

RESEARCH ARTICLE

Open Access

Massively parallel gene expression variation measurement of a synonymous codon library



Alexander Schmitz¹ and Fuzhong Zhang^{1,2,3*}

Abstract

Background: Cell-to-cell variation in gene expression strongly affects population behavior and is key to multiple biological processes. While codon usage is known to affect ensemble gene expression, how codon usage influences variation in gene expression between single cells is not well understood.

Results: Here, we used a Sort-seq based massively parallel strategy to quantify gene expression variation from a green fluorescent protein (GFP) library containing synonymous codons in *Escherichia coli*. We found that sequences containing codons with higher tRNA Adaptation Index (TAI) scores, and higher codon adaptation index (CAI) scores, have higher GFP variance. This trend is not observed for codons with high Normalized Translation Efficiency Index (nTE) scores nor from the free energy of folding of the mRNA secondary structure. GFP noise, or squared coefficient of variance (CV^2), scales with mean protein abundance for low-abundant proteins but does not change at high mean protein abundance.

Conclusions: Our results suggest that the main source of noise for high-abundance proteins is likely not originating at translation elongation. Additionally, the drastic change in mean protein abundance with small changes in protein noise seen from our library implies that codon optimization can be performed without concerning gene expression noise for biotechnology applications.

Keywords: Sort-seq, Protein abundance, Codon usage, Single-cell, Gene expression variation

Background

Gene expression can vary significantly from cell to cell in an isogenic bacterial population, giving rise to phenotypic variation that affects population survival and fitness, ensemble performance, persistence, bacterial-host interaction, and probabilistic differentiation [1–5]. The underlying causes of gene expression variation are of particular importance to the fundamental understanding of cellular processes, which may enable the development

of methods to control such variation, leading to more effective antibacterial treatments and more efficient bacteria-based biotechnology [6–10].

Cell-to-cell variation in protein abundance can arise from transcriptional, translational, and other processes that govern gene expression. How transcriptional processes affect the variability of gene expression between single-cells has been extensively studied [11–13]. Promoter strength, transcriptional bursting, transcription factor binding strength, as well as the copy number of RNA polymerase and mRNA degradation rate have all been shown to affect variability in mRNA copy numbers, which further affect the variability of protein abundance [14–16]. Parameters in translational processes such as mean translational rate and cell-to-cell variability in

* Correspondence: fzhang@seas.wustl.edu

¹Department of Energy, Environmental and Chemical Engineering, Washington University in St. Louis, Saint Louis, MO 63130, USA

²Division of Biological & Biomedical Sciences, Washington University in St. Louis, Saint Louis, MO 63130, USA

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

translational rate could both, in theory, contribute to variation in single-cell protein abundance [17]. Mean translational rate can be affected by multiple genetic elements, including the strength of ribosome binding sites, mRNA secondary structures, and codon usage, as well as growth-related factors such as charged tRNA concentrations and the copy number of free ribosomes [18]. These genetic elements and growth-related factors may also affect the variability of translational rate between single cells, which further influence variability of protein abundance. Due to this complexity, it is difficult to isolate how each individual parameter affects the variability of protein abundance. Codon usage, for example, has been shown to influence both translational efficiency and transcript stability, with suboptimal codons hindering translation and affecting mRNA stability [18–28]. Codon usage and bias also affect translational dynamics with low abundance tRNA isoacceptors pausing ribosomes [29] and controlling ribosomal traffic [30], particularly at the start of a gene sequence [31]. Despite significant knowledge on the effects of codon usage on mean gene expression, how and to what extent codon usage affects cell-to-cell variability in protein abundance is poorly understood. With codon optimization used as a popular method for enhancing and controlling expression [32], determining any additional consequences, such as on the variability, is important.

In this work, we constructed a library of green fluorescent protein (GFP) reporters with different synonymous codons at their 5' coding sequence and expressed this library in *Escherichia coli* growing in defined glucose medium. We developed a high-throughput method that involves fluorescence activated cell sorting followed by sequencing (Sort-seq) [33] to analyze protein variabilities of 219 different GFP coding sequences within one experiment. Multiple methods were employed to validate the Sort-seq for high-throughput variability measurement. We found that codon usage has a large influence on the mean and variance of GFP abundance. Meanwhile, the squared coefficient of variance (CV^2 , also called noise) varies with GFP mean abundance but shows little difference for sequences with high mean protein abundance. Similar trend was also observed when analyzing variability of *E. coli* native proteins. These results illuminate the influence of codon usage to variations in protein abundance and can be potentially extended to study protein variations in other growth conditions and from other microorganisms.

Results

Design of a Synthetic Gene Library with synonymous codons

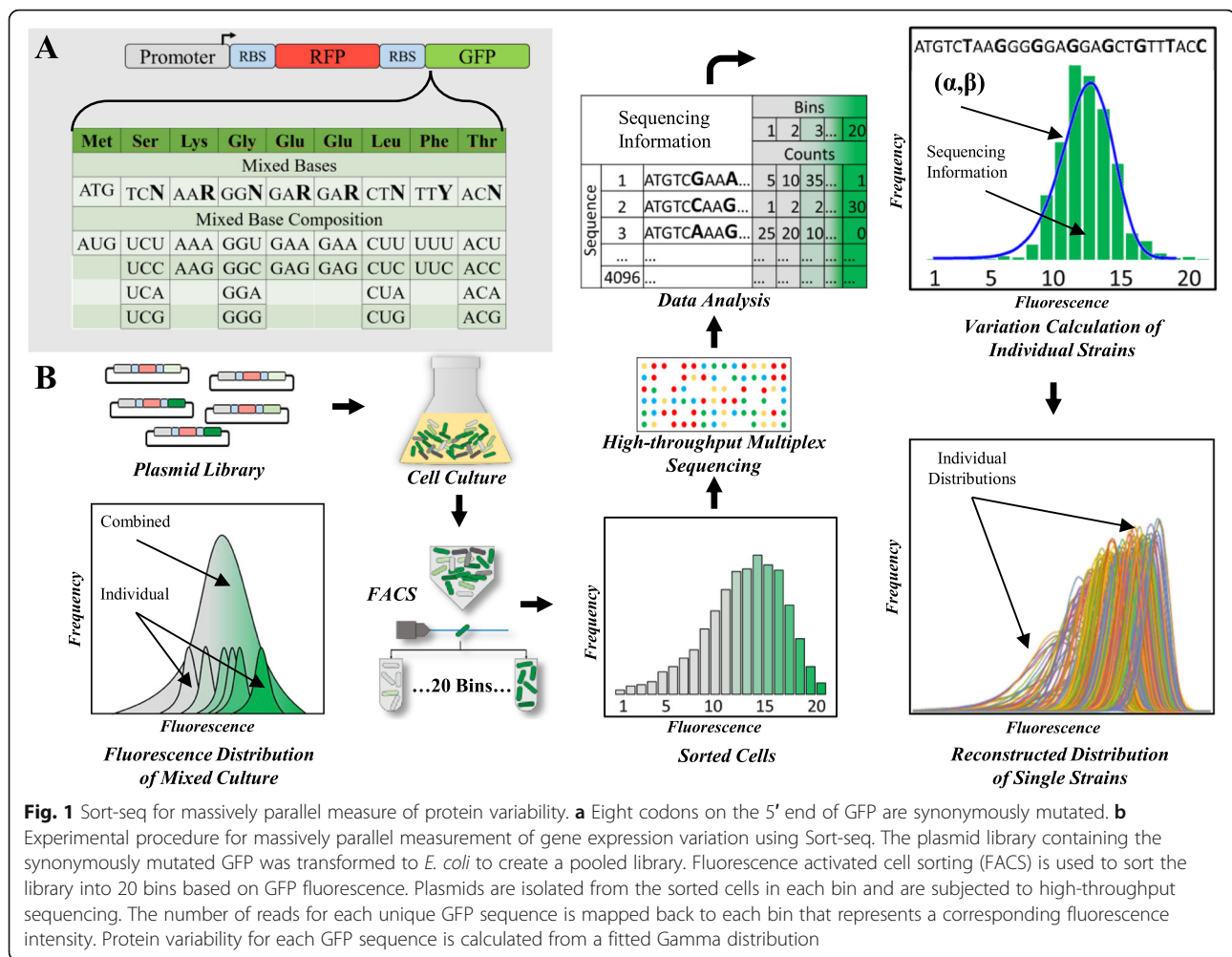
To systematically study the influence of codon usage in cell-to-cell protein variability, a GFP library was

designed with the first 8 codons after the start codon (ATG) randomly mutated to synonymous codons, resulting in a library of 4096 GFP coding sequences. All GFP coding sequences were placed to the 3' of a red fluorescent protein (RFP) with fixed codon usage in a polycistronic structure under the control of the same promoter (Fig. 1a). RFP was used as an internal control to ensure all analyzed cells have transcribed RFP, thus eliminating cells that have lost their plasmid. The synonymous codons were placed at the N-terminal of GFP coding sequence because mean protein abundance is more sensitive to codon usage in this region due to its potential to influence translation initiation, therefore allowing us to analyze protein variabilities across a wide range of protein abundance [22]. The fluorescent reporters were expressed in *E. coli* from a low copy number plasmid (SC101 origin, approximately 5 copies) to minimize burden from gene overexpression [5].

Sort-Seq for high throughput protein variability analysis

Protein variability was previously measured by quantifying single-cell fluorescence of a fluorescent protein using either microscopy or flow cytometry. These methods can measure variability for only one protein sequence at a time. Such low throughputs are insufficient for characterizing large reporter libraries. To solve this problem, we aimed to use Sort-seq [34] to quantify the variations of the GFP library in a massively parallel fashion (Fig. 1b). In this method, single cells are first sorted into different bins based on their GFP fluorescence. Sorted cell mixture in each bin is then sequenced using a distinctive barcode to indicate the bin. The number of reads for each unique GFP sequence is mapped to each bin that represents a corresponding fluorescence intensity. From the distribution of reads, protein variability for each GFP sequence can be calculated.

To validate the method, we first tested the number of bins that allow accurate determination of protein variability. A total of 10 testing strains from the library were randomly selected and their single-cell fluorescence distribution was measured using flow cytometry (Supplementary Figure S1). An increasing number of virtual bins were applied to each sample based on single-cell fluorescence intensity using either linear or exponential fluorescence scales to simulate the bins used in Sort-seq. The mean fluorescence for cells in each virtual bin was applied to all cells within that bin and the GFP variabilities for each strain were computed as the CV^2_{binned} . These values were compared to the variability directly calculated using the un-binned raw fluorescence distribution (CV^2_{real}). We found a consistent lower error rate when cells were binned using log-spaced fluorescence scales compared to those using linear fluorescence scales (Supplementary Figure S2) and is consistent with



previous work that used log-spaced bins [35]. The percent error of CV^2_{bin} to CV^2_{real} also decreases as the number of bins increases (Supplementary Figure S2). With 20 bins, 8 out of 10 randomly selected strains had errors less than 5%. The other two strains with greater than 5% error at 20 bins showed less than 5% error when using fewer than 20 bins due to flow-cytometer measurement noise. Therefore, to obtain accurate quantification of protein variability, we sorted our library into 20 bins divided using an exponential fluorescence scale (Supplementary Figure S3A, S4). Compared to previous Sort-seq work for measuring mean protein abundance, a much higher number of bins are used here, reflecting the challenge in accurate quantifying of gene expression variations [34].

After sorting, plasmids from each bin were extracted, PCR-amplified using primers containing bin-specific barcodes, and sequenced. The Sort-seq experiment was performed three times to examine consistencies between experiments. A total of 5.7 million reads from 3421 unique GFP coding sequences (out of 4096 possible

members in the designed library) were sequenced, representing 83% coverage of the library. For each unique GFP sequence, the number of cells distributed across different bins is calculated and fitted to a Gamma distribution based on the linearly scaled GFP fluorescence, from which mean, variance, and CV^2 in GFP abundance was calculated (Methods). Here we calculated variabilities from a fitted Gamma distribution, instead of directly from the binned distribution, to reduce the error caused by treating fluorescence as a discrete value at each of the individual bins. The number of cells sorted per unique GFP sequence varies broadly (Supplementary Figure S3B) potentially because different GFP sequences led to different cell growth rates and thus different library member representation prior to cell sorting. We hypothesized that for sequences with too few cells-per-sequence (CPS), its variation calculation may not be accurate due to small sampling sizes. To identify the minimum CPS that provide accurate variability measurements, we grouped GFP sequences using different CPS cut-offs and compared calculated GFP

fluorescence from independent Sort-seq measurements (Supplementary Figure S5). With a minimum CPS cut-off of 20, we obtain good correlation between two separate Sort-seq measurements for both mean GFP fluorescence ($R^2 > 0.94$), variance ($R^2 > 0.81$), and CV^2 ($R^2 > 0.68$). As the CPS cut-off drops below 20, both mean GFP fluorescence and CV^2 correlation decrease dramatically (Supplementary Figure S5). Gating based on the CPS value excluded 92% of available GFP sequences because many GFP sequences have less than 20 cells detected. Additionally, for sequences with CPS greater than 20, we examined GFP mean and CV^2 values measured from three independent Sort-seq experiments. GFP sequences with large percent error in either GFP mean or CV^2 were treated as inaccurately measured and were excluded from further analysis (Supplementary Figure S6). Gating based on percent error in GFP mean and CV^2 removed an additional 1.4% of available GFP sequences. The gating resulted in a total of 219 unique GFP sequences used in our analysis.

The reconstructed Gamma distribution of the remaining sequences overlaps closely with Sort-seq measured fluorescence distribution across replicates (Fig. 2a and b) (Supplementary Figure S7). Additionally, we compared the mean GFP fluorescence measured from Sort-seq with those measured from flow cytometry for 16 randomly-selected individual GFP sequences which showed strong correlation ($R^2 = 0.94$) for mean GFP fluorescence, further validating our method (Fig. 2c).

Codon usage correlates with mean and variance but not CV^2

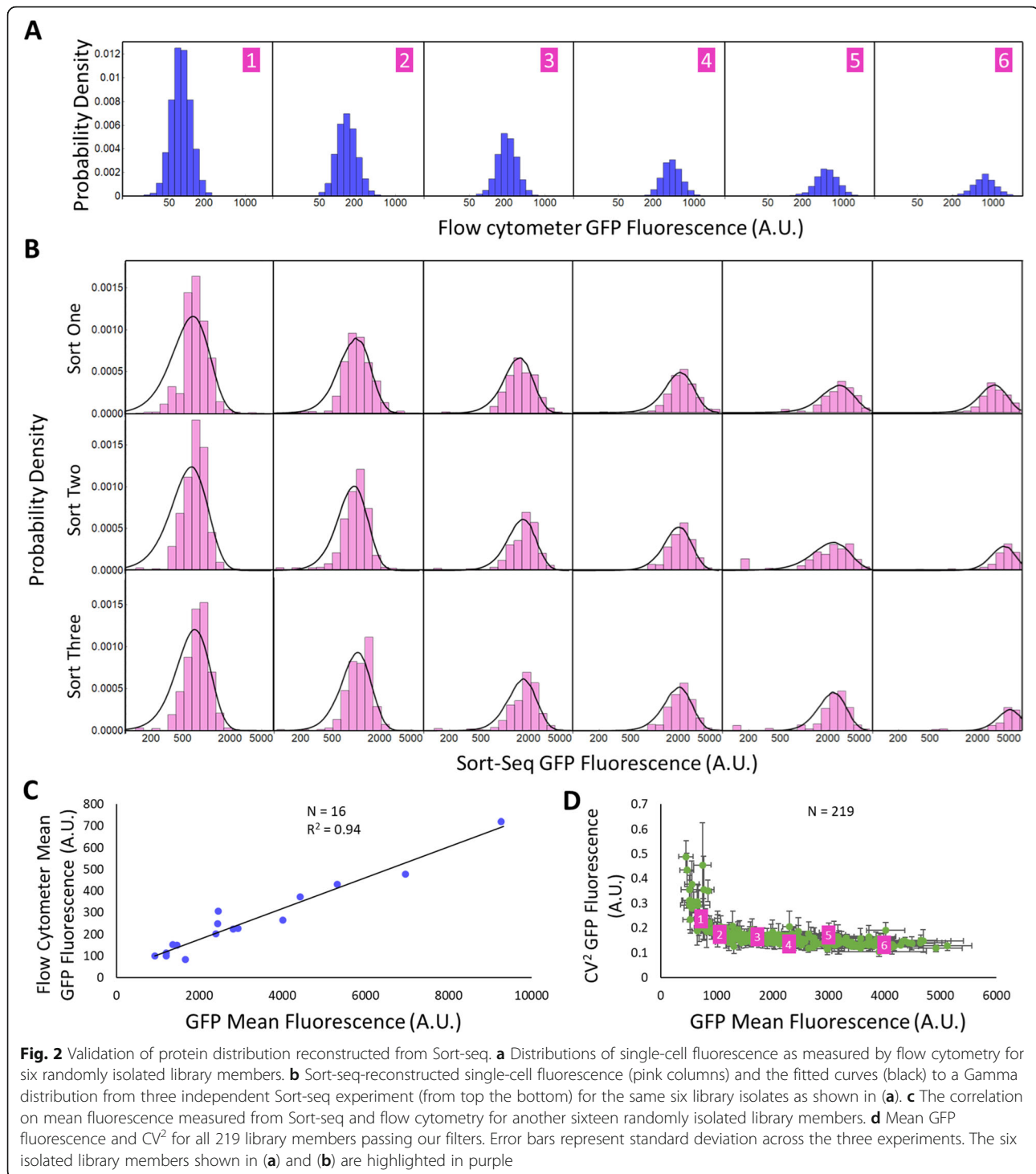
To understand how codon usage affects protein variability, GFP sequences were analyzed based on a few commonly used quantitative metrics of the 8 variable codons, including the tRNA Adaptation Index (TAI), the Codon Adaptation Index (CAI), the Normalized Translation Efficiency Index (nTE) scores and the folding free energy of the mRNA secondary structure (Fig. 3) [36–38]. The measured mean GFP abundance, variance, and CV^2 are compared for each scored group. The mean GFP level correlates weakly with either TAI ($R^2 = 0.23$, $p < 0.001$) or CAI scores ($R^2 = 0.10$, $p < 0.001$), consistent with previous measurements from GFP codon libraries [39]. However, we did not observe significant correlation with the nTE score ($p > 0.1$). Because the nTE score is a measure of cellular competition for tRNAs, the lack of correlation suggests that tRNAs are likely not the rate-limiting factor for GFP translation under our experimental condition (minimal medium with 1% glucose as carbon source). We also did not observe significant correlation between mean GFP fluorescence and the folding energy of 5' GFP mRNA ($p > 0.1$) as previously suggested [39, 40] This is potentially because GFP is the

second coding sequence on the mRNA. In our construct, the GFP start codon is located 22 base pairs after the RFP stop codon, and the ribosome is known to prevent mRNA folding for a region 21 base pairs away from the ribosome A site [41]. Thus, it is likely that the folding energy of GFP mRNA is affected by ribosome translation of the 5' RFP sequence. Similar weak positive correlations are observed between variance of GFP levels with TAI ($R^2 = 0.22$, $p < 0.001$) and CAI scores ($R^2 = 0.08$, $p < 0.001$), but not with the nTE score nor folding energy of 5' GFP mRNA ($p > 0.05$). CV^2 correlates weakly with either TAI ($R^2 = 0.07$, $p < 0.001$) or CAI score ($R^2 = 0.07$, $p < 0.001$), likely due to the fact that CV^2 is large at low GFP levels (Fig. 2d).

Altering the codon usage has a significant effect on the mean expression level, which in turn affects variance and CV^2 . To isolate the influence of codon usage through mean expression level, GFP variance and CV^2 are plotted against mean GFP level. While GFP variance increases with GFP mean (Fig. 4a), GFP CV^2 generally decreases with mean at low GFP abundance and levels off at high GFP abundance (Fig. 4b), consistent with previous observations from genome-wide *E. coli* native gene expression [17]. At high GFP abundance, several sequences with the same mean displayed different CV^2 values, but the differences are within experimental error (Fig. 2d). At high protein abundance, codon usage has little effect on protein CV^2 . Thus, codon usage affects CV^2 mostly via affecting mean GFP level. Meanwhile, codon usage affects protein variance at all gene expression levels. Codons with high TAI or CAI scores increased both GFP mean and variance (Figs. 3 and 4).

Codon usage Bias in the *E. coli* genome

In addition to testing a synonymous codon library of a synthetic gene, we also examined whether similar trends exist for native genes in the *E. coli* genome. Using protein variability of native genes measured from previous work [17], we calculated the TAI, CAI, and nTE scores of their coding sequences for 735 genes for which we had both noise information provided by a previous study [17] and sequence information provided by UniProt [42] (*E. coli* strain K-12) (Fig. 5). From the analyzed genes, weak positive correlations ($p < 0.001$) between mean expression level and TAI or CAI scores was observed, consistent with previous works [37, 38, 43]. No significant correlation ($p > 0.05$) between protein CV^2 with any of the used codon metrics was observed (Fig. 5a). The observations from analyzing *E. coli* native genes are in agreement with results from Sort-seq analysis of our GFP library. Therefore, we conclude that codon usage only influences protein noise by affecting their mean

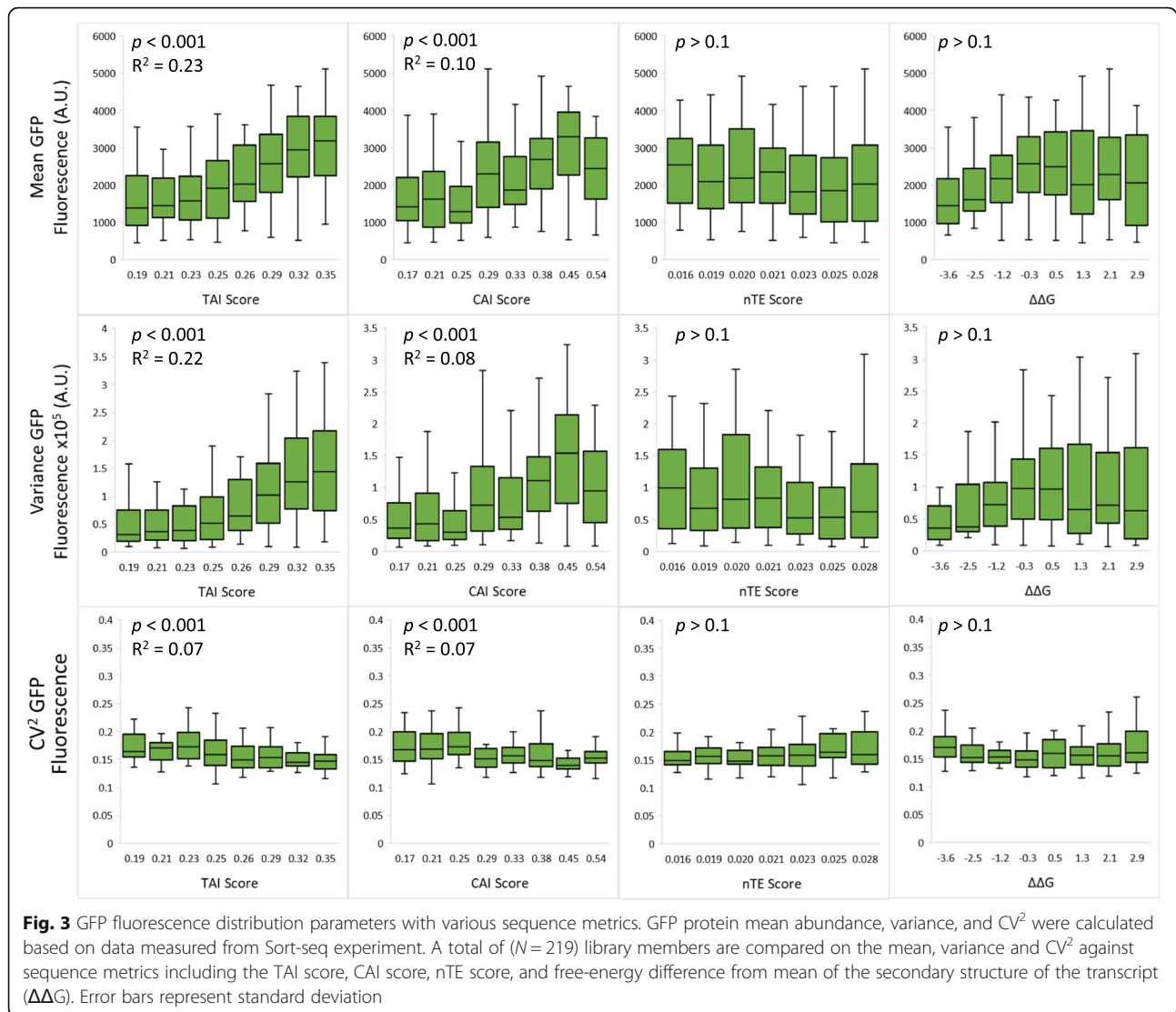


expression levels, with little influence on highly abundant proteins.

Discussion

The analyses performed in this study show that codon usage has a strong influence on the mean protein abundance and variance, with little influence on cell-

to-cell protein variation under the same mean. The altered mean protein expression does not arise from changes in GC content (Supplementary Figure S8) or from mRNA secondary structure (Fig. 3) that could alter translation initiation. For high-abundance proteins, the lack of change in protein variability suggests that cell-to-cell variation in translational rate is not



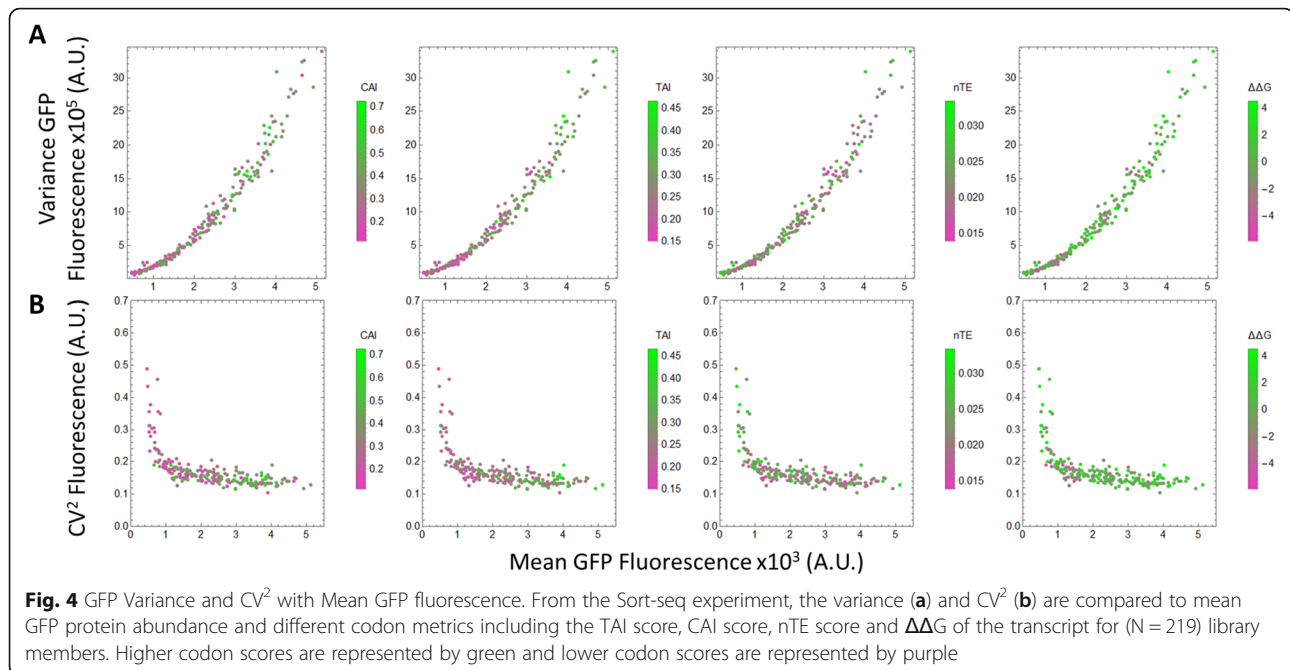
changed significantly when swapping synonymous codons. Rare codons (codons with low CAI scores) tend to decrease the mean protein abundance but only have a small effect on CV^2 . For proteins with codons requiring low-abundant tRNAs (codons with low TAI scores), their overexpression can deplete the availability of charged tRNAs. The lack of change in protein CV^2 when swapping to codons with low TAI scores suggests that the decreased availability of tRNAs does not lead to an increase in cell-to-cell variation of charged tRNAs. This is potentially caused by the tight feedback regulation of tRNAs that would maintain tRNA levels [44]. Furthermore, our results also suggest that the main source of protein noise for high-abundance proteins is likely not translational in origin but rather due to variations in transcription, such as

cell-to-cell variation in RNA polymerase as previously suggested [45].

Conclusions

We observe that synonymously mutation of just eight codons on the GFP changed mean protein abundance by as much as five-fold (Fig. 4) with little to no change in protein noise. The drastic change in protein abundance with small changes in variation indicates that for biotechnology applications, codon optimization can be performed to control gene expression levels without concerning gene expression noise [46].

Our Sort-seq based method represents a high-throughput strategy for measuring gene expression variability. A key parameter to obtain high accuracy in variability measurement is to sort cells into a large enough



number of bins to increase distribution resolution. This method can be potentially extended to other libraries, such as libraries of different promoters or RBSs, and to other organisms, illuminating genetic mechanisms that control cell variability.

Methods

Materials

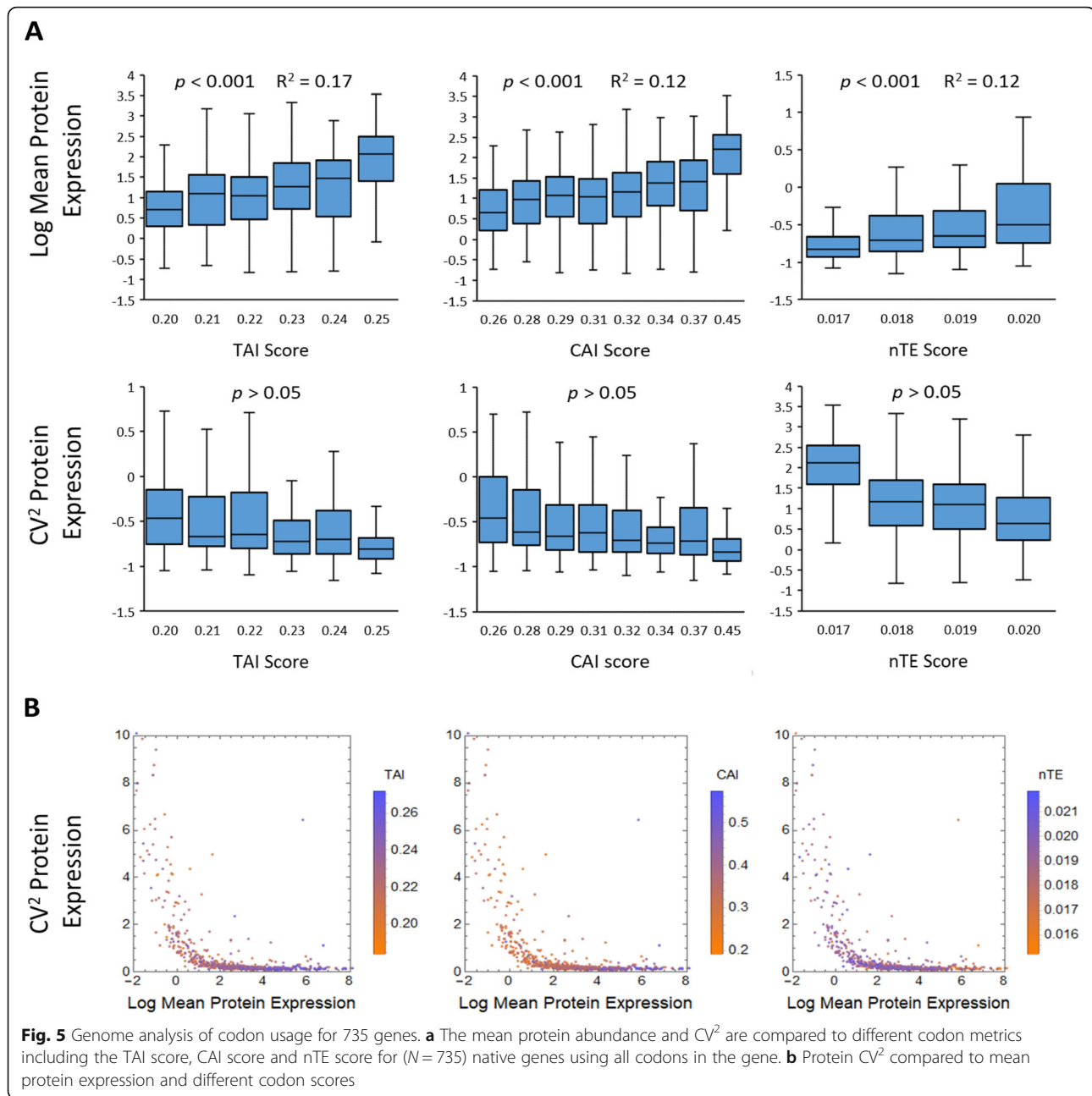
All primers were synthesized by Integrated DNA Technologies (Coralville, IA, U.S.A.). Eco31I and T4 DNA ligase were purchased from Thermo Scientific (Waltham, MA, U.S.A.). All other reagents were purchased from Sigma Aldrich (St. Louis, MO, U.S.A.). All M9 medium was supplemented with 75 mM MOPS, 2 mM $MgSO_4$, 1 mg/L thiamine, 10 μM $FeSO_4$, 0.1 mM $CaCl_2$ and micronutrients including 3 μM $(NH_4)_6Mo_7O_{24}$, 0.4 mM boric acid, 30 μM $CoCl_2$, 15 μM $CuSO_4$, 80 μM $MnCl_2$, and 10 μM $ZnSO_4$. Plasmid DNA purification kits and fragment DNA purification kits were purchased from iNTRON Biotechnology (Seoul, South Korea). High-throughput sequencing was conducted using a MiSeq 2 × 250 standard flow cell from Illumina Inc. (San Diego, CA, U.S.A.). Sanger sequencing was conducted by Eurofins Scientific (Luxembourg). Flow-cytometry was conducted on a Guava easyCyte HT system (Luminex Corp., Austin, TX, U.S.A.) using a 488 nm laser in combination with a 525/30 filter for GFP and a 532 nm laser in combination with a 583/26 filter for RFP. Cell libraries were sorted using a BD FACS Ariall-2 cell sorter (BD Biosciences, Franklin Lakes, NJ, U.S.A.) equipped with a 488 nm laser and a 530/30 nm filter for GFP and a 561 nm laser and a 582/12 nm filter for RFP.

Library construction

To ensure that all library members are synonymously mutated rather than randomly mutated, degenerate primers that allow specific base mutations were used to amplify a super-folder GFP (sfGFP) (Supplementary Table 1A). Both primers contain a Eco31I site for cloning purposes. Plasmid pS5c-RFP-sfGFPlibrary was constructed using one-step Golden-Gate DNA assembly [47]. The GFP library was inserted to the 3' of a RFP coding sequence in a BglBrick plasmid pS5c-RFP [48], which contains a p15A replication origin, a chloramphenicol resistance marker, and a P_{LacUV5} promoter driving the expression of RFP. To do so, the vector backbone was PCR amplified with primers containing Eco31I sites (Supplementary Table 1B). The two PCR amplicons were digested with Eco31I, followed by ligation with T4 ligase following the Golden-Gate protocol [47]. The ligated plasmid library was then chemically transformed into *E. coli* DH10 β competent cells. The transformed library was recovered in 5 mL Luria-Bertani (LB) medium for 2 h at 37 °C and then supplemented with chloramphenicol at 30 mg/mL and grown at 37 °C until reaching OD_{600} 0.08. The culture was then divided into 500 μL aliquots, mixed with 500 μL of 50% glycerol, and stored at -80 °C until use.

Optimizing sorting parameters

The number of bins used for the Sort-Seq protocol was determined using the flow-cytometer data from the ten individual library members. The distribution of GFP fluorescence was divided into different number of virtual bins, and the CV^2 was calculated from the both the bins



and the flow-cytometer data to determine the percent error between the two calculations (Supplementary Figure S2).

Library sorting

Cell libraries were cultivated and treated with ice and rifampicin to halt growth and transcription and additional time was given to allow translated fluorescent protein to mature. Cells were then sorted based on both GFP and RFP fluorescence values. Gates were applied to exclude cells that did not fluoresce at the RFP channel above background, which was set using the fluorescence of

wild type *E. coli* cells. Cells are only included that fluoresced RFP above the maximum RFP fluorescence of the wild type *E. coli* cells. Cells from the GFP library were sorted into 20 bins spaced based on their logarithm of GFP fluorescence. The cells were sorted for a total of eight hours during the second Sort-seq experiment, until a total of 2.16 million cells had been sorted across all 20 bins (Supplementary Figure S3). Fewer cells were sorted during the first and third Sort-seq experiments due to sorting time constraints. 269,000 cells were sorted during the first experiment and 1.89 million were sorted during the third experiment.

High-throughput sequencing

Cells from each bin were subjected to plasmid extraction. Using plasmid DNA from each bin as templates, PCR was performed to amplify the GFP coding sequence containing the variable synonymous codons using primers containing both the Illumina Multiplex sequences (Supplementary Table 2A) with a specific index for each bin (Supplementary Table 2B). PCR was performed for 16 cycles, and the PCR products were gel purified. Purified DNA samples were combined at equal concentrations to produce a 10 nM sample that was then subjected to high-throughput sequencing using a MiSeq system 2 × 250 standard flow cell. A total of 3.9 million reads were generated on the second Sort-seq experiment.

Examining of individual library members

To examine individual library members, 1 μL of the library aliquots was plated onto an agar-LB plate containing 30 mg/mL of chloramphenicol. From the overnight plate, 10 colonies were randomly-selected and cultivated. Their plasmid DNA were then extracted, followed by Sanger sequencing. All 10 plasmids contained the correct GFP coding sequences with non-identical synonymous codons at the expected sites. The 10 overnight cultures were also used to inoculate M9 minimal media containing 1% glycerol and 30 mg/mL chloramphenicol with a starting OD₆₀₀ of 0.0025 and grown at 37 °C. After 2 h, the cultures were induced with 1 mM IPTG and grown until OD₆₀₀ reached 0.08. A low OD is used to prevent clogging the flow-cytometer. Cells were then transferred to ice and incubated for 10 min followed by adding 2 μL of 50 mg/mL rifampicin to halt transcription. The culture was then moved back to 37 °C and incubated for 1 h to allow synthesized fluorescent proteins to fold and mature before flow-cytometry.

Library quality testing

The quality of the library was confirmed by high-throughput sequencing prior to sorting to ensure proper library construction and transformation. In detail, an aliquot of the library culture was grown in 5 mL LB medium overnight. The overnight culture was then used to inoculate 10 mL of minimal M9 medium containing 1% glycerol and 30 mg/mL chloramphenicol with a starting OD₆₀₀ of 0.0025 and grown at 37 °C. After 2 h of growth, the culture was induced with 1 mM of isopropyl β-D-1-thiogalactopyranoside (IPTG). When the culture reached an OD₆₀₀ of 0.08, cells were harvested and treated with ice and rifampicin in a similar way as described above. After sorting, the cells are subjected to plasmid extraction. The GFP coding sequence containing the variable synonymous codons was PCR amplified from the plasmid DNA mixture using primers containing both the Illumina Multiplex sequences and a unique

index for each bin. The Primers used are listed in (Supplementary Table 2A), and the index used is a 9 base pair region in the forward primer and listed in (Supplementary Table 2B). Index 1 was used for the initial library confirmation. PCR was performed for 16 cycles, and PCR products were gel purified. The gel extracted DNA samples were diluted to 10 nM and subjected to high-throughput sequencing. High-throughput sequencing produced 2 million reads with 85% of reads as correct members of the library. From 2 million reads, all possible library members were observed, representing 100% coverage and validating the library construction.

Data processing

Using the index of each read, GFP sequences were sorted into their respective bins. For each unique GFP sequence, the number of reads found in a bin was first normalized by the total number of reads in that bin. The fraction of per unique GFP sequence in each bin was then multiplied by the number of cells sorted into that bin to obtain the number of cells in each bin. GFP sequences that were distributed into less than 2 bins or with less than 20 cells per sequence (CPS) (Supplementary Figure S3B) were removed without analysis. Using all three Sort-seq experiments, error bars are calculated for each library member for both mean GFP fluorescence and CV² of GFP fluorescence. Any library members with above 30% error in mean GFP fluorescence and with above 40% error CV² of GFP fluorescence are excluded from further analysis. Cut-offs in percent error were determined by natural cut-offs in the distribution of the percent error (Supplementary Figure S6). Finally, a total of 219 different GFP sequences were used for protein variability analysis.

To calculate protein variability, each of the 20 bins was assigned a relative protein abundance value based on the fluorescence of each bin. The bins are sorted on their logarithm scale and so are converted to linear scale before fitting. For each GFP sequence, its distribution across 20 bins were fitted to a continuous Gamma distribution using eq. 1.

$$P(x) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha} \Gamma(\alpha)} \quad (1)$$

where x is the fluorescence of each bin, α is the shape parameter and β is the scale parameter representing a gamma distribution. The mean, variance, and CV² of each GFP sequence were calculated from the fitted α and β values using eqs. 2, 3, and 4 respectively.

$$\text{Mean}(x) = \alpha\beta \quad (2)$$

$$\text{Var}(x) = \alpha\beta^2 \quad (3)$$

$$CV^2(x) = \frac{\alpha\beta^2}{(\alpha\beta)^2} = \frac{1}{\alpha} \quad (4)$$

Codon metrics and mRNA folding energy calculations

The CAI score for each GFP sequence was calculated from the eight variable codons using eq. (5) as described previously [36]:

$$CAI(sequence) = \left(\prod_{k=1}^L w_k \right)^{1/L} \quad (5)$$

where L represents the length of the sequence in the number of codons, and w_k is the weight of the k th codon in the gene sequence. The weight for each codon was obtained from previous work [36]. The TAI score was calculated for the same region using eqs. 6, 7, and 8 as previously described [37]. Specifically, utilizing tRNA gene copy as an approximation for tRNA abundance and assuming the tRNA usage of a gene is a measure of how well that gene is adapted to the available tRNA pool. W_i , the absolute adaptiveness value, was first calculated as:

$$W_i = \sum_{j=1}^{n_i} (1 - s_{ij}) tGCN_{ij} \quad (6)$$

where n_i is the number of tRNA isoacceptors that recognize the i th codon. $tGCN_{ij}$ is the gene copy number of the j th tRNA that recognizes the i th codon and s_{ij} is a selective constraint on the efficiency of codon-anticodon coupling as reported previously [37].

$$w_i = \begin{cases} \frac{W_i}{W_{max}} & \text{if } W_i \neq 0 \\ w_{mean} & \text{else} \end{cases} \quad (7)$$

where W_{max} is the maximum W_i value and w_{mean} is the geometric mean of all w_i with $W_i \neq 0$.

$$TAI(sequence) = \left(\prod_{k=1}^L w_{ik} \right)^{\frac{1}{L}} \quad (8)$$

where L is the length of the sequence in number of codons and w_{ik} is the weight of the k th codon.

The nTE score was calculated for the same region for each library member using the method described [38] and is shown in eqs. 9, 10, 11, 12 and 13:

$$U_i = \sum_{j=1}^g a_j c_{ij} \quad (9)$$

$$cu_i = \frac{U_i}{U_{max}} \quad (10)$$

where c_{ij} is the sum of the counts of codon i in gene j and a_j is the transcript abundance of gene j considering all genes in genome g . cu_i is the relative estimate of how often each codon is translated in the genome.

$$nTE'_i = \frac{w_i}{cu_i} \quad (11)$$

$$nTE_i = \frac{nTE'_i}{nTE'_{max}} \quad (12)$$

$$nTE(sequence) = \left(\prod_{k=1}^L nTE_{ik} \right)^{\frac{1}{L}} \quad (13)$$

where w_i is the same as calculated by eqs. 6 and 7. L is the length of the sequence in number of codons and nTE_{ik} is the weight of the k th codon. These same methods were used to calculate the CAI, TAI and nTE for the *E. coli* genome analysis utilizing all codons in each gene (Fig. 5). The free energy of folding for the secondary structure of the transcript was calculated using NUPACK [40] for a region 42 base pairs before and after the codon library. The difference was calculated for each library member compared to the mean free energy of folding of all the library members analyzed. This computation is the same as previously described [39].

Abbreviations

TAI: tRNA Adaptation Index; CAI: Codon Adaptation Index; nTE: Normalized Translation Efficiency Index; GFP: Green Fluorescent Protein; RFP: Red Fluorescent Protein; CPS: Cells-Per-Sequence; FACS: Fluorescence Activated Cell Sorting; LB: Luria-Bertani; IPTG: Isopropyl beta-D-1-thiogalactopyranoside

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-021-07462-z>.

Additional file 1 Supplementary tables and figures. **Table S1:** Primers used for cloning synonymously mutated GFP library. **Table S2:** Primers and indexes used for high-throughput sequencing. **Fig. S1:** GFP fluorescence distribution for 10 isolated library members. **Fig. S2:** Virtual binning of 10 isolated from the GFP library. **Fig. S3:** Sorting cells based on GFP fluorescence. Results for each of the bins from one of the sort-seq experiments. **Fig. S4:** Sort-seq RFP fluorescence. **Fig. S5:** Finding the minimum number of cells to use per sequence (CPS). **Fig. S6:** Percent error between three sort-seq experiments. Using all three sort-seq experiments, percent error is calculated in the measurement of both mean GFP fluorescence and CV^2 . **Fig. S7:** Sort-seq reconstructed single cell fluorescence and the fitted curves to a Gamma distribution for six library isolates. **Fig. S8:** The GC percent content of the synonymously mutated sequence is compared to the mean and CV^2 GFP fluorescence of each sequence.

Acknowledgements

We thank Dr. Barak Cohen for helpful discussions and advice on Sort-seq.

Authors' contributions

A.S. designed and conducted the experiments and wrote the paper. F.Z. assisted with experimental design and writing the paper. All authors have read and approved this manuscript.

Funding

This work is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM133797. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. AS is supported by a T32 training grant from the National Human Genome Research Institute [HG000045 to A.S.]. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data, and in writing the manuscript.

Availability of data and materials

All data generated during this study are included in this published article and its supplementary information files. Datasets including sequences, squared variances, and plasmid and strain information is available at the Figshare database. <https://figshare.com/s/51c4820b2bee85c94007>

Declarations**Ethics approval and consent to participate**

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Energy, Environmental and Chemical Engineering, Washington University in St. Louis, Saint Louis, MO 63130, USA. ²Division of Biological & Biomedical Sciences, Washington University in St. Louis, Saint Louis, MO 63130, USA. ³Institute of Materials Science & Engineering, Washington University in St. Louis, Saint Louis, MO 63130, USA.

Received: 11 September 2020 Accepted: 22 February 2021

Published online: 02 March 2021

References

- Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. *Nature*. 2010;467:167–73.
- Ackermann M. A functional perspective on phenotypic heterogeneity in microorganisms. *Nat Rev Microbiol*. 2015;13:497–508.
- Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. *Science*. 2002;297:1183–6.
- Raser JM, O'Shea EK. Noise in gene expression: origins, consequences, and control. *Science*. 2005;309:2010–3.
- Han Y, Zhang F. Heterogeneity coordinates bacterial multi-gene expression in single cells. *PLoS Comput Biol*. 2020;16:1–17.
- Martins BM, Locke JC. Microbial individuality: how single-cell heterogeneity enables population level strategies. *Curr Opin Microbiol*. 2015;24:104–12.
- Delvigne F, Goffin P. Microbial heterogeneity affects bioprocess robustness: dynamic single-cell analysis contributes to understanding of microbial populations. *Biotechnol J*. 2014;9:61–72.
- Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*. 2008;135:216–26.
- Xiao Y, Bowen CH, Liu D, Zhang F. Exploiting nongenetic cell-to-cell variation for enhanced biosynthesis. *Nat Chem Biol*. 2016;12:339–44.
- Schmitz AC, Hartline CJ, Zhang F. Engineering microbial metabolite dynamics and heterogeneity. *Biotechnol J*. 2017;12:1700422.
- Guimaraes JC, Rocha M, Arkin AP. Transcript level and sequence determinants of protein abundance and noise in *Escherichia coli*. *Nucleic Acids Res*. 2014;42:4791–9.
- Sherman MS, Lorenz K, Lanier MH, Cohen BA. Cell-to-cell variability in the propensity to transcribe explains correlated fluctuations in gene expression. *Cell Syst*. 2015;1:315.
- Jones DL, Brewster RC, Phillips R. Promoter architecture dictates cell-to-cell variability in gene expression. *Science*. 2014;346:1533–6.
- Tuller T, Waldman YY, Kupiec M, Ruppert E. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci*. 2010;107:3645–50.
- Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*. 2009;324:255–8.
- Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*. 2007;25:117–24.
- Taniguchi Y, Choi PJ, Li G-W, Chen H, Babu M, Hearn J, et al. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*. 2010;329:533–8.
- Zhou Z, Dang Y, Zhou M, Li L, Yu C-H, Fu J, et al. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci*. 2016;113:E6117–25.
- Srinivasan S, Cluett WR, Mahadevan R. Constructing kinetic models of metabolism at genome-scales: a review. *Biotechnol J*. 2015;10:1345–59.
- Hanson G, Collier J. Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol*. 2017;19:20–30.
- Quax TEF, Claessens NJ, Söll D, van der Oost J. Codon Bias as a means to fine-tune gene expression. *Mol Cell*. 2015;59:149–61.
- Boël G, Letso R, Neely H, Price WN, Wong KH, Su M, et al. Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature*. 2016;529:358–63.
- Li GW, Oh E, Weissman JS. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*. 2012;484:538–41.
- Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*. 2011;12:32–42.
- Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y. Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proc Natl Acad Sci*. 2018;115:E4940–9.
- Gorochowski TE, Ignatova Z, Bovenberg RAL, Roubos JA. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic Acids Res*. 2015;43:3022–32.
- Cambrey G, Guimaraes JC, Arkin AP. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*. *Nat Biotechnol*. 2018;36:1005.
- Roymondal U, Das S, Sahoo S. Predicting gene expression level from relative codon usage bias: An application to *Escherichia coli* genome. *DNA Res*. 2009;16:13–30.
- Zhang G, Hubalewska M, Ignatova Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat Struct Mol Biol*. 2009;16:274–80.
- Yu C-H, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, et al. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Mol Cell*. 2015;59:744–54.
- Bentele K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N. Efficient translation initiation dictates codon usage at gene start. *Mol Syst Biol*. 2014;9:675.
- Burgess-Brown NA, Sharma S, Sobott F, Loenarz C, Oppermann U, Gileadi O. Codon optimization can improve expression of human genes in *Escherichia coli*: a multi-gene study. *Protein Expr Purif*. 2008;59:94–102.
- Peterman N, Levine E. Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics*. 2016;17:206.
- Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol*. 2012;30:521–30.
- Kosuri S, Goodman DB, Cambrey G, Mutalik VK, Gao Y, Arkin AP, et al. Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc Natl Acad Sci U S A*. 2013;110:14024–9.
- Sharp PM, Li W-H. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*. 1987;15:1281–95.
- Reis M d. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res*. 2004;32:5036–44.
- Pechmann S, Frydman J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat Struct Mol Biol*. 2012;20:237–43.
- Goodman DB, Church GM, Kosuri S. Causes and effects of N-terminal codon bias in bacterial genes. *Science*. 2013;342:475–9.
- Zadeh J, Steenberg C, Bois J, Wolfe B, Pierce M, Khan A, et al. Software news and updates NUPACK: analysis and Design of Nucleic Acid Systems. *J Comput Chem*. 2011;32:170–3.
- Mao Y, Liu H, Liu Y, Tao S. Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in *Saccharomyces cerevisiae*. *Nucleic Acids Res*. 2014;42:4813–22.
- Consortium TU. UniProt : a worldwide hub of protein knowledge; 2019;47 November 2018. p. 506–15.
- Carbone A, Zinovyev A, Kepes F. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics*. 2003;19:2005–15.
- Lo SS, Kongstad M, Stenum TS, Muñoz-Gómez AJ, Sørensen MA. Transfer RNA is highly unstable during early amino acid starvation in *Escherichia coli*. *Nucleic Acids Res*. 2017;45:793–804.
- Yang S, Kim S, Rim Lim Y, Kim C, An HJ, Kim JH, et al. Contribution of RNA polymerase concentration variation to protein expression noise. *Nat Commun*. 2014;5:1–9.

46. Liu D, Mannan AA, Han Y, Oyarzún DA, Zhang F. Dynamic metabolic control: towards precision engineering of metabolism. *J Ind Microbiol Biotechnol.* 2018;45:535–43.
47. Engler C, Kandzia R, Marillonnet S. A one pot, one step, Precision Cloning Method with High Throughput Capability. *PLoS One.* 2008;3:3647.
48. Lee TS, Krupa RA, Zhang F, Hajimorad M, Holtz WJ, Prasad N, et al. BglBrick vectors and datasheets: a synthetic biology platform for gene expression. *J Biol Eng.* 2011;5:12.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

