nature structural & molecular biology

ARTICLES

https://doi.org/10.1038/s41594-018-0080-2

Accurate design of translational output by a neural network model of ribosome distribution

Robert Tunney[1,6], Nicholas J. McGlincy[2,3,6], Monica E. Graham[3], Nicki Naddaf[3], Lior Pachter[2,4,5] and Liana F. Lareau[3*]

Synonymous codon choice can have dramatic effects on ribosome speed and protein expression. Ribosome profiling experiments have underscored that ribosomes do not move uniformly along mRNAs. Here, we have modeled this variation in translation elongation by using a feed-forward neural network to predict the ribosome density at each codon as a function of its sequence neighborhood. Our approach revealed sequence features affecting translation elongation and characterized large technical biases in ribosome profiling. We applied our model to design synonymous variants of a fluorescent protein spanning the range of translation speeds predicted with our model. Levels of the fluorescent protein in budding yeast closely tracked the predicted translation speeds across their full range. We therefore demonstrate that our model captures information determining translation dynamics in vivo; that this information can be harnessed to design coding sequences; and that control of translation elongation alone is sufficient to produce large quantitative differences in protein output.

As the ribosome moves along a transcript, it encounters diverse codons, tRNAs, and amino acids. This diversity affects translation elongation and ultimately gene expression. For instance, exogenous gene expression can be severely hampered by a mismatch between the choice of synonymous codons and the availability of tRNAs. The consequences of endogenous variation in codon use have been more elusive, but new methods have revealed that synonymous coding mutations, upregulation of tRNAs, and mutations within tRNAs can have dramatic effects on protein expression, folding, and stability[1–3]. Codon usage can directly affect the speed of translation elongation[4]. However, translation initiation has been considered the rate-limiting step in translation, thus implying that changes in elongation speed should have limited effects[5]. Recent work has suggested a relationship between codon use and RNA stability: slower translation may destabilize mRNAs and thus decrease protein expression[6,7]. Because these opposing viewpoints have yet to be fully reconciled, the definition of a favorable sequence for translation remains unsettled.

The advent of high-throughput methods to measure translation elongation in vivo has elucidated the functional implications of codon usage. Ribosome profiling measures translation across an entire transcriptome by capturing and sequencing the regions of mRNA protected within ribosomes, called ribosome footprints[8]. Each footprint reflects the position of an individual ribosome on a transcript, and the aminoacyl (A)-site codon—the site of tRNA decoding—can be reliably inferred in each footprint (Fig. 1a). This codon-level resolution yields the distribution of ribosomes along mRNAs from each gene. The counts of footprints on each codon can be used to infer translation-elongation rates: slowly translated codons yield more footprints, and quickly translated codons yield fewer footprints (Fig. 1b). Analyses of ribosome profiling data have shown a relationship between translation-elongation rate and biochemical features such as tRNA abundance, wobble base-pairing, amino acid polarity, and mRNA structure[9–18]. Expanded probabilistic and machine-learning models have shown that the sequence context of a ribosome contributes to its elongation rate, both directly and through higher-order features such as nascent protein sequence[15–17,19]. Computational modeling has also indicated that technical artifacts and biases contribute to the distribution of ribosome footprints[18–21]. However, distinguishing experimental artifacts from the biological determinants of elongation rate remains a challenge. Here, we used neural networks to model ribosome distribution along transcripts. The model captures both biological variation in translation-elongation speed and technical biases affecting footprint count, which we confirmed experimentally. We implemented a tool, Iχnos, that applies our model to design coding sequences, and we used this tool to design sequences spanning a range of predicted translation-elongation speeds. We found that the predicted elongation speeds accurately tracked protein expression, thus supporting a role of the elongation phase of translation in modulating gene expression.

Results

Design and performance of a neural network model of translation elongation. First, we developed a regression framework to model the distribution of ribosomes along transcripts as a function of local sequence features. As our measure of ribosome density on individual codon positions, we calculated scaled footprint counts by dividing the raw footprint count at each codon position by the average footprint count on its transcript (Fig. 1b). This normalization controls for variable mRNA abundance and translation-initiation rates across transcripts. The scaled count thus reflects the relative speed of translation elongation at each position. We used a sequence neighborhood around the A site as the predictive region for scaled counts, and we encoded this neighborhood as input to a regression model via one-hot encoding of the codons and nucleotides in this

[1]Graduate Group in Computational Biology, University of California, Berkeley, Berkeley, CA, USA. [2]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. [3]California Institute for Quantitative Biosciences, University of California, Berkeley, Berkeley, CA, USA. [4]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. [5]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA. [6]These authors contributed equally: Robert Tunney, Nicholas J. McGlincy.
*e-mail: lareau@berkeley.edu

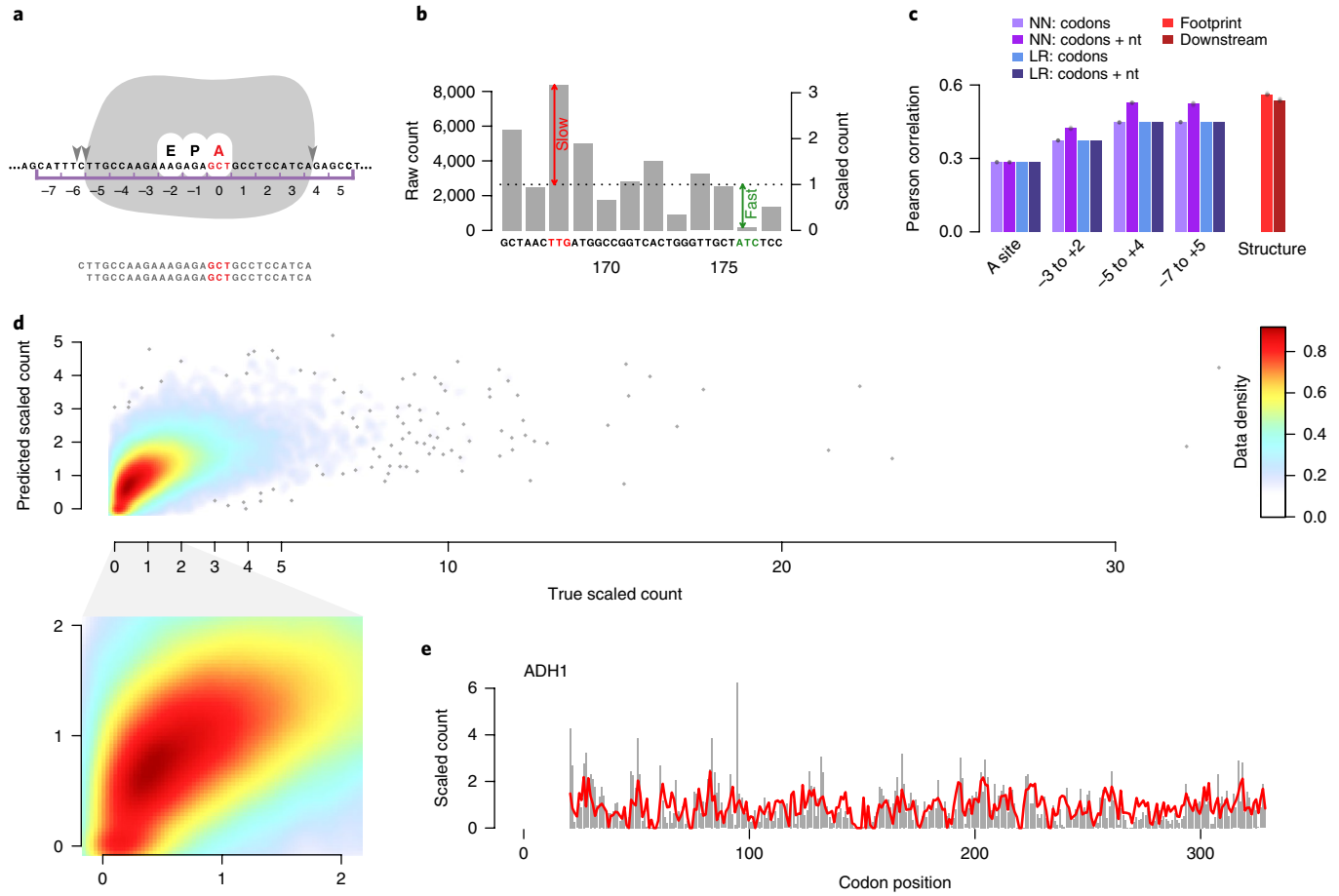

Fig. 1 | Design and performance of a neural network model of translation elongation. a, Each ribosome protects an mRNA footprint of approximately 28 or 29 nt. Sequence coordinates in a neighborhood around a ribosome are indexed relative to the codon in the A site of the ribosome. **b**, Read-count rescaling. For each gene, the counts of footprints assigned to each A-site codon are divided by the average counts per codon over that gene. The resulting scaled footprint counts are used for model training and prediction. **c**, Model performances (Pearson correlations between predicted and true scaled counts over the test set) for neural network (NN) and linear regression (LR) models over a range of sequence neighborhoods, with and without nucleotide features, as well as correlations for models that also incorporate structure scores of the three 30-nt windows overlapping the footprint region or the maximum structure score within 59 nt downstream of the ribosome. Bars show the mean of ten runs of each model; the ten individual runs for each model are overlaid as gray points. **d**, True versus predicted scaled counts for the test set, under a model with codon and nucleotide features spanning codon positions −5 to +4. Color scale shows density of data points. **e**, True scaled counts (gray bars) and predicted scaled counts (red line) for a highly translated gene.

region (Supplementary Fig. 1). Then we learned a regression function with a feed-forward neural network trained on a large, high-quality ribosome profiling dataset from *Saccharomyces cerevisiae*[22]. We chose the top 500 genes on the basis of footprint density and coverage criteria and sorted them into training and test sets of 333 and 167 genes, respectively.

We determined the sequence neighborhood that best predicted ribosome density by comparing a series of models ranging from an A-site-only model to a model spanning codon positions −7 to +5 (Fig. 1c). The identity of the A-site codon was an important but limited predictor of the distribution of ribosome footprints (Pearson's $r = 0.28$). Expanding the sequence context around the A site steadily improved the predictive performance, up to the full span of a ribosome footprint (codons −5 to +4). Additional sequence context beyond the boundaries of the ribosome did not improve performance. We also observed a large boost in predictive performance by including redundant nucleotide features in addition to codon features over the same sequence neighborhood, especially near the ends of the ribosome footprint (Fig. 1c; $r = 0.53$ for the −5 to +4 model including nucleotide features, $\Delta r = 0.08$ relative to the no-nucleotide model). Linear regression models that included only codon features performed similarly to the neural networks tested, but they did not improve with the inclusion of nucleotide features. This result suggests that the neural network models learn a meaningful and nonlinear predictive relationship in nucleotide features, particularly toward the flanking ends of footprints, thus making them more successful than linear models.

Next, we assessed the contribution of local mRNA structure to footprint distributions. We computed mRNA folding energies in sliding 30-nt windows over all transcripts, then trained a series of models that each included one window from nucleotide positions −45 to +72 relative to the A site. The model performance improved after inclusion of structure scores at nucleotide positions −17, −16, and −15, i.e., the windows that span the actual ribosome footprint ($\Delta r = 0.03$; Fig. 1c and Supplementary Fig. 2). No individual windows downstream of the footprint improved our predictions, and the maximum structure score over 30 sliding windows downstream of the ribosome had only a slight effect ($\Delta r < 0.01$; Fig. 1c). Thus, our approach does not capture a conclusive effect of downstream mRNA structure on elongation rate. Because we

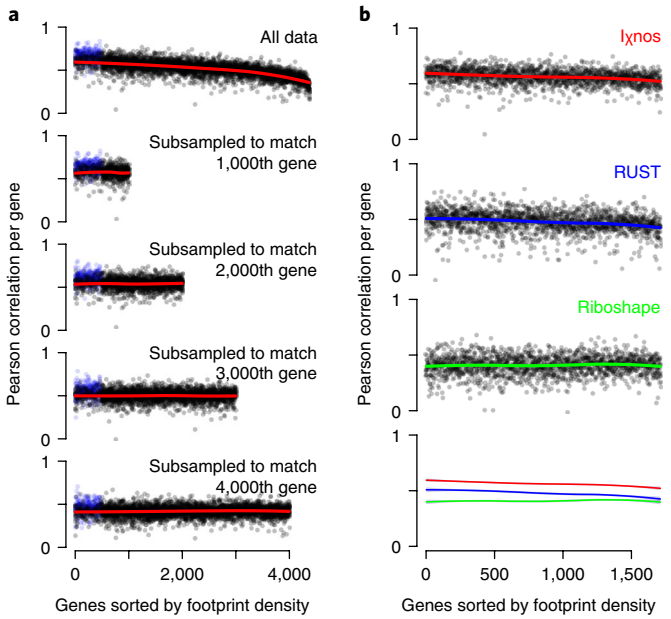

Fig. 2 | Performance comparisons on low-coverage genes and with competing models. a, Top, per-gene correlations between true and predicted scaled counts, for all 4,375 genes in our transcriptome that passed filtering criteria. Training set genes in blue (333/top 500 genes by footprint density). Loess curve on test-set genes are shown in red. Below, as above, with footprint counts on the top 1,000, 2,000, 3,000, and 4,000 genes subsampled to the density of footprint counts on the one thousandth, two thousandth, three thousandth, and four thousandth gene, respectively, and 'true' scaled counts recomputed. **b**, Comparison of Iχnos with the similar models RUST[19] and riboshape[15]. Shown are per-gene correlations between true and predicted scaled counts, on 1,711 genes passing the filtering criteria from all three methods. Training-set genes from Iχnos are excluded. Colored lines are loess curves, which are also compared at bottom.

unexpectedly observed an effect of structure within the ribosome, we tested the direction of the effect and found that more structure in these windows led to lower predicted footprint counts. This finding suggests that stable mRNA structure in the footprint fragments themselves inhibits their in vitro recovery in ribosome profiling experiments, and our model captures the bias consequently introduced to the data.

Our best model incorporated a sequence window from codons −5 to +4, represented as both codons and nucleotides, as well as structural features of the three windows spanning the footprint. It captured sufficient information to accurately predict footprint distributions on individual genes (Fig. 1e), and it yielded a correlation of 0.57 (Pearson's r) between predicted and true scaled counts over all positions in the test set (Fig. 1d). Although our model performed well across a range of scaled counts, it had difficulty in predicting very high scaled footprint counts at a small number of sites. These sites may represent ribosome stalling determined by biological factors encoded outside of this local sequence neighborhood[16].

Our model was trained on highly expressed genes, because abundant ribosome footprints enable more accurate sampling of ribosome positions. However, highly expressed genes can have biased codon usage[23]. To ensure that our model accurately predicts translation on genes across the full range of expression and codon usage, we computed the correlation between the observed and predicted scaled counts for all yeast genes. Performance decreased with lower expression (Fig. 2a), but we hypothesized that the decreased performance reflected noisier observed footprint counts arising from less abundant mRNAs rather than differences in codon composition. To test this possibility, we downsampled the footprints for each of the 1,000 highest-expression genes to match the average counts per codon of the thousandth gene, then repeated this procedure for the top 2,000, 3,000, and 4,000 genes. We then compared the predictions of our model, which had been trained on the full data from highly expressed genes, against the downsampled data. At each coverage level, our method performed equally well on high- and low-expression genes. Thus, our model has no decrease in performance on genes that tend to have less favored codon content, after controlling for data density.

We also compared the performance of our model against two earlier approaches that incorporate information from the sequence neighborhood of each codon to predict ribosome distributions: RUST, which computes the expected ribosome density at each codon, using the sequence window around that codon as input[19], and riboshape, which uses wavelet decomposition to denoise the observed counts by projecting them into different subspaces at different levels of resolution (smoothness), then predicts ribosome density after transformation into these subspaces[15]. To compare riboshape to our own method and to RUST, we evaluated how well its predictions in the highest-resolution subspace (i.e., closest to the raw data) correlated with the observed footprint counts. Our model outperformed both models, with an average Pearson correlation per gene of 0.56 versus 0.48 (RUST) and 0.41 (riboshape) across all genes included in all three analyses (Fig. 2b). We also found that our predictions of the raw data were better than riboshape's predictions of the transformed data at each resolution (Supplementary Table 1).

Sequences near the A site and at the ends of footprints contribute to footprint density. To quantify the influence of distinct positions in the sequence neighborhood on elongation rate, we trained a series of leave-one-out models that excluded individual codon positions from the input sequence neighborhood, then compared their performance to a reference model that included all positions. We found that the A-site codon contributed the most to predictive performance ($\Delta r = 0.13$), but we also observed contributions from the surrounding sequence context, including the peptidyl (P) and exit (E) sites ($\Delta r = 0.03$ and 0.03) (Fig. 3a). Each codon position from −5 to +4, the span of a typical 28-nt ribosome footprint, improved the performance of the full model, whereas positions outside the span of a footprint decreased performance. Contributions from the E and P sites suggested that the continued presence of tRNAs at these positions modulates elongation rate. In contrast, the large contribution of the +3 codon ($\Delta r = 0.06$), at the 3′ end of the footprint, probably reflects artifactual biases arising from the ribosome profiling process, thus corroborating previous reports of fragment end biases[19,20].

We were also interested in understanding the relative influence of the A-site codon and its immediate environment. Overall, the A-site codon and its immediate environment predicted ribosome density similarly well (Pearson's $r = 0.28$ for the A site only; $r = 0.26$ for the codons from −3 to +2 excluding the A site). To identify A-site codons that tended to dominate the prediction by contributing relatively more than their context, we compared the performance of a −3 to +2 model and a model with codons −3 to +2 but excluding the A site (Supplementary Fig. 3). We found that the presence of lysine codons AAA and AAG in the A site led to the strongest predictions, in agreement with a major effect of charged lysine residues on translation[11]. In contrast, we also identified several sequence contexts that tended to dominate the prediction, by examining the sequence contexts of the positions with higher squared error arising from the A-site-only prediction at that position than the no-A-site context (Supplementary Fig. 3).

Next, we examined what our model had learned about the relationship between sequence and ribosome density. The raw parameters of a neural network can be difficult to interpret, so we

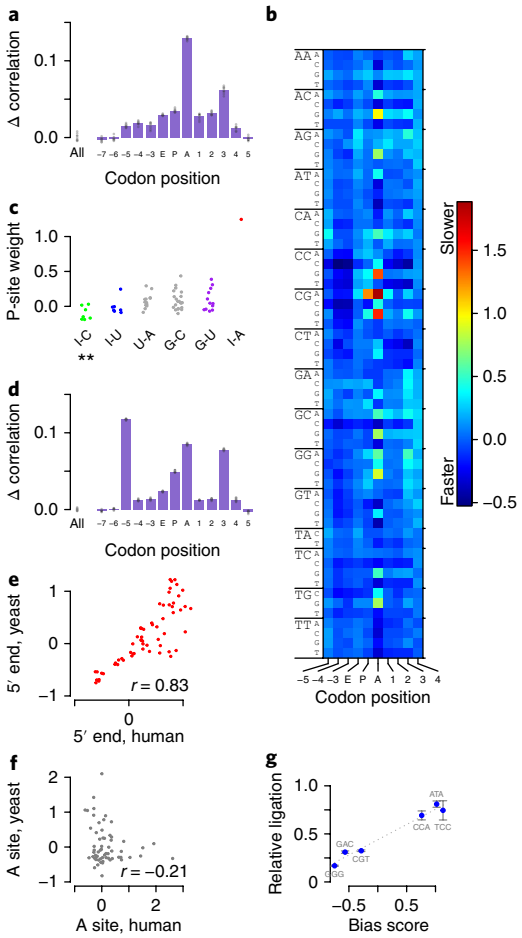

Fig. 3 | Interpretation of models of translation-elongation rates. **a**, Predictive value of codon positions in a yeast ribosome profiling dataset[22]. We computed Pearson correlations between true and predicted scaled counts on the test set, for a reference model including codon and nucleotide features from codon positions -7 to +5, and for a series of leave-one-out models, each excluding one codon position. Gray points show differences between Pearson's *r* for ten runs of each leave-one-out model and the mean *r* of ten runs of the reference model. Bars represent the means of these values. **b**, Mean contributions to scaled counts by codon identity and position. Codon distributions and weights learned for each codon in positions −5 to +4 are shown in Supplementary Table 3. **c**, P-site codon contributions grouped by the codon-anticodon base pair formed by the third nucleotide of each codon. Asterisks indicate $P < 0.05$ after Bonferroni correction, unpaired two-sided Mann–Whitney *U* test between each group and all other codons. I-C, $P = 0.014$. **d**, Predictive value of codon positions as in **a**, from a yeast ribosome profiling library constructed with CircLigase II, as described in ref. [29]. **e**,**f**, Contributions from codon position −5, at the 5′ ends of footprints (**e**) and the A site in human ribosome profiling data[30] (**f**) versus our yeast ribosome profiling data, both constructed with CircLigaseII. Analysis was limited to 28-nt footprints to avoid frame biases. **g**, Ligation efficiency of CircLigase II. Oligonucleotide substrates resembling ribosome footprints at the circularization step of the protocol, with different 3-nt end sequences, were ligated by both enzymes. Circularization was assayed by qPCR with primers spanning the ligation, as compared with primers in a contiguous region of the oligonucleotide. Ligation was calculated relative to CircLigase I ligation of the best-ligated substrate. Each point represents the ratio of the means of three qPCR replicates; error bars represent the standard error of that ratio.

determined a score for each codon at each position by computing the mean increase in predicted scaled counts due to that codon (Fig. 3b and Supplementary Table 3). The time spent finding the correct tRNA is considered to be a main driver of elongation speed and consequently of footprint counts[24]. Indeed, the A-site-codon scores exhibited the widest range, and scores at this position but not other positions correlated with the tRNA Adaptation Index (tAI), a measure of tRNA availability[25], as has been widely observed (Pearson's $r = 0.50$; $P = 0.0005$ after Bonferroni correction). Our results highlighted the well-characterized slow translation of CCG (proline), CGA (arginine), and CGG (arginine) codons at the A site[26]. Our data also underscore that sequences in the P site contribute to elongation speed. The CGA codon showed a particularly strong inhibitory effect in the P site, in agreement with recent results[26,27]. We noted that this codon forms a disfavored I-A wobble pair with its cognate tRNA, thus distorting the anticodon loop[28], whereas the four fastest P-site codons all form I-C wobble pairs (Fig. 3c). Overall, I-C base pairs in the P site contributed to faster translation (Mann–Whitney $P = 0.014$ after Bonferroni correction; Fig. 3c). From these results, we concluded that the conformation of the tRNA–mRNA duplex can influence its passage through the ribosome, not just initial recognition in the A site.

We also observed strong sequence preferences at the 3′ ends of ribosome footprints. Sequence bias has previously been noted in the 5′ and 3′ ends of ribosome footprints, and this bias has been suggested to arise from ligase preferences during library preparation[19,20]. To compare features of ribosome profiling data generated in different experiments, we applied our model to a large ribosome profiling dataset that we generated from yeast through a standard ribosome profiling protocol[29]. Models trained on these data learned disconcertingly high weights for both the −5 and +3 codon positions (Fig. 3d). The −5 codon, i.e., the 5′ end of a footprint, was the single strongest predictor of footprint counts, exceeding even those of the A site. We found similarly large 5′-end contributions in published yeast and human datasets generated through similar protocols[30,31] (Supplementary Fig. 4). These experiments, like our own, made use of CircLigase enzymes to circularize ribosome footprints after reverse transcription. In contrast, the experiment that we first modeled used T4 RNA ligase to attach 5′ linkers directly onto ribosome-footprint fragments[22]. To compare end-sequence preferences between experiments, we trained models on only 28-nt footprints so that the ends of the footprints corresponded to the −5 codon position. Comparing the T4 ligase yeast data with CircLigase yeast data[31], we observed no relationship between the scores learned at 5′ footprint ends ($r = 0.05$) but a high correlation between scores at the A site, where we would expect biological similarity ($r = 0.86$). In contrast, we observed a high correlation at the −5 position between our CircLigase yeast data and the CircLigase-generated human dataset[30] ($r = 0.83$, Fig. 3e) but no significant relationship at the A site, where we would expect species-specific codon bias ($r = -0.21$, $P = 0.11$; Fig. 3f). This result suggested that the fragment end scores reflected experimental artifacts rather than in vivo biology.

To directly test the effects of enzyme biases on recovery of ribosome-protected fragments, we experimentally measured the ligation of synthetic oligonucleotides with end sequences shown to be favored or disfavored in our model. The relative ligation efficiency of each substrate closely mirrored the end-sequence scores learned by our model for both CircLigase I and CircLigase II (Fig. 3g and Supplementary Fig. 5). The least favored sequences were ligated by CircLigase II with only 20% the efficiency of the most favored sequences, thus indicating that some ribosome footprints would be represented at five times the frequency of other footprints for purely technical reasons. This biased recovery of fragments might skew the results of ribosome profiling experiments, thereby affecting estimates of elongation and overall per-gene translation.

Expression of synonymous reporters closely tracks predicted translation speeds. Our model captured the quantitative preferences

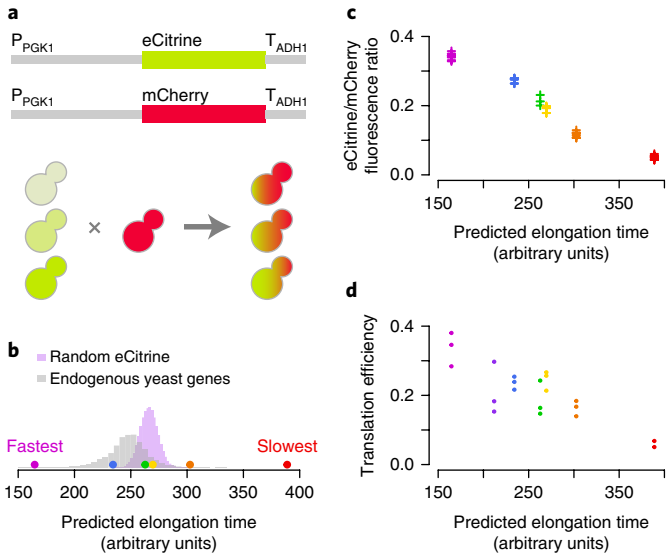

Fig. 4 | Design of synonymous sequences showing that elongation rate affects translation output. a, Six reporter constructs with distinct synonymous eCitrine coding sequences were inserted into the *his3Δ1* locus of BY4742 yeast, and an equivalent construct with a constant mCherry coding sequence was inserted into the *his3Δ1* locus of BY4741 yeast. The haploids were mated to produce diploid yeast with both reporters, whose fluorescence was then measured with flow cytometry. P_{PGK1}, PGK1 promoter; T_{ADH1}, ADH1 terminator. **b**, The synonymous eCitrine sequences included the fastest and slowest predicted sequences under our model (magenta and red), plus sequences with predicted translation-elongation times at percentiles 0, 33, 67, and 100 of a randomly generated set of 100,000 synonymous eCitrine sequences (blue, green, yellow, and orange, respectively). The score distribution of 100,000 random eCitrine sequences is shown in lavender. The scores of endogenous yeast genes, rescaled by length to compare with eCitrine, are shown in gray. **c**, eCitrine/mCherry fluorescence ratio, as measured by flow cytometry of 11,000–18,000 yeast cells, versus the predicted elongation time of each sequence. Each plus symbol represents the median ratio of yellow and red fluorescence from one biological replicate of the given eCitrine strain. Eight biological replicates, each an independent integration of the reporter construct, are included for each strain, except for the strains shown in blue and orange, which have seven, and the strain shown in green, which has three. Colors are as in **b**. **d**, Translation efficiency, or median eCitrine/mCherry fluorescence ratio divided by the relative eCitrine/mCherry mRNA ratio (ratio of medians of three qPCR replicates) for each eCitrine variant, versus the predicted elongation time of each sequence. Purple, yECitrine sequence; other colors are as in **b**. Each point represents one biological replicate of the given eCitrine strain; three biological replicates were measured for each strain, except for the strain shown in red, for which two biological replicates were measured.

of ligases for footprint end sequences and established that a substantial portion of the predictive information of these end regions is due to technical artifacts. However, the biologically sensible weights learned for codons in the A site showed that the model captured substantial biology as well. We reasoned that if our model captures biological aspects of translation elongation, we could use the parameters learned by the model to design sequences that would be translated at different rates. To focus on the biological contributions and reduce the influence of biases from the ends of footprints, we relied on the information found in the codons closer to the A site (discussed further in Supplementary Note 1).

To test our model's ability to predict translation, we expressed synonymous variants of the yellow fluorescent protein eCitrine in yeast (Fig. 4a). First, using the yeast ribosome profiling data from Weinberg et al.[22], we trained a neural network model with a sequence neighborhood extending from codon positions −3 to +2. Next, we designed a dynamic programming algorithm to compute the maximum- and minimum-translation-time synonymous versions of eCitrine according to our model. We defined the overall translation time (in arbitrary units) of a gene as the sum of the predicted scaled counts over all codons in the gene. We also generated and scored a set of 100,000 random synonymous eCitrine coding sequences and selected the sequences at percentiles 0, 33, 67, and 99 of predicted translation time within that set (Fig. 4b). We used flow cytometry to measure the fluorescence of diploid yeast, each containing an eCitrine variant along with the red fluorescent protein mCherry as a control, and calculated the relative fluorescence of each variant (Fig. 4c and Supplementary Fig. 6).

The expression of eCitrine in each yeast strain closely tracked its predicted elongation rate, with the predicted fastest sequence producing sixfold-higher fluorescence than the predicted slowest sequence (Fig. 4c). However, the existing yeast-optimized yECitrine sequence[32] produced threefold-higher fluorescence than our predicted fastest sequence (Supplementary Fig. 7). To understand the source of this discrepancy, we measured eCitrine mRNA from all strains and found that sequences designed by our method had approximately equivalent mRNA levels, whereas yECitrine had fivefold more mRNA (Supplementary Fig. 7). Calculating translation efficiencies, or protein produced per mRNA, reconciled this lack of agreement. We observed a clear linear relationship between predicted elongation rate and translation efficiency (Fig. 4d).

Discussion

These experiments demonstrate that our model is able to predict large quantitative differences in protein production, by using only information about translation elongation. The sequences that we designed and tested have predicted translation speeds that span the range of natural yeast genes (Fig. 4b). This result supports an effect of elongation rate on the translation efficiency and protein output of endogenous genes. Initiation rather than elongation is usually thought to be rate limiting for protein production of most endogenous genes[5,24]. Models have suggested that highly expressed transgenes might deplete the effective supply of ribosomes, thereby lowering initiation and causing elongation to be rate limiting; however, our reporter is expressed at the level of many endogenous genes and should represent well under 1% of mRNA. How translation speed can control translation efficiency remains to be determined. One contribution could come from pileups behind stalled or slow-moving ribosomes, thus diminishing the maximum throughput of protein production[17]. In particular, codon choice near the beginning of a gene, affecting elongation speed, can interfere with translation initiation and therefore control protein output[33]. Although codon choice can also affect mRNA stability and thus total protein output[6,7], our fast and slow predicted sequences have equivalent steady-state mRNA. Further, an effect arising purely from mRNA stability would affect protein output but not translation efficiency, counter to our observations. Instead, our results indicate that optimized elongation rates do result in more protein per mRNA, and this effect does not depend entirely on mRNA stability. The landscape of factors affecting codon optimality is complex[34], and codon preferences vary across species, tissues, and conditions. Our approach can capture empirical information about codon preferences in any system in which translation can be measured by ribosome profiling and can apply this information to the design of sequences for quantitative expression in that system.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41594-018-0080-2.

Received: 21 November 2017; Accepted: 11 May 2018;
Published online: 2 July 2018

References

1. Ishimura, R. et al. Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. *Science* **345**, 455–459 (2014).
2. Goodarzi, H. et al. Modulated expression of specific tRNAs drives gene expression and cancer progression. *Cell* **165**, 1416–1427 (2016).
3. Kirchner, S. et al. Alteration of protein function by a silent polymorphism linked to tRNA abundance. *PLoS Biol.* **15**, e2000779 (2017).
4. Zhao, F., Yu, C.-H. & Liu, Y. Codon usage regulates protein structure and function by affecting translation elongation speed in *Drosophila* cells. *Nucleic Acids Res.* **45**, 8484–8492 (2017).
5. Shah, P., Ding, Y., Niemczyk, M., Kudla, G. & Plotkin, J. B. Rate-limiting steps in yeast protein translation. *Cell* **153**, 1589–1601 (2013).
6. Presnyak, V. et al. Codon optimality is a major determinant of mRNA stability. *Cell* **160**, 1111–1124 (2015).
7. Bazzini, A. A. et al. Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *EMBO J.* **35**, 2087–2103 (2016).
8. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
9. Stadler, M. & Fire, A. Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA* **17**, 2063–2073 (2011).
10. Dana, A. & Tuller, T. Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. *PLOS Comput. Biol.* **8**, e1002755 (2012).
11. Charneski, C. A. & Hurst, L. D. Positively charged residues are the major determinants of ribosomal velocity. *PLoS Biol.* **11**, e1001508 (2013).
12. Gardin, J. et al. Measurement of average decoding rates of the 61 sense codons in vivo. *eLife* **3**, e03735 (2014).
13. Lareau, L. F., Hite, D. H., Hogan, G. J. & Brown, P. O. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *eLife* **3**, e01257 (2014).
14. Pop, C. et al. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol. Syst. Biol.* **10**, 770 (2014).
15. Liu, T.-Y. & Song, Y. S. Prediction of ribosome footprint profile shapes from transcript sequences. *Bioinformatics* **32**, i183–i191 (2016).
16. Zhang, S. et al. Analysis of ribosome stalling and translation elongation dynamics by deep learning. *Cell Syst.* **5**, 212–220.e6 (2017).
17. Dao Duc, K. & Song, Y. S. The impact of ribosomal interference, codon usage, and exit tunnel interactions on translation elongation rate variation. *PLoS Genet.* **14**, e1007166 (2018).
18. Fang, H. et al. Scikit-ribo enables accurate estimation and robust modeling of translation dynamics at codon resolution. *Cell Syst.* **6**, 180–191.e4 (2018).
19. O'Connor, P. B. F., Andreev, D. E. & Baranov, P. V. Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. *Nat. Commun.* **7**, 12915 (2016).
20. Artieri, C. G. & Fraser, H. B. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome Res.* **24**, 2011–2021 (2014).
21. Hussmann, J. A., Patchett, S., Johnson, A., Sawyer, S. & Press, W. H. Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. *PLoS Genet.* **11**, e1005732 (2015).
22. Weinberg, D. E. et al. Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell Rep.* **14**, 1787–1799 (2016).
23. Sharp, P. M., Tuohy, T. M. & Mosurski, K. R. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* **14**, 5125–5143 (1986).
24. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12**, 32–42 (2011).
25. dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**, 5036–5044 (2004).
26. Letzring, D. P., Dean, K. M. & Grayhack, E. J. Control of translation efficiency in yeast by codon-anticodon interactions. *RNA* **16**, 2516–2528 (2010).
27. Gamble, C. E., Brule, C. E., Dean, K. M., Fields, S. & Grayhack, E. J. Adjacent codons act in concert to modulate translation efficiency in yeast. *Cell* **166**, 679–690 (2016).
28. Murphy, F. V. IV & Ramakrishnan, V. Structure of a purine-purine wobble base pair in the decoding center of the ribosome. *Nat. Struct. Mol. Biol.* **11**, 1251–1252 (2004).
29. McGlincy, N. J. & Ingolia, N. T. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**, 112–129 (2017).
30. Iwasaki, S., Floor, S. N. & Ingolia, N. T. Rocaglates convert DEAD-box protein eIF4A into a sequence-selective translational repressor. *Nature* **534**, 558–561 (2016).
31. Schuller, A. P., Wu, C. C.-C., Dever, T. E., Buskirk, A. R. & Green, R. eIF5A functions globally in translation elongation and termination. *Mol. Cell* **66**, 194–205.e5 (2017).
32. Sheff, M. A. & Thorn, K. S. Optimized cassettes for fluorescent protein tagging in *Saccharomyces cerevisiae*. *Yeast* **21**, 661–670 (2004).
33. Chu, D. et al. Translation elongation can control translation initiation on eukaryotic mRNAs. *EMBO J.* **33**, 21–34 (2014).
34. Qian, W., Yang, J.-R., Pearson, N. M., Maclean, C. & Zhang, J. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet.* **8**, e1002603 (2012).

Acknowledgements

We are grateful to N. Ingolia and S. McCurdy for discussion. This work was supported by the National Cancer Institute of the National Institutes of Health, under award R21CA202960 to L.F.L., and by the National Institute of General Medical Sciences of the National Institutes of Health, under award P50GM102706 to the Berkeley Center for RNA Systems Biology. R.T. was supported by the Department of Defense through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. This work made use of the Vincent J. Coates Genomics Sequencing Laboratory at the University of California, Berkeley, supported by National Institutes of Health S10 Instrumentation grant OD018174, and the UC Berkeley flow cytometry core facilities.

Author contributions

L.F.L., R.T., and N.J.M. designed the study, with input from L.P. R.T. developed the software and performed modeling, and R.T., L.P., and L.F.L. analyzed and interpreted the computational results. N.J.M. designed and created the yeast strains and performed expression experiments, with assistance from M.E.G. and N.N. M.E.G. performed yeast ribosome profiling. N.J.M. and L.F.L. analyzed and interpreted the experimental data. R.T. and L.F.L. wrote the manuscript, with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41594-018-0080-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.F.L.

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

Methods

Ribosome profiling. Yeast ribosome profiling was performed exactly according to ref. [29] with the following modifications:

250 ml of YEPD medium was inoculated from an overnight culture of BY474 to an optical density at 600 nm (OD_{600}) of 0.1. Yeast were grown to mid-log phase and harvested at an OD_{600} of 0.565. Lysis proceeded according to ref. [29] except with no cycloheximide in the lysis buffer (20 mM Tris, pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT, 1% (vol/vol) Triton X-100, and 25 U/ml Turbo DNase I). To quantify the RNA content of the lysate, total RNA was purified from 200 µl of lysate with a Direct-zol RNA MiniPrep Kit (Zymo Research), and the concentration of RNA was measured with a NanoDrop 2000 spectrophotometer (Thermo Fisher).

Lysate containing 30 µg of total RNA was thawed on ice and diluted to 200 µl with polysome buffer with no cycloheximide (20 mM Tris, pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$, and 1 mM DTT). 0.1 µl (1 U) of RNase I (Epicentre) was added to the diluted cell lysate, which was then incubated at room temperature for 45 min. Digestion and monosome isolation proceeded according to ref. [29], except with no cycloheximide in the sucrose cushion.

Purified RNA was separated on a 15% TBE/urea gel, and fragments of 18–34 nt were gel extracted. Size was determined relative to RNA size markers NI-NI-800 and NI-NI-801 (ref. [29]) and NEB microRNA size marker (New England Biolabs). Library preparation proceeded according to ref. [29]. The library was made with downstream linker NI-NI-811 (/5Phos/NNNNNA GCTAAGATCGGAAGAGCACACGTCTGAA/3ddC/) and a modified RT primer with a preferred CircLigase II substrate (AG) at the 5′ end (oLFL075, 5′-/5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAG/iSp18/ GTGACTGGAGTTCAGACGTGTGCTC). Library-amplification PCR used primers NI-NI-798 and NI-NI-825 (Illumina index ACAGTG). The resulting library was sequenced as single-end 51-nt reads on an Illumina HiSeq4000 instrument according to the manufacturer's protocol by the Vincent J. Coates Genomics Sequencing Laboratory at the University of California, Berkeley.

Sequencing data processing and mapping. A custom yeast transcriptome file was generated on the basis of all chromosomal ORF coding sequences in orf_coding. fasta from the *Saccharomyces* Genome Database genome annotation R64-2-1 for reference genome version R64-1-1 (UCSC sacCer3) for *S. cerevisiae* strain S288C. A human transcriptome file was generated from GRCh38.p2, Gencode v. 22, to include one transcript per gene, on the basis of the ENSEMBL 'canonical transcript' tag. In both human and yeast transcriptome files, 13 nt of 5′-UTR sequence and 10 nt of 3′-UTR sequence were included to accommodate footprint reads from ribosomes at the first and last codons. For yeast transcripts with no annotated UTR, the flanking genomic sequence was included. For human transcripts with no annotated UTR, or UTRs shorter than 13 or 10 nt, the sequence was padded with N.

Yeast ribosome profiling reads from ref. [22] (SRR1049521) were trimmed to remove the ligated 3′ linker (TCGTATGCCGTCTTCTGCTTG) from any read that ended with any prefix of that string, and to remove eight random nucleotides at the 5′ end (added as part of the 5′ linker). Yeast ribosome profiling reads generated in our own experiments (GEO GSE106572) were trimmed to remove the ligated 3′ linker, which included five random nucleotides and a 5-nt index of AGCTA (NNNNNIIIIIAGATCGGAAGAGCACACGTCTGAAC). Human ribosome profiling reads from ref. [30] (SRR2075925 and SRR2075926) were trimmed to remove the ligated 3′ linker (CTGTAGGCACCATCAAT). Yeast ribosome profiling reads from ref. [31] (SRR5008134 and SRR5008135) were trimmed to remove the ligated 3′ linker (CTGTAGGCACCATCAAT).

Trimmed fastq sequences of longer than 10 nt were aligned to yeast or human ribosomal and noncoding RNA sequences with bowtie v. 1.2.1.1 (ref. [35]), with options 'bowtie -v 2 --S'. Reads that did not match rRNA or ncRNA were mapped to the transcriptome with options 'bowtie --a --norc -v 2 --S'. Mapping weights for multimapping reads were computed with RSEM v. 1.2.31 (ref. [36]).

Assignment of A sites. A-site codons were identified in each footprint with simple rules for the offset of the A site from the 5′ end of the footprint. These rules were based on the length of the footprint and the frame of the 5′ terminal nucleotide. For each dataset, the set of lengths that included appreciable footprint counts was determined (for example, Weinberg 27–31 nt). For each length, the counts of footprints mapping to each frame were computed. The canonical 28-nt footprint starts coherently in frame 0, with the 5′ end 15 nt upstream of the A site[8]. For all other lengths, rules were defined if footprints mapped primarily to one or two frames, and offsets were chosen to be consistent with overdigestion or underdigestion relative to a 28-nt footprint. Footprints mapping to other frames were discarded.

Scaled counts. For each codon, the raw footprint counts were computed by summing the RSEM mapping weights of each footprint with its A site at that codon. Scaled footprint counts were computed by dividing the raw counts at each codon by the average raw counts over all codons in its transcript. This procedure controlled for variable initiation rates and copy numbers across transcripts. The resulting scaled counts are mean-centered at 1, with scaled counts higher than 1 indicating slower-than-average translation. The first 20 and last 20 codons in each gene were excluded from all computations and datasets, to avoid the atypical footprint counts observed at the beginnings and ends of genes.

Genes were excluded from analysis if they had fewer than 200 raw footprint counts in the truncated coding sequences, or fewer than 100 codons with mapped footprints in this region. Because many highly expressed yeast genes have close paralogs, only one gene was retained from each set of nearly identical paralogs. Then the top 500 genes were selected according to footprint density (average footprint counts per codon). Two-thirds of these genes were selected at random as the training set, and the remaining one-third of genes were used as the test set.

Input features. The model accepts user-defined sets of codon and nucleotide positions around the A site to encode as input features for predicting ribosome density. The A site is indexed as codon 0, and its first nucleotide is indexed as nucleotide 0, with negative indices in the 5′ direction and positive indices in the 3′ direction. Each codon and nucleotide feature is converted to a binary vector via one-hot encoding, and these vectors are concatenated as input into the regression models. The model also accepts RNA folding energies from the RNAfold package and allows the user to define window sizes and positions to score RNA structure to be included as inputs into the regression models.

In our best-performing model, codons −5 to +4 and nucleotides −15 to + 14 were chosen, as well as folding energies from three 30-nt windows starting at nucleotides −17, −16, and −15.

Model construction. All models were constructed as feed-forward artificial neural networks with the Python packages Lasagne v. 0.2.dev1 (ref. [37]) and Theano v. 0.9.0 (ref. [38]). Each network contained one fully connected hidden layer of 200 units with a tanh activation function and an output layer of one unit with a ReLU activation function. Models were trained with minibatch stochastic gradient descent with Nesterov momentum (batch size 500).

Comparisons to other models. RUST[19] was run via https://ribogalaxy.ucc.ie/ according to the authors' instructions. First, we computed a codon metafootprint on the Weinberg dataset, which we aligned to the transcriptome as described above. We used an A-site offset of 15 and limited the analysis to 28-nt footprints (the most abundant), in keeping with the authors' analysis. Then, we ran the 'similarity of observed and expected profiles' analysis with that codon metafootprint and retrieved the correlation of the observed and expected footprint distribution for each individual gene.

Riboshape[15] was downloaded from https://sourceforge.net/projects/riboshape/ on 1 February 2018. We generated the riboshape data structure according to the README file, with custom scripts (process_data.py and make_data_structure.m, available on GitHub), on our processed footprint-count data from the Weinberg dataset. We restricted the analysis to the 2,170 genes present in both our transcriptome and the chxdata.mat data structure shipped with riboshape. We binned our genes according to truncated lengths 100–210, 211–460, 461–710, 711–960, and 961–4,871, which matched the bins in ref. [15] after accounting for our 20 codon-truncation regions at either end of genes. Then we trained riboshape models on these bins, with σ parameters of 1, 3, 5, 12.5, 25, 37.5, 50, and 75. We report the per-gene correlations between the true footprint data and their regression fits (waveforms) in their wavelet-decomposition subspace with the least amount of denoising. The values in this subspace are closest to the observed footprint data, and their model trained for this subspace performs the best at predicting observed footprint density. We also report for each subspace the correlation between their denoised footprint data and the regression fits in that subspace. The prior is more directly comparable to our work.

Feature importance measurements. A series of leave-one-out models were trained, with exclusion of one codon position at a time from the sequence neighborhood. The importance of each codon position in predictive performance was computed as the difference in MSE between the reduced and full models.

The contribution of codon c at position i to predicted scaled counts was calculated as the average increase in predicted scaled counts due to having that codon at that position, over all instances in which codon c was observed at position i in the test set. This increase was computed relative to the expected predicted scaled counts when the codon at position i was varied according to its empirical frequency in the test set (Supplementary Note 2).

Sequence optimization. The overall translation time of a coding sequence was computed as the sum of the predicted scaled counts over all codons in that coding sequence. This quantity corresponds to the total translation time in arbitrary units. A dynamic programming algorithm was developed to find the fastest and slowest translated coding sequences in the set of synonymous coding sequences for a given protein, under a predictive model of scaled counts (Supplementary Note 3). This algorithm runs in $O(CM^L)$ time, where C is the length of the coding sequence in codons, M is the maximum multiplicity of synonymous codons (i.e., 6), and L is the length in codons of the predictive model's sequence neighborhood. This procedure achieves considerable efficiency over the naive $O(C^L)$ model, on the basis of the assumption that only codons within the sequence neighborhood influence scaled counts.

This algorithm was used to determine the fastest- and slowest-translating coding sequences for eCitrine, under a predictive model with a sequence window from codons −3 to +2, and with no structure features. Then 100,000 synonymous coding sequences for eCitrine were generated by selecting a synonymous codon uniformly at random for each amino acid. These coding sequences were scored, and the sequences at percentiles 0, 33, 67, and 100 were selected for expression experiments.

Measuring circularization efficiency. We designed oligonucleotides that mimicked the structure of the single-stranded cDNA molecule circularized by CircLigase during the ribosome profiling protocol in ref. [29]. These oligonucleotides have the structure /5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAG/iSp-18/GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCACAGTCATCGTTCG CATTACCCTGTTATCCCTAAJJJ, where /5Phos/ indicates 5′ phosphorylation; /iSP18/ indicates an 18-atom hexaethylene glycol spacer; and JJJ indicates the reverse complement of the nucleotides at the 5′ end of the footprint favored or disfavored under the model (oligonucleotides defined in Supplementary Table 2). Circularization reactions were performed with CircLigase I or II (Epicentre), as described in the manufacturer's instructions, with 1 pmol oligonucleotide in each reaction. Circularization reactions were diluted 1/20 before being subjected to qPCR with a DyNAmo HS SYBR Green qPCR Kit (Thermo Scientific) on a CFX96 Touch Real Time PCR Detection System (Bio-Rad). For each circularization reaction, two qPCR reactions were performed: one in which the formation of a product was dependent upon oligonucleotide circularization and one in which it was not (oligonucleotides defined in Supplementary Table 2). qPCR data were analyzed with custom R scripts whose core functionality is based on the packages qpcR[39] and dpcR[40] (qpcr_functions.R, available on Github). The signal from the circularization-dependent amplicon was normalized to that from the circularization-independent amplicon, then expressed as a fold change relative to the mean of the oligonucleotide with the most favored sequence under the model.

Plasmid and yeast-strain construction. Yeast integrating plasmids expressing either mCherry or a differentially optimized version of eCitrine were constructed. The differentially optimized versions of eCitrine were synthesized as gBlocks by Integrated DNA Technologies and inserted into the plasmid backbone through Gibson assembly[41]. Transcription of both mCherry and eCitrine was directed by a PGK1 promoter and an ADH1 terminator. To enable yeast transformants to grow in the absence of leucine, the plasmids contained the LEU2 expression cassette from *Kluyveromyces lactis* taken from pUG73 (ref. [42]), which was obtained from EUROSCARF. To enable integration into the yeast genome, the plasmids contained two 300-bp sequences from the *his3Δ1* locus of BY4742. Genbank files describing the plasmids are provided in Supplementary Dataset 1. To construct yeast strains expressing these plasmids, the plasmids were linearized at the *SbfI* site, and ~1 µg linearized plasmid was used to transform yeast with the high-efficiency lithium acetate/single-stranded carrier DNA/PEG method, as previously described[43]. Transformants were selected by growth on SCD –Leu plates, and plasmid integration into the genome was confirmed by yeast colony PCR with primers flanking both the upstream and downstream junctions between the plasmid sequence and the genome (oligonucleotides defined in Supplementary Table 2). PCR was performed with GoTaq DNA polymerase (Promega M8295). Haploid BY4742 and BY4741 strains expressing the eCitrine variants and mCherry, respectively, were then mated. For each eCitrine variant, eight transformants were mated to a single mCherry transformant. Diploids were isolated according to their ability to grow on SCD –Met –Lys plates. Strains with sequence-confirmed mutations or copy-number variation were excluded from further analysis.

Assessing fluorescent-protein expression by flow cytometry. Overnight cultures of diploid yeast in YEPD were diluted in YEPD so that their OD_{600} was equal to 0.1 in a 1-ml culture, then grown for 6 h in a 2-ml deep-well plate supplemented with a sterile glass bead, at 30 °C with shaking at 250 r.p.m. This culture was pelleted by 5-min centrifugation at 3,000 *g*, fixed by resuspension in 16% paraformaldehyde and incubated 30 min in the dark at room temperature. Cells were washed twice in DPBS (Gibco 14190-44) and stored in DPBS at 4 °C until analysis. Before analysis, cells were diluted ~1:4 in DPBS and subjected to flow cytometry measurements on a BD Biosciences LSR Fortessa X20 analyzer. Forward light-scatter measurements (FSC) for relative size and side-scatter measurements (SSC) for intracellular refractive index were made with a 488-nm laser. eCitrine fluorescence was measured with 488-nm (blue) laser excitation and detected with a 505-nm long-pass optical filter followed by a 530/30-nm optical filter with a bandwidth of 30 nm (530/30 or 515–545 nm). mCherry fluorescence was measured with a 561-nm (yellow–green) laser for excitation and a 595-nm long-pass optical filter followed by 610/20-nm band-pass optical filter with a bandwidth of 20 nm (or 600–620 nm). PMT values for each color channel were adjusted such that the mean of a sample of BY4743 yeast was 100. 50,000 events were collected for each sample. Flow cytometry data were analyzed with a custom R script (gateFlowData.R, available on Github) whose core functionality is based on the Bioconductor packages flowCore[44], flowStats, and flowViz[45]. In summary, for each sample, events that had values for red or yellow fluorescence that were less that 1 had those values set to 1. Then, to select events that represented normal cells, we used the curv2filter method to extract events that had FSC and SSC values within the values of the region of highest local density of all events, on the basis of their FSC and SSC values. For these events, the red fluorescence intensity was considered a measure of mCherry protein expression, and the yellow fluorescence intensity was considered a measure of eCitrine protein expression.

Measuring eCitrine and mCherry mRNA expression by qRT–PCR. Overnight cultures of diploid yeast in YEPD were diluted in YEPD so that their OD_{600} was equal to 0.1 in a 20-ml culture, then grown at 30 °C with shaking at 250 r.p.m. until their OD_{600} reached 0.4–0.6. 10 ml of culture was then pelleted by centrifugation for 5 min at 3,000 *g* and snap frozen in liquid nitrogen. Total RNA was extracted from pelleted yeast cultures according to ref. [46]. Thereafter, 10 µg of this RNA was treated with Turbo DNase I (Ambion) according to the manufacturer's instructions, and 1 µg of DNase-treated RNA was reverse transcribed with anchored oligo(dT) and Protoscript II (NEB) according to the manufacturer's instructions. One-twentieth of this reaction was then subjected to qPCR with a DyNAmo HS SYBR Green qPCR Kit (Thermo Scientific) on a CFX96 Touch Real Time PCR Detection System (Bio-Rad). For each reverse-transcription reaction, two qPCR reactions were performed: one with primers specific to the mCherry ORF and one with primers specific to the eCitrine variant ORF in question (oligonucleotides defined in Supplementary Table 2). qPCR data were analyzed with custom R scripts whose core functionality is based on the packages qpcR[39,46] and dpcR[40] (qpcr_functions.R, available on Github). The signal from each eCitrine variant ORF was normalized to that from the mCherry ORF in the same sample, then expressed as a fold change compared with the median of these values for the MIN (fastest predicted sequence) eCitrine variant.

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

Data availability. Ribosome profiling sequence data generated in this study have been deposited in the NCBI GEO database under accession number GSE106572. All Iχnos software and analysis scripts, including a complete workflow of analyses in this paper and all analyzed data used to create figures, can be found at https://github.com/lareaulab/iXnos/.

References

35. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
36. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
37. Battenberg, E. et al. Lasagne: first release. https://doi.org/10.5281/zenodo.27878 (2015).
38. The Theano Development Team et al. Theano: a Python framework for fast computation of mathematical expressions. https://arxiv.org/abs/1605.02688 (2016).
39. Ritz, C. & Spiess, A.-N. qpcR: an R package for sigmoidal model selection in quantitative real-time polymerase chain reaction analysis. *Bioinformatics* **24**, 1549–1551 (2008).
40. Burdukiewicz, M. et al. Methods for comparing multiple digital PCR experiments. *Biomol. Detect. Quantif.* **9**, 14–19 (2016).
41. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
42. Gueldener, U., Heinisch, J., Koehler, G. J., Voss, D. & Hegemann, J. H. A second set of loxP marker cassettes for Cre-mediated multiple gene knockouts in budding yeast. *Nucleic Acids Res.* **30**, e23 (2002).
43. Daniel Gietz, R. & Woods, R. A. Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. *Methods Enzymol.* **350**, 87–96 (2002).
44. Hahne, F. et al. flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics* **10**, 106 (2009).
45. Sarkar, D., Le Meur, N. & Gentleman, R. Using flowViz to visualize flow cytometry data. *Bioinformatics* **24**, 878–879 (2008).
46. Ares, M. Isolation of total RNA from yeast cell cultures. *Cold Spring Harb. Protoc.* **2012**, 1082–1086 (2012).

nature research

Corresponding author(s): Liana F Lareau

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

We grew up strains from 8 yeast colonies for each version of eCitrine, representing 8 independent integrations of the eCitrine reporter into a fixed locus.

2. Data exclusions

Describe any data exclusions.

qPCR: We measured mRNA levels in three biological replicates (independent integrations of the eCitrine reporter) for each version of eCitrine. One integrant of eCitrineMAX was excluded because the RNA pellet was aspirated during the experiment, leaving two biological replicates of that strain.
Flow cytometry: Biological replicates of yeast strains were excluded for the following reasons, documented in our analysis scripts on GitHub:
1. eCitrine333: we discovered a polymorphism in the donor plasmid and tested each integrant for the mutation. We discarded 4 of 8 integrants because they were not wildtype at that position.
2. eCitrine000, eCitrine333, yECitrine: one integrant of each eCitrine variant was discarded because the FACS data did not pass pre-determined filters for number of cells or had bimodal fluorescence.
3. eCitrine999: one integrant was discarded because genomic copy number qPCR gave results inconsistent with a single copy integration.

3. Replication

Describe whether the experimental findings were reliably reproduced.

All attempts at replication were successful.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

For neural network training, genes were randomly allocated into the training set (2/3) or test set (1/3).

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Blinding was not relevant to our study.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

All iXnos software, transcriptome files used for alignments, and a complete script to reproduce our entire analysis including figures are available at: https://github.com/lareaulab/iXnos
Linkers were trimmed from ribosome footprints with custom code as documented in the paper and on GitHub.
Alignments were performed with bowtie v. 1.2.1.1 and RSEM v. 1.2.31.
The neural network method, iXnos, was implemented in python and is fully documented in the paper and on GitHub. It relied on python packages Lasagne v. 0.2.dev1 and Theano v. 0.9.0.
Flow cytometry was analyzed with custom R scripts fully documented on GitHub. They incorporated bioconductor packages flowCore v. 1.40.3, flowStats v. 3.32.0, and flowViz v. 1.38.0.
All methods used to create the figures, including statistics and qPCR analsysis, are included as R scripts on GitHub.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

All materials and strains are readily available. Yeast strains will be provided on request. Commercial sources:
Direct-zol RNA MiniPrep kit (Zymo Research)
CircLigase I and II (Epicentre)
DyNAmo HS SYBR Green qPCR Kit (Thermo Scientific)
GoTaq DNA polymerase (Promega)
Turbo DNase I (Ambion)
Protoscript II (NEB)

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

S. cerevisiae strains were constructed from frozen stocks of BY4741 and BY4742 obtained from Thermo Fisher.

b. Describe the method of cell line authentication used.

Mating type and marker genotype were authenticated by growth on selective plates.

c. Report whether the cell lines were tested for mycoplasma contamination.

Yeast are not susceptible to mycoplasma contamination.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

The study did not involve human research participants.