GFP construct produced in BL21-Gold (DE3) cells.

percentage was lower, meaning that the gate was aiming for more strict selection of the cells, the overall yield was not higher and the samples showed more contamination. During cell sorting, we sorted 50 000 cells, which was optimal for covering a 9 cm Petri dish while still allowing for single colony selection if needed.

Solubility of optimized protein variants

Another outcome of this experiment is a notable improvement of the solubility of expressed variants, which is often overlooked in the literature. The solubility improvement is demonstrated in Fig. 10C for TrxA protein. TrxA is present in the Ctrl0 sample (lane labeled Ctrl0, Cells), but there is a very low amount of the construct in the soluble fraction (lane labeled Ctrl0, SOL). After cell sorting, it is evident that the ratio of the protein in the digested cells and soluble fraction has shifted towards the soluble fraction. See lanes containing cells from the P2 gate. For example, in lanes L4 P2, Cells and L4 P2, SOL, there is a significantly higher amount of the product in the L4 P2, SOL lane compared to the C0 lanes.

The N-terminal sequences cannot be used universally

N-terminal sequences used to benchmark success of the yield optimization

To investigate whether the same N-terminal sequence can be used to optimize yield of multiple proteins we designed a set of 10 sequences that can be placed after the start codon of produced proteins. These sequences are taken from literature, copied from common vectors, and from well-expressed genes as indicated in Table 1. These sequences serve for benchmarking the success of the yield optimization.

Vector selection for testing the benchmarking sequences

We designed two different vectors called pKIK and pQIK. The first expression vector, pKIK, is based on commercially available vector pET22 and the cassette contains either a DNA library or a benchmark sequence, cloning site with *NotI* and *AscI* restriction enzyme sites, followed by a HRV3C protease cleavage site to cleave the following GFP with C-terminal His-Tag. The second vector, pQIK, based on Qiagene's pQE30, contains a simpler expression cassette without multiple cloning sites and it also lacks the HRV3C cleavage site. Cloning procedures were designed in a way, which does not require any restriction enzyme sites, so there is a library or a benchmark sequence after the start codon, followed by C-terminal His-Tagged GFP. See Data S1 (Plasmid_sequences) for detailed description of vector sequences.

Originally, we started this project with the pKIK vector, and we decided to implement a different plasmid while trying to prevent the GFP cleavage (discussed above). This is also one of the reasons why the expression cassette of the newer pQIK vector is simpler compared to the original pKIK vector. Vector of choice for the FACS experiments was the pQIK vector for its simpler composition and ability to be used in more different *E. coli* cell strains. The expression cassettes of both vectors are shown in Fig. 12. Main

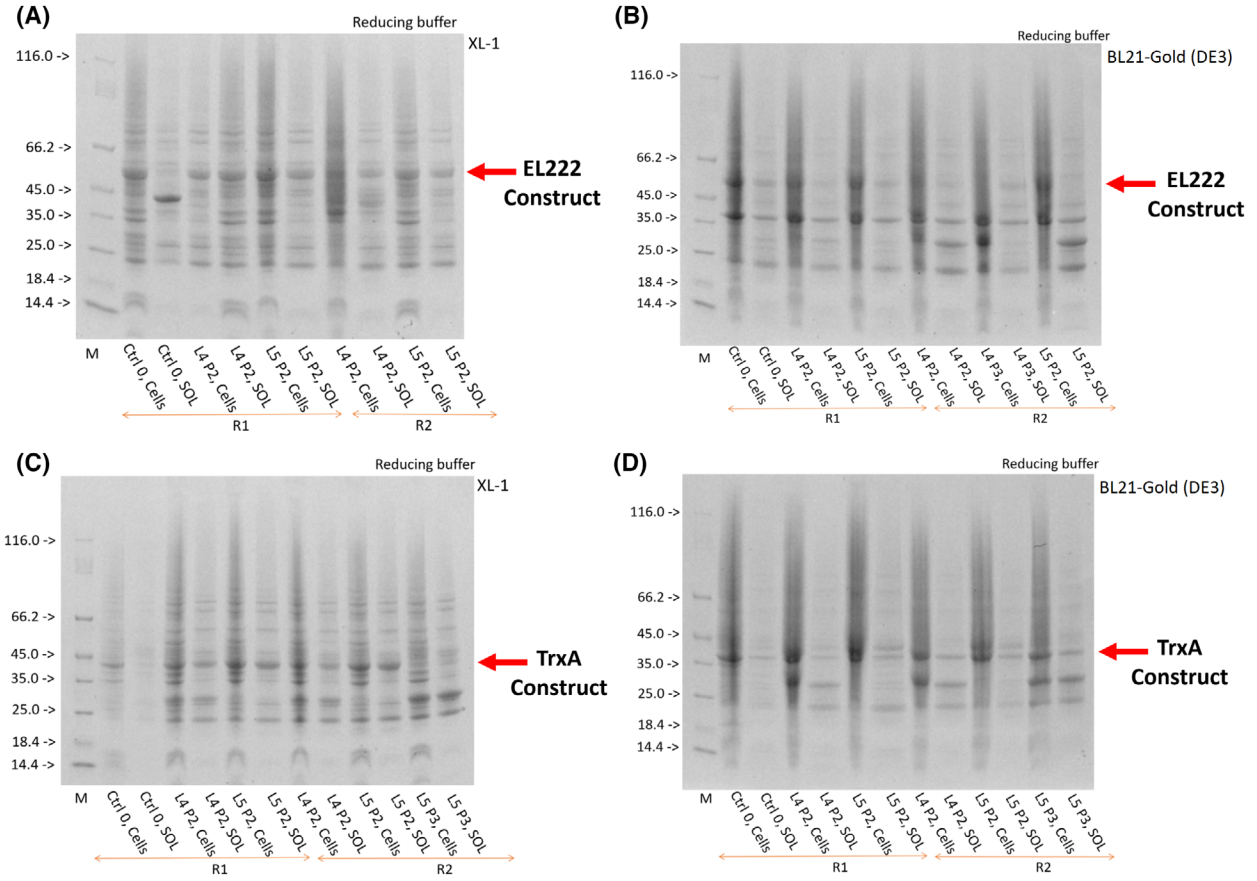


Fig. 10. SDS/PAGE analysis of the produced constructs. Digested cells (labeled "Cells") and soluble fractions (labeled "SOL") of desired constructs. Expected size of the cleaved GFP is 27.5 kDa. (A) EL222 produced in XL-1 cells, expected size of the whole construct (EL222 + GFP) is 50.8 kDa (plus four or five amino acids derived from the N-terminal library). (B) EL222 produced in BL21-Gold (DE3) cells, expected size of the whole construct (EL222 + GFP) is 50.8 kDa (plus four or five amino acids derived from the N-terminal library). (C) TrxA produced in XL-1 cells, expected size of the whole construct (TrxA+GFP) is 39.5 kDa (plus four or five amino acids derived from the N-terminal library). (D) TrxA produced in BL21-Gold (DE3) cells, expected size of the whole construct (TrxA+GFP) is 39.5 kDa (plus four or five amino acids derived from the N-terminal library). Molecular weight markers (M, kDa) are indicated on the left. R1 and R2 refer to rounds 1 and 2 of selection, respectively.

differences between pKIK and pQIK vectors are also shown in Table 2.

To test the influence of the benchmark sequences on the recombinant protein yield, we performed protein production at 37 °C for 3 h and harvested cells were used for a FACS experiment. The comparison of fluorescence among all samples was based on FITC-A median values (Fig. 13).

We decided not to report specific sequences isolated from sorted cells that exhibited higher protein yields. These sequences are highly construct-dependent and are not effective for all constructs. Therefore, their general use could potentially mislead researchers by suggesting a universal solution to this complex problem, which involves multiple molecular mechanisms.

Conclusions

This study presents a novel approach for enhancing recombinant protein yield in *E. coli* expression systems. Utilizing a randomized N-terminal library and fluorescence-activated cell sorting (FACS) selection, we achieved up to 30-fold yield increases for various recombinant proteins (Fig. 7). This approach allows for high-throughput screening of tens of thousands of cells per second containing target protein variants. Significantly, the protocol prioritizes the soluble fraction of produced proteins leading to products more suitable for downstream applications compared to methods relying solely on rational design.

Our findings highlight the potential of N-terminal optimization for improving protein expression.

Color

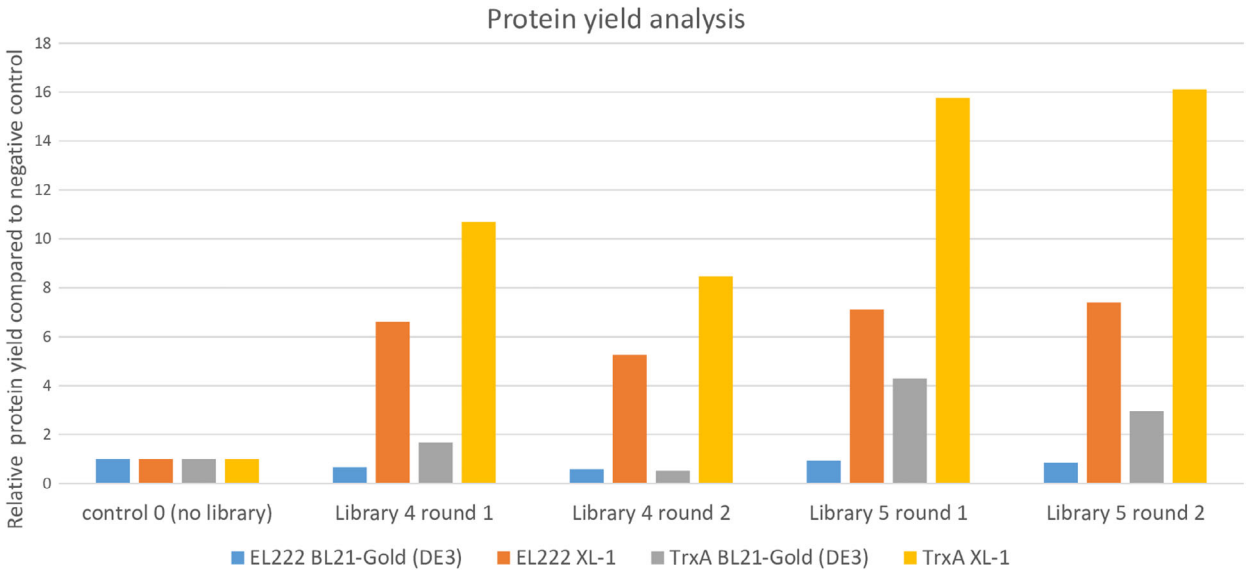


Fig. 11. Chart comparing relative yields of produced proteins (TrxA and EL222) in both BL21-Gold (DE3) and XL-1 *E. coli* cells. The yield was calculated from the soluble fraction on the SDS/PAGE analysis (Fig. 10). Both versions of the library are shown, library 4 referring to a library version containing 12 nucleotides and library 5 referring to a 15 nucleotides library. The set of expressions was performed once prioritizing the exploration of diverse constructs, cell types, and library versions.

Table 1. Benchmark sequences: composition and sources.

Name	Residues	Sequence	Source of the sequence
S01	MGSS	ATGGGCAGCAGC	pET28 vector
S02	MRGS	ATGAGAGGATCG	pQE30 vector
S03	MSKIK	ATGTCTAAAATAAAA	[9,18]
S04	MSKIK	ATGTCTAAAATTAAA	Inspired by [9] but changed according to [12]
S05	MGSDKI	ATGGGTTCAGACAAAATT	ThioHis peptide (M – G – TRX protein)
S06	MSDKI	ATGAGCGATAAAATT	Trx protein, pETM20 based on Addgene plasmid # 176223; http://n2t.net/addgene:176223; RRID: Addgene_176 223; UniProt: P0AA25; *E. coli*
S07	MKIEE	ATGAAAATCGAAGAA	MBP protein, pETM41 based on Addgene plasmid # 38334; http://n2t.net/addgene:38334; RRID: Addgene_38334
S08	MSKEK	ATGTCTAAAGAAAAA	EF-Tu protein; UniProt: P0CE47; AAA50993.1
S09	MKKIA	ATGAAAAAGATTGCA	From rpiB protein (Ribose-5-phosphate isomerase B); UniProt: P37351; *E. coli*; CAA57688.1
S10	MVKIH	ATGGTGAAGATTCAT	[12]

However, finding a universally optimized N-terminal sequence is unlikely and each protein requires its own optimization process, which can be time-consuming and may not always be successful. The method is not suitable for all proteins, particularly those inherently toxic to *E. coli*, those with structures incompatible with bacterial expression. Indeed, we observed limited yield increase for proteins erythroferrone, interleukin 9, and certain thioredoxin A constructs.

In conclusion, this protocol offers a valuable tool for increasing the production of proteins with low yields. This optimization process can also help to get a better ratio of soluble versus insoluble fraction of a protein of interest.

Materials and methods

Cloning

We used several methods of plasmid cloning. The first one was a homologous recombination-based commercial In-Fusion cloning technique (TaKaRa Bio, San Jose,

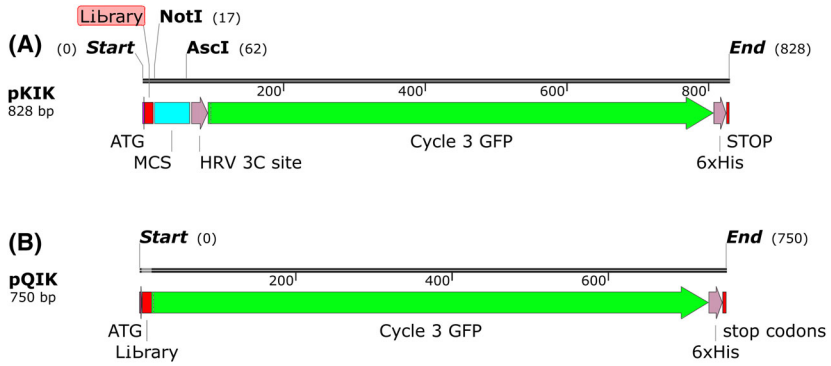


Fig. 12. Scheme of the expression cassettes. (A) pKIK vector set. Protein of interest can be cloned using *NotI* and *AscI* sites into the multiple cloning site (MCS) (B) pQIK vector set. Proteins of interest can be cloned using In-Fusion Restriction-Free cloning techniques between the N-terminal library and the GFP gene.

Table 2. pKIK and pQIK vector set main differences.

Vector pKIK	Vector pQIK
Based on pET22	Based on pQE30
T7 promoter (strong)	T5 promoter (weak)
Restricted *E. coli* strains usage (limited to DE3 cell strains)	Suitable for more *E. coli* strains (TOP10, XL-1, BL21-Gold (DE3)...)
Lac repressor LaqI	No Lac repressor (leaky system)
Production-focused expression cassette design	Clean proof of concept-focused expression cassette

CA, USA). The second method we used was a Restriction-Free (RF) cloning (protocol described elsewhere [43]). Genes coding the examined proteins were ordered in a form of lyophilized DNA string (Thermo Fisher, Waltham, MA, USA). Primers were designed according to TaKaRa Bio Infusion reaction protocol [44] and synthesized (Generi Biotech, Hradec Králové, Czech Republic). Primers for the RF cloning and In-Fusion differed only in their length. Primers for infusion were shorter, having 18–25 bp annealing parts and overhangs, while primers for RF cloning were approximately 31–38 bp long in both regions. DNA libraries were placed directly in the primers, so the library complexity would not get lower during the PCR reaction and primers were designed with a preference to have a G or C nucleotide on their ends, while the melting temperature for both forward and reverse primers was matched. Typical set of primers used in this case for RF cloning the GST gene into the pQIK vector looked like this:

GST-Q-4-FW

5′ *CACAGAATTCATTAAAGAGGAGAAATTAACTATG* NNNNNNNNNNNN**ATGTCCCCTATACTAGGTTA TTGGAAAATTAAGGGC** 3′

GST-Q-REV

5′ **GTGAAAAGTTCTTCTCCTTTGCTAGC***TGGAGGA TGGTCGCCACCACC* 3′

Primers for In-Fusion cloning of the EL222 gene into the pQIK vector looked like this:

EL2-Q-4-FW

5′ *CACAGAATTCATTAAAGAGGAGAAATTAACTATG* NNNNNNNNNNNN**GGGGCAGACGACACACG** 3′

EL2-Q-REV

5′ **GTGAAAAGTTCTTCTCCTTTGCTAGC***GATTCCGG CTTCGACGGC* 3′

where the library (in this case 12 random nucleotides coding four random amino acids) is labeled by letter "N", the annealing part of the primer for the insert is shown in bold and the italic part shows overhang for the infusion cloning, which is complementary to the pQIK vector.

Running an overlap extension PCR with the GFP gene and inserting the whole construct (instead of just inserting the protein of interest) usually delivered higher efficiency, so this extra step was included to obtain samples with higher library complexity.

To replace, delete or add a library or a benchmark sequence in an existing sample, we utilized a KLD cloning technique (New England Biolabs, Ipswich, MA, USA), following the manufacturer's protocol [45]. Typical set of primers (in this case for replacing the benchmark sequences) looked like this:

QIK_T-c1_KLD_FW

5′ *GGCAGCAGC***ATGAGCGATAAAATTATTCACCTG AC** 3′

SH-pQ_KLD_REV

5′ **AGTTAATTTCTCCTCTTTAATGAATTCTGTG** 3′

where the annealing part of the primer is shown in **bold** and the *italic* part shows the new sequence replacement/insertion (in this case benchmark sequence S01 without the start codon, which is already present in the vector).

Traditional cloning based on restriction enzymes cleavage and T4 DNA ligase (New England Biolabs, Ipswich, MA, USA) was used just for a minor part of the samples without DNA library.

Name	TrxA pQIK, XL-1	TrxA, pQIK, BL21-Gold (DE3)	TrxA, pKIK, BL21-Gold (DE3)	GST, pKIK, BL21-Gold (DE3)	only GFP, pKIK, BL21-Gold (DE3)	ERFE, pKIK, BL21-Gold (DE3)
S01	1,165	790	1,029	832	2,838	5
S02	626	902	1,023	698	2,540	2
S03	439	1,040	736	462	2,565	4
S04	1,339	473	733	490	2,440	5
S05	1,243	339	1,070	668	2,845	4
S06	1,499	823	900	495	2,577	5
S07	1,585	663	741	352	2,614	4
S08	1,856	428	1,040	544	2,890	4
S09	1,437	816	625	297	2,300	5
S10	605	480	328	248	1,241	4

Fig. 13. FITC-A levels of fluorescence measurements of the benchmark samples, their sequences are listed in Table 1. Each value represents the median calculated from measurements of hundred thousand analyzed singlet cells, measured once; the highest values are displayed in dark green, decreasing to white towards lower values for each data set. The last column shows erythroferrone (ERFE) constructs that we were not able to produce. Changing the composition of the N-terminus did not boost the production of this construct.

Primers used for the project are listed in a separate table, see Data S2 (-Primers).

Cloned vectors were transformed into electrocompetent *E. coli* TOP10 cells (Thermo Fisher, Waltham, MA, USA) and plated onto solid LB-agar plates containing carbenicillin (50 $\mu g \cdot mL^{-1}$). We selected carbenicillin over ampicillin for its better stability and to minimize the occurrence of satellite colonies.

For benchmark samples and samples without any library single colonies were picked and plasmid DNA was isolated using E.Z.N.A. Plasmid Mini Kit I (Omega Bio-tek, Norcross, GA, USA). For samples containing libraries, plates were washed using 2.4 mL of PBS to resuspend colonies. 2 mL of the suspension were recovered, and plasmid DNA was isolated using the same kit. Isolated plasmids were sequenced (Eurofins Ebersberg, Germany).

Protein expression

Protein expression

This method of protein production was used only for the samples without DNA library. Single colony was picked from a solid agar LB medium containing carbenicillin (50 $\mu g \cdot mL^{-1}$), transferred into fresh LB medium containing carbenicillin (50 $\mu g \cdot mL^{-1}$) and incubated using Duetz microtiter plate system microplates (24-deepwell) at 220 RPM at 37 °C overnight. Overnight culture was then used to inoculate fresh LB carbenicillin medium in 1:100 ratio. This cell suspension was then incubated under the same conditions as the ON culture. When OD_{600} reached 0.6–0.8, induction was performed by addition of IPTG to a final concentration of 1 mM. The cell suspension was then incubated either at 37 °C for 3 h or at 18 °C overnight. The production suspensions were harvested using centrifugation at 5000 ***g*** for 10 min.

For libraries containing samples, the procedure was the same with these exceptions. Instead of picking a single colony from the plate, the plate was washed using 2.4 mL of production media and this mixture was combined with 8 mL of fresh LB media containing carbenicillin, creating a suspension, which was used directly for protein production without overnight culture. Induction for library containing samples was performed by addition of IPTG to a final concentration 1 mM after 60 min of incubation regardless of the OD of the cell culture.

SDS/PAGE and solubility analysis

SDS/PAGE analysis was done using 4–20% Mini-PROTEAN® TGX™ Precast Gels (BioRad, Hercules, CA, USA). After running the electrophoresis, gels were dyed using the Coomassie Brilliant Blue R-250 Staining Solutions Kit (BioRad) and captured using the Transwhite function of the Azure 300 (Azure Bio Systems, Montréal, Canada). ImageJ software was used to analyze the protein yields using the analyze gel function [42]. To show only the soluble fraction of proteins from the harvested cells, we used a B-PER Complete Bacterial Protein Extraction Reagent (Thermo Fisher). Harvested cells were resuspended in this solution and incubated at RT for 30 min. Then the mixture was centrifuged at 20 000 ***g*** and 4 °C for 20 min to separate the soluble fraction from the bacterial debris. Supernatant (soluble fraction) was taken to a fresh tube for further analysis.

Flow cytometry analysis

Flow cytometry experiments were performed by staff of the Imaging Methods Core Facility at BIOCEV using FACSAria™ Fusion SORP (BD Biosciences, New Jersey, USA). Briefly, 0.5 mL of harvested cells were washed and resuspended in 1.4 mL of PBS buffer. This suspension was used for the flow cytometry and diluted so the number of cells passing through the flow cell did not exceed 20 000 $cells \cdot s^{-1}$ during the cell sorting. We utilized 1.0 ND filter and 70 μm nozzle for sorting of the *E. coli* cells.

During the sorting procedure we sorted 50 000 cells and plated them onto solid agar LB medium containing carbenicillin antibiotics (50 $\mu g \cdot mL^{-1}$). For control samples, we used the FACS for sorting the same amount (50 000) of singlet cells. This way our control samples underwent the same procedure, as our experiment samples and we were able to match the amount of *E. coli* cells plated for each sample.

Gating strategy

Since the size of *E. coli* cell is close to the detection limit of the BD FACSAria™ Fusion Flow Cytometer, which we used for this experiment, we designed a P1 gate, shown in the Fig. 3A, to make sure that the smallest detected particles were living cells not just cell debris. Since we were discarding the majority of the cells running through the sort, we decided to design stricter gating and discard the smallest particles (gate P1) (see Fig. 3A), since we expected this gate to contain several contaminants such as bacterial debris. This was in concordance with our previous finding, which showed that if the gating of live cells was less strict, the P2 gated cells that we wanted to collect from the experiment had shown higher contamination.

Regarding the gating of singlet cells in case of *E. coli* cells, gating the singlet cells is usually not accurate, since bacteria are close to the detection limit of BD FACSAria™ Flow Cytometer. Similarly, as discussed above, we cut out the most extreme hits to minimize the contamination in the final cell suspension.

The gates designed to sort cells emitting the highest fluorescence (gates P2, P3 etc.) follow the limitations regarding sample cleavage discussed in the results section and shown in Fig. 8, which revealed, that if the population emitting the highest fluorescence was not uniform, the GFP may have been cleaved from the desired construct and there should be additional gate designed aiming for the top of population with the second highest fluorescence.

Acknowledgements

The authors acknowledge Imaging Methods Core Facility at BIOCEV, namely Mgr. Ondřej Honc and Michaela Kolařík Zázvorková for their support & assistance with the flow cytometry experiments. This research was supported by project nr. LX22NPO5102 (MEYS): Financed by European Union – Next Generation EU and the institutional grant to the Institute of Biotechnology of the Czech Academy of Sciences RVO 86652036.

Conflict of interest

The authors declare no conflicts of interest associated with this work.

Author contributions

SH, PM and BS conceived the project; SH, YP and PM planned experiments; SH, JS, LS and MH performed experiments and analyzed data; SH, BS, LS and YP wrote the paper; JS created the graphics for the publication.

Peer review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/febs.17376.

Data availability statement

The data that support the findings of this study are openly available in the Zenodo server at https://zenodo.org/records/12772716.

References

1 Rosano GL & Ceccarelli EA (2014) Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front Microbiol* **5**, 172.
2 Overton TW (2014) Recombinant protein production in bacterial hosts. *Drug Discov Today* **19**, 590–601.
3 Gopal GJ & Kumar A (2013) Strategies for the production of recombinant protein in *Escherichia coli*. *Protein J* **32**, 419–425.
4 Jia B & Jeon CO (2016) High-throughput recombinant protein expression in *Escherichia coli*: current status and future perspectives. *Open Biol* **6**, 160196.
5 Makrides SC (1996) Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiol Rev* **60**, 512–538.
6 Bivona L, Zou ZC, Stutzman N & Sun PD (2010) Influence of the second amino acid on recombinant protein expression. *Protein Expr Purif* **74**, 248–256.
7 Goodman DB, Church GM & Kosuri S (2013) Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342**, 475–479.
8 Gutierrez G, Marquez L & Marin A (1996) Preference for guanosine at first codon position in highly expressed *Escherichia coli* genes. A relationship with translational efficiency. *Nucleic Acids Res* **24**, 2525–2527.
9 Ojima-Kato T, Nagai S & Nakano H (2017) N-terminal SKIK peptide tag markedly improves expression of difficult-to-express proteins in *Escherichia coli* and *Saccharomyces cerevisiae*. *J Biosci Bioeng* **123**, 540–546.
10 Stenstrom CM, Jin HN, Major LL, Tate WP & Isaksson LA (2001) Codon bias at the 3′-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. *Gene* **263**, 273–284.

11 Thangadurai C, Suthakaran P, Barfal P, Anandaraj B, Pradhan SN, Boneya HK, Ramalingam S & Murugan V (2008) Rare codon priority and its position specificity at the 5′ of the gene modulates heterologous protein expression in *Escherichia coli*. *Biochem Biophys Res Commun* **376**, 647–652.
12 Verma M, Choi J, Cottrell KA, Lavagnino Z, Thomas EN, Pavlovic-Djuranovic S, Szczesny P, Piston DW, Zaher HS, Puglisi JD *et al*. (2019) A short translational ramp determines the efficiency of protein synthesis. *Nature Comm* **10**, 5774.
13 Stenstrom CM & Isaksson LA (2002) Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3′ side. *Gene* **288**, 1–8.
14 Allert M, Cox JC & Hellinga HW (2010) Multifactorial determinants of protein expression in prokaryotic open Reading frames. *J Mol Biol* **402**, 905–918.
15 Yu ZB & Jin JP (2007) Removing the regulatory N-terminal domain of cardiac troponin I diminishes incompatibility during bacterial expression. *Arch Biochem Biophys* **461**, 138–145.
16 Charneski CA & Hurst LD (2013) Positively charged residues are the major determinants of ribosomal velocity. *PLoS Biol* **11**, e1001508.
17 Looman AC, Bodlaender J, Comstock LJ, Eaton D, Jhurani P, Deboer HA & Vanknippenberg PH (1987) Influence of the codon following the Aug initiation codon on the expression of a modified Lacz gene in *Escherichia coli*. *EMBO J* **6**, 2489–2492.
18 Ojima-Kato T, Nishikawa Y, Furukawa Y, Kojima T & Nakano H (2023) Nascent MSKIK peptide cancels ribosomal stalling by arrest peptides in *Escherichia coli*. *J Biol Chem* **299**, 104676.
19 Kozak M (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361**, 13–37.
20 Tuller T, Waldman YY, Kupiec M & Ruppin E (2010) Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci USA* **107**, 3645–3650.
21 Del Campo C, Bartholomaus A, Fedyunin I & Ignatova Z (2015) Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function. *PLoS Genet* **11**, e1005613.
22 Keller TE, Mis SD, Jia KE & Wilke CO (2012) Reduced mRNA secondary-structure stability near the start codon indicates functional genes in prokaryotes. *Genome Biol Evol* **4**, 80–88.
23 Katz L & Burge CB (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res* **13**, 2042–2051.
24 Kudla G, Murray AW, Tollervey D & Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255–258.
25 de Smit MH & van Duin J (1990) Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc Natl Acad Sci USA* **87**, 7668–7672.
26 Bhandari BK, Lim CS & Gardner PP (2021) TISIGNER.Com: web services for improving recombinant protein production. *Nucleic Acids Res* **49** (W1), W654–W661.
27 Gu W, Zhou T & Wilke CO (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol* **6**, e1000664.
28 Zhou T & Wilke CO (2011) Reduced stability of mRNA secondary structure near the translation-initiation site in dsDNA viruses. *BMC Evol Biol* **11**, 59.
29 dos Reis M, Savva R & Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* **32**, 5036–5044.
30 Pop C, Rouskin S, Ingolia NT, Han L, Phizicky EM, Weissman JS & Koller D (2014) Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol Syst Biol* **10**, 770.
31 Chen GFT & Inouye M (1990) Suppression of the negative effect of minor arginine codons on gene-expression - preferential usage of minor codons within the 1st 25 codons of the Escherichia-Coli genes. *Nucleic Acids Res* **18**, 1465–1473.
32 Ikemura T (1985) Codon usage and transfer-Rna content in unicellular and multicellular organisms. *Mol Biol Evol* **2**, 13–34.
33 Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer-Rnas and the occurrence of the respective codons in its protein genes – a proposal for a synonymous codon choice that is optimal for the *Escherichia coli* translational system. *J Mol Biol* **151**, 389–409.
34 Zamora-Romo E, Cruz-Vera LR, Vivanco-Dominguez S, Magos-Castro MA & Guarneros G (2007) Efficient expression of gene variants that harbour AGA codons next to the initiation codon. *Nucleic Acids Res* **35**, 5966–5974.
35 de Valdivia EIG & Isaksson LA (2004) A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in *Escherichia coli*. *Nucleic Acids Res* **32**, 5198–5205.
36 Puri N, Appa Rao KB, Menon S, Panda AK, Tiwari G, Garg LC & Totey SM (1999) Effect of the codon following the ATG start site on the expression of ovine growth hormone in *Escherichia coli*. *Protein Expr Purif* **17**, 215–223.
37 Cambray G, Guimaraes JC & Arkin AP (2018) Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*. *Nat Biotechnol* **36**, 1005-+.

38 Pham PN, Huliciak M, Biedermannova L, Cerny J, Charnavets T, Fuertes G, Herynek S, Kolarova L, Kolenko P, Pavlicek J *et al*. (2021) Protein binder (ProBi) as a new class of structurally robust non-antibody protein scaffold for directed evolution. *Viruses* **13**, 190.
39 Huliciak M, Biedermanová L, Berdár D, Herynek S, Kolárová L, Tomala J, Mikulecky P & Schneider B (2023) Combined in vitro and cell-based selection display method producing specific binders against IL-9 receptor in high yields. *FEBS J* **290**, 2993–3005.
40 Crameri A, Whitehorn EA, Tate E & Stemmer WPC (1996) Improved green fluorescence protein by molecular evolution using DNA shuffling. *Nat Biotechnol* **14**, 5–319.
41 Fukuda H, Arai M & Kuwajima K (2000) Folding of green fluorescent protein and the cycle3 mutant. *Biochemistry* **39**, 12025–12032.
42 Rasband WS (1997-2018) ImageJ. U.S. National Institutes of Health, Bethesda, MA, USA. https://imagej.nih.gov/ij/
43 Unger T, Jacobovitch Y, Dantes A, Bernheim R & Peleg Y (2010) Applications of the restriction free (RF) cloning procedure for molecular manipulations and protein expression. *J Struct Biol* **172**, 34–44.
44 Takara Bio Inc. USA (2024) In-Fusion® Snap Assembly User Manual. https://www.takarabio.com/documents/User%20Manual/In/In-Fusion%20Snap%20Assembly%20User%20Manual.pdf
45 New England Biolabs (2024) KLD Enzyme Mix Reaction Protocol (M0554). https://www.neb.com/en/protocols/2017/03/31/kld-enzyme-mix-reaction-protocol-m0554

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.
Data S1. Sequences of plasmids.
Data S2. Primers used.