nature biotechnology

Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*

Guillaume Cambray[1,2], Joao C Guimaraes[1,3] & Adam Paul Arkin[3,4]

Comparative analyses of natural and mutated sequences have been used to probe mechanisms of gene expression, but small sample sizes may produce biased outcomes. We applied an unbiased design-of-experiments approach to disentangle factors suspected to affect translation efficiency in *E. coli*. We precisely designed 244,000 DNA sequences implementing 56 replicates of a full factorial design to evaluate nucleotide, secondary structure, codon and amino acid properties in combination. For each sequence, we measured reporter transcript abundance and decay, polysome profiles, protein production and growth rates. Associations between designed sequences properties and these consequent phenotypes were dominated by secondary structures and their interactions within transcripts. We confirmed that transcript structure generally limits translation initiation and demonstrated its physiological cost using an epigenetic assay. Codon composition has a sizable impact on translatability, but only in comparatively rare elongation-limited transcripts. We propose a set of design principles to improve translation efficiency that would benefit from more accurate prediction of secondary structures *in vivo*.

Translation is one of the most conserved and energy-intensive processes in any cell. In bacteria, synthesis and maintenance of ribosomes and related translational machinery accounts for more than 40% of the cellular energy budget[1]. This means that, for any translated sequence, the determinants of translational efficiency are likely to be under selection to minimize cost[2]. Aside from specific regulatory mechanisms, optimization of translation is thought to involve fine-tuning intrinsic sequence properties. Although hundreds of studies have linked such features to protein production, very few have also measured physiological outcomes of sequence variation or attempted to integrate translation efficiencies and physiology[3–5].

Highly expressed genes are enriched in codons cognate to abundant tRNAs[6], and gene expression levels usually correlate with indices describing these enrichments[7,8]. Biased codon usage is thought to optimize elongation by preventing wasteful ribosome collisions and stacking[9]. Similarly, incorporation of specific amino acids[10] and secondary structures in transcripts[11,12] has been linked with translation speed. Experimental studies on endogenous and designed transcripts have identified the importance of RNA structures in hindering initiation by blocking ribosome binding sites[13–16]. The sequence of the 5′ untranslated region and the first 30–50 codons of a gene is thought to contain most determinants of translation efficiency[17]. Selection for a translation speed ramp encoded at the 5′ end of coding sequences via the above mechanisms has been proposed to balance the flow of initiating ribosomes with the elongation capacity of the downstream sequence[18,19].

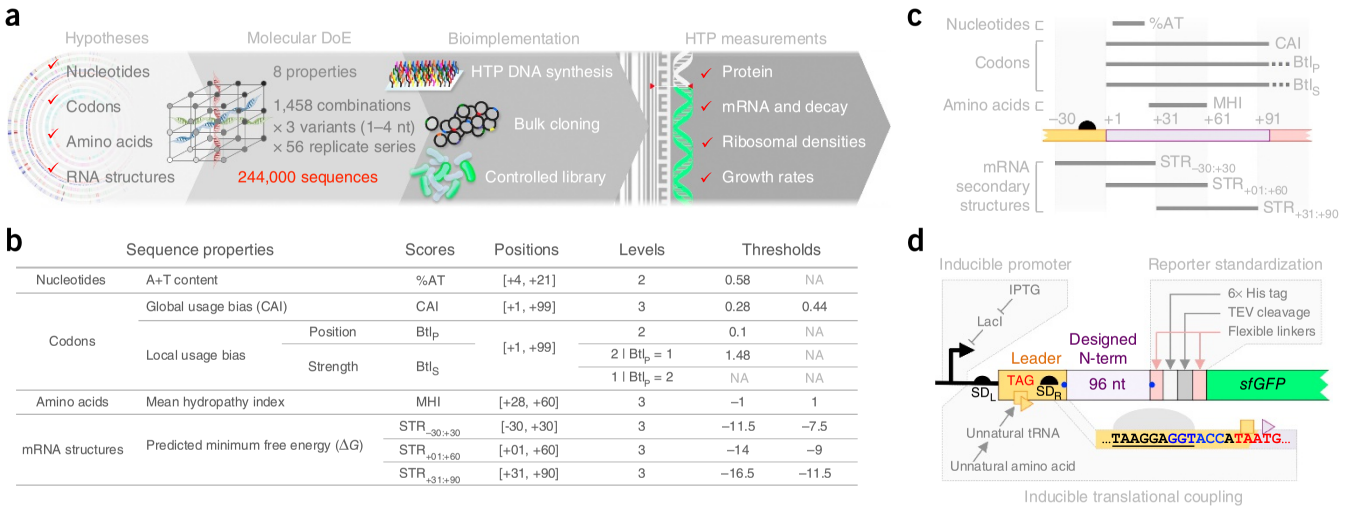
In practice, discerning the relative importance of sequence-intrinsic features and the interactions among them has proven difficult. This is because most sequence–activity studies involve analysis of endogenous genes, which have evolved highly scrambled control signals. For example, a bias for N-terminal positive charge is found almost exclusively in weakly expressed membrane proteins. This type of protein sequence enables association with the cytosolic membrane leaflet[20] rather than ribosomal slowing[19], but it is associated with both.

Though intended for tests of causation, studies using designed sequences might suffer from similar confounding effects. For example, occurrence of rare codons in a synthetic library comprising 13 synonymous variants of the first 10 codons from 137 *E. coli* genes fused to a standard reporter was found to correlate with increased protein production. Although this finding seemed to support the codon ramp hypothesis, *post hoc* analysis revealed that these rare codons were also A+T-rich, resulting in weaker secondary structures in mRNAs and presumably higher initiation[21]. Notwithstanding caveats associated with its simple design, this codon variant library was further used to show various effects of transcript structure, codon and amino acid properties on cell growth[5]. To our knowledge, the largest study to formally attempt to disentangle multiple sequence features by design used a library of 285 synthetic genes and concluded that A+T content over the first 35 nucleotides (nt) rather than structures or codons has the largest effect on translation efficiency[22]. This study was done *in vitro* and could not assess physiological effects *in vivo*.

We undertook a comprehensive analysis of factors affecting translation efficiency in *E. coli* by implementing a massive design of experiments (DoE) at the molecular level. DoE is widely used in many fields (for example, agronomy, clinical trials, chemical process

[1]California Institute for Quantitative Biosciences, University of California, Berkeley, Berkeley, California, USA. [2]DGIMI, Univ. Montpellier, INRA, Montpellier, France. [3]Department of Bioengineering, University of California, Berkeley, Berkeley, California, USA. [4]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA. Correspondence should be addressed to G.C. (guillaume.cambray@inra.fr) or A.P.A. (aparkin@lbl.gov).

Received 24 November 2017; accepted 2 August 2018; published online 24 September 2018; doi:10.1038/nbt.4238



b

Sequence properties			Scores	Positions	Levels	Thresholds	
Nucleotides	A+T content		%AT	[+4, +21]	2	0.58	NA
Codons	Global usage bias (CAI)		CAI	[+1, +99]	3	0.28	0.44
	Local usage bias	Position	Btl_P	[+1, +99]	2	0.1	NA
		Strength	Btl_S		2 | Btl_P = 1	1.48	NA
					1 | Btl_P = 2	NA	NA
Amino acids	Mean hydropathy index		MHI	[+28, +60]	3	–1	1
mRNA structures	Predicted minimum free energy (ΔG)		$STR_{-30:+30}$	[-30, +30]	3	–11.5	–7.5
			$STR_{+01:+60}$	[+01, +60]	3	–14	–9
			$STR_{+31:+90}$	[+31, +90]	3	–16.5	–11.5

Figure 1 High-throughput design of experiments. (**a**) A workflow for hypothesis-driven functional genomics applied to translation. 244,000 digital DNA sequences were designed programmatically to produce 56 independent libraries. Each library covers every combination of eight sequence properties previously found to correlate with protein production in *E. coli*. Designer sequences were synthesized and cloned in bulk. Multiple phenotypes affording an integrative study of the genetic perturbations' impact on the cellular system were characterized using high-throughput (HTP) sequencing as a generic quantitative readout. DoE, design of experiments. (**b**) Sequence properties varied in the factorial design. (**c**) Topological mapping of sequence properties onto the reporter system. As the functional contributions of properties defined from the same underlying sequence are intricately intertwined, dissecting their separate contributions requires systematic sequence and experimental design. (**d**) A standard reporter for translational output. Synthetic sequences are cloned as N-terminal fusions to a modified *sfGFP* gene, which leads to the production of an invariant fluorescent reporter upon post-translational processing by the TEV protease. Translation of the reporter is driven by a perfect Shine–Dalgarno motif (SD_R, underlined) embedded into the leader sequence of an inducible translational coupling device (yellow), itself driven by its own upstream SD (SD_L). Translation of the leader cistron can only occur when we add a specific unnatural amino acid to the culture, which suppresses a stop codon placed early in the sequence through a cognate synthetic tRNA[32]. Ribosomes able to translate past this point unwind RNA structure ahead of them, thus making the embedded SD_R more accessible. Ribosomes terminating on the leader need not dissociate from the transcript to reinitiate translation of the reporter, further improving initiation rates[13,35]. The system is borne on a medium-copy plasmid (colE1 origin).

optimization, and mechanical design)[23] and aims to explain the variance in a set of experimental observables by systematically planned variations of a set of explanatory variables[24]. In the framework of studying sequence–activity relationships, explanatory variables are not extrinsic and independent treatments (for example, temperature and osmolarity) but rather intrinsic and interdependent properties of the DNA molecule (for example, synonymous sequences varying A+T content but not transcript secondary structures nor codon usage). As such constraints significantly complicate implementation, very few studies have applied DoE to optimize[25,26] or characterize[22,27,28] genetic systems.

Here we design 244,000 synthetic sequences to systematically explore an eight-dimensional space defined by the main sequence properties introduced above and use high-throughput DNA sequencing to measure the consequences of these sequence perturbations on several phenotypes linked to translation efficiency (**Fig. 1a**).

RESULTS

Design of experiments to understand translation

We selected eight intrinsic sequence properties reported to be the main factors affecting translation efficiency (**Fig. 1b**,**c**). These predictors describe sequence nucleotide content (A+T content (%AT)), patterns of codon usage (codon adaptation index (CAI), codon ramp bottleneck position (Btl_P) and strength (Btl_S)), hydrophobicity of the coded polypeptide (mean hydrophobicity index (MHI)) and stability of secondary structures tiled along the transcript ($STR_{-30:+30}$, $STR_{+01:+60}$, $STR_{+31:+90}$). These explanatory variables were arrayed into a statistical full factorial design (**Fig. 1b**). Full factorial designs vary all predictors (over a discrete range of levels) in all combinations. A set of observations is then performed on each combination. The data are treated with analysis of variance (ANOVA) and other regression analyses to quantify the effect of each predictor, and all possible interactions between them, on the observables, given all possible values of the other predictors.

To derive relevant physiological ranges for our eight predictors, we calculated values of each sequence properties for all 3,990 protein coding genes annotated in a reference *E. coli* genome and examined their associations with matched published measures of mRNA and protein abundance[29] (Online Methods and **Supplementary Data 1–4**). On the basis of this analysis (**Supplementary Fig. 1**), we assigned either two or three discrete levels over which to vary each parameter, leading to 1,458 unique combinations of parameter levels (**Fig. 1b**). Nearly half these combinations are not represented in the genome, and randomly generated sequences produce a very skewed sampling of this discretized property space (**Supplementary Fig. 2a**,**b** and **Supplementary Data 5** and **6**).

We used D-Tailor[30] to derive 56 fully independent mutational series, each representing a complete replicate of the full factorial design. For each sequence, we further generated two closely related replicates that differ by one to four random points mutations but retain the same particular combination of properties (Online Methods, **Supplementary Figs. 2c–e** and **3**, and **Supplementary Data 7** and **8**). The resulting set of 244,000 sequences was synthesized on a high-density oligonucleotide array, PCR-amplified and cloned into a reporter plasmid, and the resulting library was introduced into *E. coli* MDS42 by transformation (Online Methods). Amplicon sequencing

of the cloned sequences readily permitted observation of 99.4% of the library (242,516 strains with >10 reads; **Supplementary Fig. 4** and **Supplementary Data 9** and **10**).

Several features of our expression system were chosen to minimize potential bias in experimental outcomes. The expression cassette (**Fig. 1d**) accepts the library as a specific protease-cleavable N-terminal translational fusion to the superfolder GFP gene (*sfGFP*)[31]. This cleavage produces a processed fluorescent reporter that is invariant across strains, thus reducing potential post-translational effects of the designed sequence (for example, differential protein stability). Sequences are transcribed from an IPTG-inducible promoter to avoid strain competition during library construction and propagation. Translation initiates at a Shine–Dalgarno signal (SD_R) embedded in a short leader sequence that overlap the fused reporter by a single nucleotide (**Fig. 1d**). Translation of that leader cistron permits unwinding of potential mRNA structures around the SD_R to mitigate their impact on reporter gene translation[15]. We derived the pEVOL unnatural suppressor system[32] to enable fully tunable control of leader sequence translation (**Supplementary Fig. 5** and **Supplementary Data 11**–**13**). This allows us to assess structural control of initiation without confounding effects of sequence modification.

Quantification of effects on protein production

We used fluorescence-activated cell sorting followed by high-throughput targeted sequencing of the designed region to measure fluorescent protein production as a proxy for translation rate[28,33]. We quantified protein production under normal initiation (P_{NI}) conditions for 242,269 strains in the library, aggregating four highly reproducible measurements of independent biological replicates (Online Methods, **Supplementary Fig. 6a**–**c**, **Supplementary Code 1**–**6** and **Supplementary Data 14** and **15**).

We conducted a multiway ANOVA[27,28] to quantify the relative contribution of each sequence property and its first-order interactions with P_{NI} (**Supplementary Code 7** and **Supplementary Data 16**). First-order interactions quantify the dependence of one property's effect on another property's levels. This analysis revealed that mRNA secondary structures around the start codon (30 nt in either direction; $STR_{-30:+30}$) have the biggest effect on translation efficiency, accounting for 83% of design properties' total effect (**Fig. 2a**, top). The main other contributors to variability of translation are %AT (7%); $STR_{+01:+60}$ (4%) and its interaction with $STR_{-30:+30}$ (3%); and $STR_{+31:+90}$ (2%) and its interaction with $STR_{+01:+60}$ (1%). Most of these properties involve structures. Even %AT might capture unaccounted structural signal. Similar pictures emerge from conducting multiple and recursive regression analyses (**Supplementary Fig. 6d**,**e**, **Supplementary Code 8** and **9**, and **Supplementary Data 17** and **18**).

Taken together, design properties and first-order interactions could only explain 28% of P_{NI} variance, suggesting that there are important factors that are not accounted for. This prompted us to dissect different sources of error in the experimental design (**Fig. 2a**, bottom). Variations among the three combinatorial replicates built within each mutational series (design error) represent 8% of the variance and vastly exceed the experimental error (**Supplementary Fig. 6c**). This underlines an inability of property scoring algorithms to capture functionally important differences between very similar nucleotide sequences (1–4 mutations; >95% pairwise identities) and hints at the roughness of sequence–activity landscapes. This strongly suggests that we are missing informative sequences descriptors. We also observed highly variable P_{NI} profiles among mutational series (**Fig. 2b**, left). When included in the ANOVA, this serial identity and its interactions reclaim 10% and 6% of unexplained variance, respectively (**Fig. 2a**, bottom).

Performing independent ANOVAs for each series further reveal a diversity of explanation patterns (**Fig. 2b**, left; **Supplementary Code 5**; and **Supplementary Data 15**), which must be linked to differences between series' core sequences (**Supplementary Fig. 3**). Explainable variances range between 25% and 63%, with an average of 43%. These numbers correlate with series-wise means and variances in P_{NI} (**Supplementary Fig. 6f**), suggesting that poor explanation patterns largely stem from faulty design of key sequence properties and consequent failure to elicit varied phenotypic responses. Secondary structures, for example, are emergent properties whose *in vivo* formation is hard to predict[34]. Systematic discrepancies between computationally predicted and actual folding could heavily influence P_{NI} distribution for a given series. Hence, both sequence-specific misrepresentations of the design properties (design errors) and largely unidentified features shared by most members of a series (serial errors) represent important limitations to our understanding of translation efficiency.

The role of secondary structure in hindering translation initiation has recently gained renewed interest[13–16]. $STR_{-30:+30}$ is the main feature in all but two of our series (**Fig. 2b**). Its contribution is inversely correlated to that of other features (**Fig. 2c**), which suggests that otherwise relevant sequence properties may have little effect on P_{NI} if the structural context limits initiation. For example, we find little impact of CAI overall (0.05%)—but in the top P_{NI} quintile, it accounts for 13% of design properties' total effect (**Supplementary Fig. 6g**).

Increased initiation unmasks elongation-limiting properties

We used our inducible bicistronic controller (**Fig. 1d**) to examine the behavior of the same sequence library subjected to increased translation initiation rates. To translate the full bicistronic leader, ribosomes must unfold potential secondary structures within the overlapping initiation region of the downstream reporter. We reasoned this would facilitate initiation by either free or scanning ribosomes upon leader termination (translational coupling)[13,35], without modification of the underlying sequence. Thereafter, we refer to this regime as facilitated initiation. As above, we used fluorescence-activated cell sorting and amplicon sequencing to measure protein production under facilitated initiation (P_{FI}) for 238,458 strains (Online Methods, **Supplementary Fig. 7a**–**c**, **Supplementary Code 1**–**6** and **Supplementary Data 14** and **15**).

Higher initiation rates caused a global but non-uniform increase in protein production (**Fig. 2d**,**e**). ANOVAs revealed that P_{FI} was still mainly explained by $STR_{-30:+30}$ (**Fig. 2d**–**f**, **Supplementary Code 10** and **Supplementary Data 19**). We hypothesize that more stable structures might refold more efficiently between translating ribosomes, which in turn reduces SD_R accessibility to both free and scanning ribosomes. A different explanation might be that secondary structures slow elongation by leader-bound ribosomes (**Fig. 3a**). Either way, application of our epigenetic method to reduce structures around the start codon uncovered a role for two sequence features: $STR_{+01:+60}$ and CAI (**Fig. 2f**, **Supplementary Fig. 7d**–**f**, **Supplementary Code 11** and **12**, and **Supplementary Data 20** and **21**).

$STR_{+01:+60}$ (32.9%) and its interactions with both P_{NI} (21.5%) and $STR_{+31:+90}$ (1.8%) showed the strongest contributions to production improvement under facilitated initiation. $STR_{+01:+60}$ affects initiation without segregating either the SD_R or start codon (**Fig. 3a**,**e**). Interestingly, they do so most strongly when not interfered with by more stable $STR_{+31:+90}$ (**Fig. 3b**). In fact, dynamic competition among all three of our designed structures determined complex patterns of interactions that influenced translation efficiencies (**Fig. 3c**). Our analysis thus reveals that distal RNA structures can indirectly affect translation initiation by outcompeting initiation-limiting structures[36].

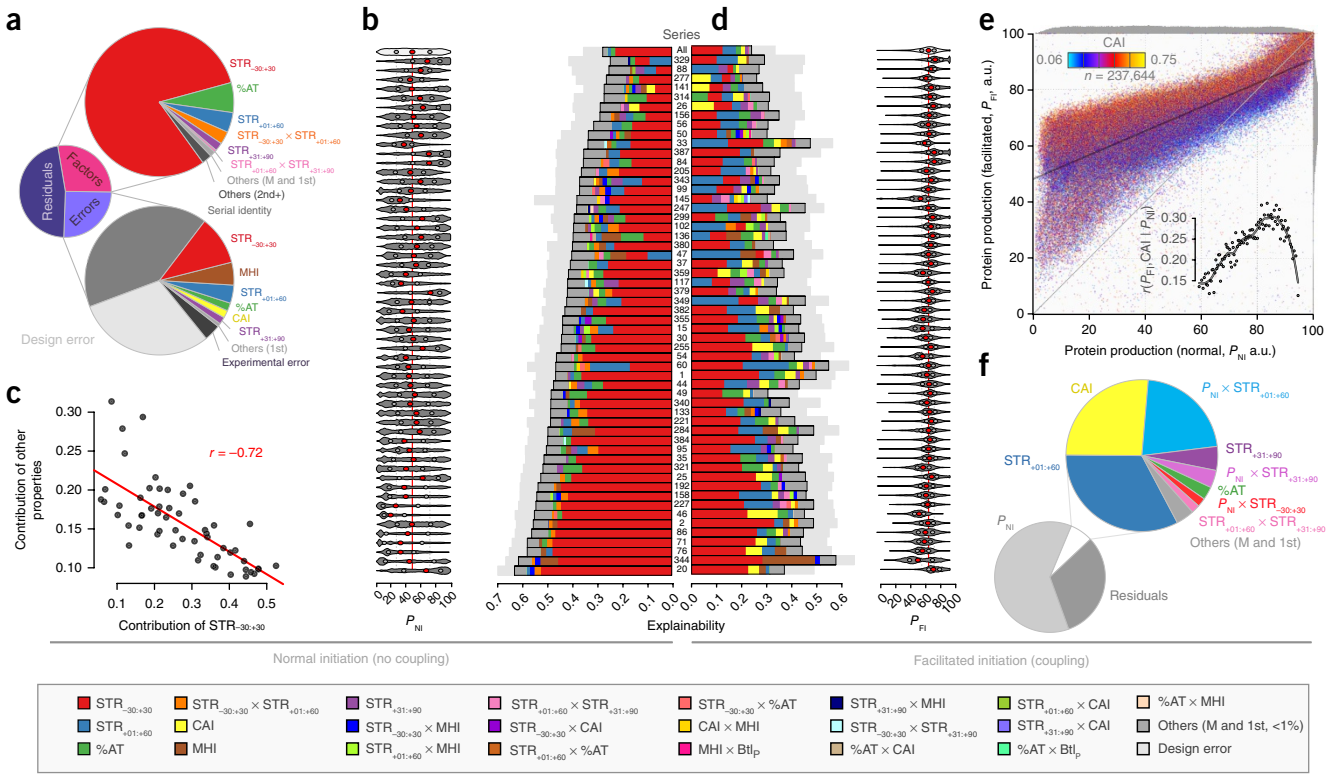


Figure 2 Protein production under normal and facilitated conditions of translation initiation. (**a**) Contributions of properties and errors to variations in protein production. Nested pie charts show effect sizes quantified by ANOVA over the whole library (n = 242,269 strains). Upper pie details the variance explained by the design properties and their interactions. Lower pie shows the contribution of various types of errors: experimental, within-series differences between close replicates (design error), between-series differences (factorial identity) and first-order interactions between factorial identity and design properties (colored slices). Small effects are grouped in gray, distinguishing interaction depth (M, main effects; 1st, first-order interactions; 2nd+, higher-order interactions). (**b**) Series-wise decomposition of variance in protein production. For each series, effects >1% are stacked by decreasing values (colors as shown) and smaller effects are grouped next (dark gray). The design error is shown at the end of each bar (light gray). Protein production distributions are shown at left, with red and gray dots marking quartiles ($n \in$ [4,429; 4,372] strains for each series, except n = 3,418 for incomplete series no. 136). (**c**) Hierarchical dominance of secondary structures in the initiation region. For each series, the combined effect sizes of all other properties are plotted against that of $STR_{-30:+30}$ (sample sizes as in **b**). Pearson's correlation $r = -0.72$. (**d**) Series-wise decomposition of explainable variance in protein production under facilitated initiation. Translational coupling increases the protein production profiles (right, colored points as in **b**) and alters the property contribution patterns (left) (n = 238,458 for the whole dataset and $n \in$ [3093; 4,368] strains for each series). (**e**) Comparing production regimes exposes the effect of codon usage. Scatter plot of P_{FI} versus P_{NI}, colored by CAI (n = 237,644 strains); a.u., arbitrary units. The linear regression of P_{FI} on P_{NI} is plotted (thick line). Inset shows rising correlations between P_{FI} and CAI with increasing percentiles of P_{NI}. Higher codon adaptation supports commensurate increases in elongation rates when initiation is improved. (**f**) Specific property contributions under facilitated initiation. Decomposition of variance in P_{FI} upon multiple linear regression including P_{NI} as a predictor (n = 237,644 strains). The effects of design properties under normal translation are captured by P_{NI}. Remaining contributions quantify differential effects elicited under coupling. Properties with effect <1% of the inset pie are grouped in gray (M, main effects; 1st, first-order interactions).

This mechanism is likely exploited by natural regulatory processes, as we found that distal structures in *E. coli* genes tend to be more stable than expected by chance and *a fortiori* more stable than those in the initiation region (**Supplementary Fig. 1k**).

Fine-grained analyses using structures computed across sliding windows yielded smooth explainability profiles but revealed dramatic diversity in optimal window length and position across mutational series (**Fig. 3d**, **Supplementary Code 13** and **Supplementary Data 22–24**). A model based on single nucleotide accessibilities instead of arbitrarily fixed windows confirmed the general importance of the Shine–Dalgarno (SD) motif (**Fig. 3e**) and provided better descriptions of individual series (**Fig. 3f**). Like current approaches, however, it failed to improve explanation of the whole dataset (**Supplementary Code 14–16** and **Supplementary Data 25–29**). These analyses suggest that static structure predictions may not be sufficient to enable robust predictions of translation efficiency (see Discussion).

While codon composition represented only 0.05% of the variance explained by the design properties under normal initiation (**Fig. 2a**), it contributed more than a quarter (26.2%) of the improvement observed when structural constraints on initiation were alleviated (**Fig. 2d**). For any given P_{NI}, sequences with higher CAI have higher P_{FI}, demonstrating that rates of protein production were formerly limited by initiation (**Fig. 2e**). The strength of the observed association increases as residual initiation-limiting impact of structures decreases (**Fig. 2e** inset and **Supplementary Fig. 7f**).

We found no measurable effect of the codon ramp on either P_{NI}, P_{FI} or their relationship (**Supplementary Fig. 8a**,**b**). We further investigated a variety of published codon metrics[18,37–41], but none outperformed the CAI[7] in describing our data (**Supplementary Fig. 8c**, **Supplementary Code 17** and **Supplementary Data 30**). We were able to derive improved codon indices that maximize correlations with our data (**Supplementary Fig. 8c**)—but their broader

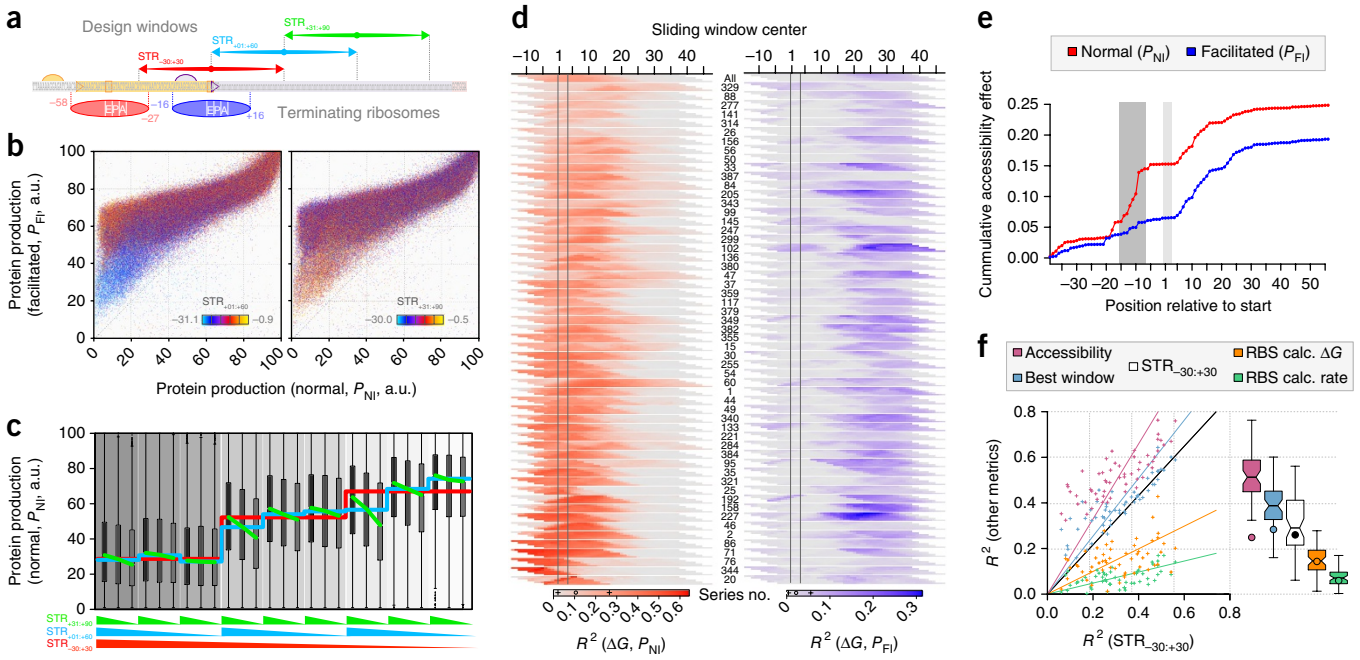


Figure 3 Dynamic structure interactions hinder functional predictions. (**a**) Steric relationships between ribosomes and secondary structures. The reporter system and designed structures are detailed at scale. Ovals show the precise footprint of leader-bound terminating ribosomes in normal (red) and facilitated (blue) initiation regimes. (**b**) Emergent codistributions of designed structures in protein production spaces. Scatter plots of P_{FI} versus P_{NI} colored by $STR_{+01:+60}$ (left) or $STR_{+31:+90}$ (right) (n = 237,644 strains); a.u., arbitrary units. These distributions highlight complex structural interaction patterns, especially among low protein producers. (**c**) Structure interactions affect protein production. Box plots outline the distribution of P_{NI} for every combination of designed structures, as shown ($n \in$ [8,902; 9,026] strains for each plot; boxes mark interquartile ranges, whiskers measure 1.5 times these ranges). Red, blue and green lines further highlight, respectively, medians per level of $STR_{-30:+30}$, $STR_{+01:+60}$ given $STR_{-30:+30}$, and $STR_{+31:+90}$ given both $STR_{-30:+30}$ and $STR_{+01:+60}$. Strong $STR_{+31:+90}$ provides a dynamic switch capable of facilitating initiation through thermodynamic competition with $STR_{+01:+60}$. (**d**) Highly variable structural profiles between series. For each series, profiles show the R^2 of regressions of P_{NI} (left, n = 242,269 strains altogether) or P_{FI} (right, n = 238,458 strains altogether) on ΔG predictions for sliding windows of different sizes (50 to 70 nt by increments of 5). Values are color-coded on a heat map (thermometer as shown) wherein columns mark windows' centers, while rows within each series correspond to increasing window length (top to bottom). (**e**) Functional impact of sequence accessibility with nucleotide resolution. Position-wise contributions are obtained by regressing average nucleotide accessibility of each positions of every transcript against P_{NI} (red) or P_{FI} (blue). Cumulative plotting of effect sizes highlights decreased contribution of SD (dark gray) accessibility under coupling and lack of start codon involvement (light gray). (**f**) Predictions of protein production are limited. R^2 from series-wise regressions of P_{NI} on $STR_{-30:+30}$ are plotted against those obtained by regressing P_{NI} on nucleotide accessibilities (pink), best series-wise structure window (blue), composed minimum free energy (orange) and initiation rate (green, both) from the RBS calculator[16] (RBS calc.) (n = 56 mutational series). Corresponding box plots are shown on the right, with circles marking R^2 over the entire dataset (boxes mark interquartile ranges, whiskers measure 1.5 times these ranges, central lines mark the medians and notches their 95% confidence intervals).

significance is questionable, as they fail to describe natural genes better than the original CAI (**Supplementary Fig. 8d**). When averaged over the whole design sequence, hydropathy yielded a weaker but similar signal to CAI (**Supplementary Fig. 9a**). Therefore, by applying our bicistronic controller to increase initiation rates without modifying the underlying sequence, we confirmed the impact of codon and to a lesser extent amino acid composition on elongation rates.

Unexpected cost of slow translation initiation

Recent hypotheses for the cost of translation have focused on the time that ribosomes spend in translation. Inspired by Opijnen *et al.*[42], we were able to batch quantify the relative growth rates of 233,846 and 229,224 strains under normal (W_{NI}) and facilitated (W_{FI}) initiation, respectively, by comparing sequencing counts of the design sequences in samples aliquoted along serial passages of the library (Online Methods, **Supplementary Fig. 10a–d**, **Supplementary Code 18** and **Supplementary Data 9**). As with protein production, we observed wide variations in the means and spreads of growth among mutational series (**Fig. 4a**,**d**).

W_{NI} has strongly non-monotonic relationships to P_{NI} that are best captured by quadratic regressions with optimal growth rates at intermediate protein production (**Fig. 4b**,**c**, **Supplementary Code 19** and **Supplementary Data 31**). This is unexpected, because growth has often been shown to follow a negative linear relationship with gratuitous protein expression[3,43]. A few outliers, exceedingly unfit considering their low protein production (underachievers), were recently noted in a large-scale study[5]. At our even larger scale, these strains were numerous enough to heavily bend the trend. Remarkably, facilitating initiation largely improved the growth of underachievers despite increasing their protein production (**Fig. 4e**,**g**, middle), yielding largely linear relationships between W_{FI} and P_{FI} (**Fig. 4e**,**f**, **Supplementary Code 20** and **Supplementary Data 32**).

We found only a small contribution of CAI to growth in either condition (**Fig. 4c**,**f** and **Supplementary Fig. 10e**). Once direct effects on protein production are accounted for, secondary

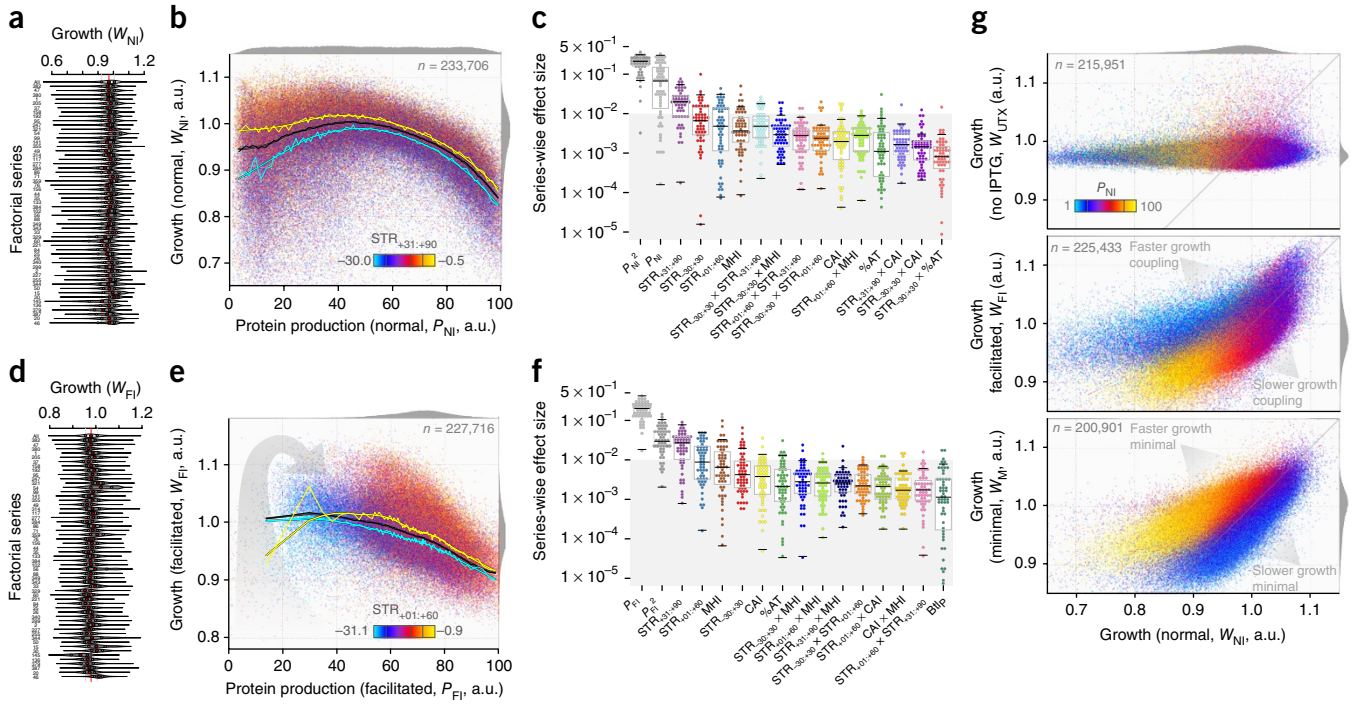


Figure 4 Unexpected growth defects associated with reduced translation initiation. (**a**) Variable growth profiles between factorial series. Library-wide and series-wise distributions of growth rate as measured through sequencing under normal initiation conditions (W_{NI}). Red and gray dots mark medians and interquartile ranges, respectively (n = 233,846 strains for the whole dataset and $n \in$ [3,282; 4,353] strains for each series). (**b**) Lowest protein productions are unexpectedly associated with lower growth rates. Scatter plot of W_{NI} as a function of P_{NI}, colored by $STR_{+31:+90}$ (n = 233,706 strains); a.u., arbitrary units. Dark lines mark the median W_{NI} for each percentile of P_{NI} (thin line) and a loess smoother (thick line) highlighting the biphasic relationship. Yellow and cyan lines show the same information for the top and bottom deciles of $STR_{+31:+90}$, respectively. $STR_{+31:+90}$ is positively correlated with W_{NI}, especially at low P_{NI}. (**c**) Specific contributions of design properties to growth rate strengthen the functional importance of secondary structures. Multiple linear regressions of W_{NI} including P_{NI} and P_{NI}^2 as explanatory variables quantify the biphasic relationship highlighted in **b** and dissociate direct impacts of design properties from those mediated by protein production. Shown are effect sizes for explanatory variables with values >1% in at least one series ($n \in$ [3,282; 4,353] strains for each series). Corresponding box plots are shown in the background (boxes mark interquartile ranges, with whisker termini drawn at 1.5 times these ranges; central lines mark the medians). (**d**–**f**) Facilitating initiation relieves the fitness cost despite increased protein production. Same as **a**–**c**, respectively, for growth (W_{FI}; n = 229,224 strains for the whole dataset and $n \in$ [3,066; 4,354] strains for each series) and protein production (P_{FI}) under condition of facilitated initiation. Points in **e** are colored by $STR_{+01:+60}$ (n = 227,716 strains). The biphasic relationship is less marked than under normal conditions. The gray arrow drawn in **d** visually highlights this transition. See also the prevalence of P_{FI} over P_{FI}^2 in **f** ($n \in$ [2,806; 4,350] strains for each series). (**g**) Low translation initiation is physiologically costly. Scatter plots of growth rates in different culture conditions as a function of W_{NI}, colored by P_{NI}. Expression of reporter transcripts is necessary to observe growth differences (top; n = 215,951 strains). Facilitating initiation through translational coupling increases growth rates among the lowest protein producers (middle; n = 225,433 strains). Lower ribosome abundance after growth in minimal medium magnifies the growth defect W_M among low protein producers (bottom; n = 200,901 strains).

structures—$STR_{+31:+90}$ in particular—have a specific, negative effect on growth (**Fig. 4b**,**c**,**e**,**f** and **Supplementary Fig. 10f**,**g**). The relative growth defect of underachievers is exacerbated in minimal medium, a condition known to make ribosomes scarce[3,44] (**Fig. 4g**, bottom). This would suggest that structural features might increase the cost of translation by increasing ribosome sequestration through slower elongation. But why then would such an effect be largely conditioned on poor initiation (**Fig. 4e**,**g**, middle)? Poor initiation is not expected to cause ribosome queuing in itself, and slow elongation through secondary structures should only matter under nonlimiting initiation (as observed for CAI). After verifying that transcript production was indeed necessary to the observed growth differences (**Fig. 4g**, top), we sought to gain further insights by collecting data on reporter transcript abundance.

Equilibrium of transcript abundance and translation efficiency

Translation efficiency is often expressed as the ratio of protein or ribosome footprints to transcript abundance[45]. However, features governing translation have repeatedly been reported to influence transcript stability[40,46,47]. If significant, such translation-to-transcription feedbacks could affect our interpretation of translation efficiency. To assess this, we used a high-throughput decay assay to quantify the abundance of reporter transcripts at steady state (RNA_{SS}), their half-life (RNA_{HL}) and the fraction resilient to degradation (RNA_{PTX}) for 233,487 strains under normal conditions of initiation (Online Methods, **Supplementary Fig. 11a**–**e**, **Supplementary Code 21** and **Supplementary Data 9**, **33** and **34**).

We found that substantial variations in RNA parameters were correlated with P_{NI}. Low protein producers had high RNA_{SS} that was associated with full mRNA degradation (**Fig. 5a**, blue); high protein producers had reduced RNA_{SS} and resilience to mRNA degradation (red). We hypothesize that these observations arise from secondary structure interactions that affect both degradation and translation.

Secondary structures in mRNA are known to hinder RNases[48]. Low protein producers tend to have strong $STR_{-30:+30}$, yielding higher

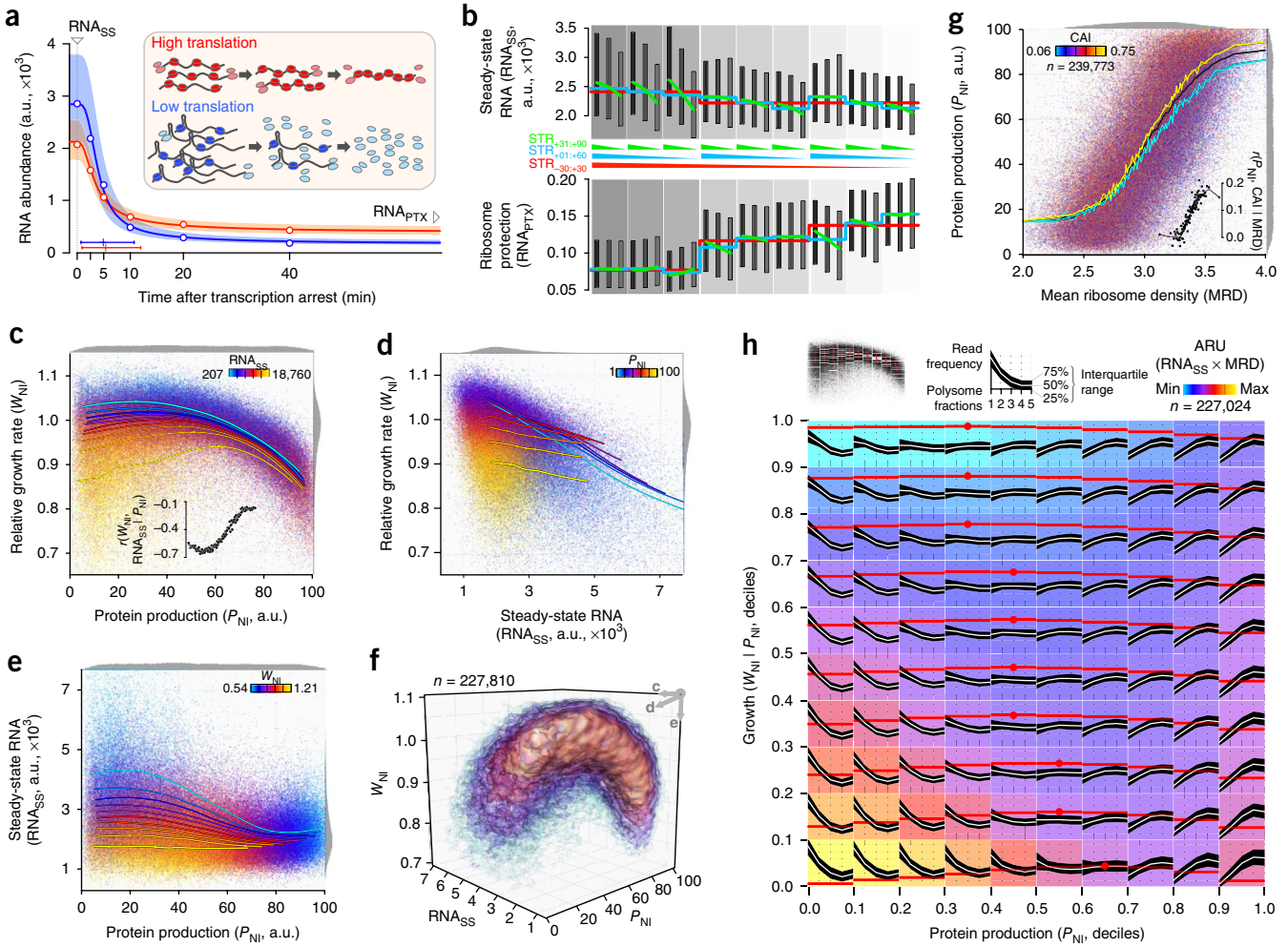


Figure 5 Pathological accumulation of stable transcripts inhibits initiation rate. (**a**) Translation regime influences RNA abundance and decay. RNA abundances are derived from the ratio of RNA to cognate DNA counts upon sequencing, and thus expressed in arbitrary units (a.u.). Red and blue points show medians for the top and bottom quartile of P_{NI}, respectively (n = 58,211 strains each). Red and blue lines and shadings mark corresponding decay fits and interquartile ranges. Ribosomal protection intensifies during the assay in highly translated transcripts (inset, red). Secondary structures limit ribosomal protection but increase transcripts stability and RNA_{SS} (inset, blue). (**b**) Interactions between secondary structures affect RNA abundance (top) and ribosome protection (bottom). Interaction plots as in **Figure 3c** ($n \in$ [8,365; 8,880] stains for each box plot; for clarity, whiskers and outliers are not plotted). (**c**) Low-producing, slow-growing strains show the highest transcript abundance. Scatter plot of W_{NI} as a function of P_{NI}, colored by RNA_{SS}. Lines show loess regressions for every decile of RNA_{SS}, following the same color code, with points marking local growth maxima. Line widths provide perspective relative to viewpoints shown in **f**. Pearson's correlation between RNA_{SS} and W_{NI} increases with P_{NI} percentiles (inset, $n \in$ [2,278; 2,279] strains for each point). (**d**) Higher transcript abundance leads to slower growth in a protein-production-dependent manner. (**e**) A biphasic relationship between RNA abundance and protein production. Highly structured, poorly translated transcripts refractory to RNases accumulate in the cell. Highly translated transcripts protected by ribosomes accumulate less. (**f**) Three-dimensional phenotypic envelope of the library. Colored layers mark increasing data densities. Arrows at to right show viewpoints for **c**–**e** (n = 227,810 strains in all panels). See also **Supplementary Video 1**. (**g**) Codon usage modulates protein production for a given ribosomal density. Scatter plot of P_{NI} versus mean ribosomal density (MRD), colored by CAI as shown. Lines as detailed in **Figure 5b,e**. The correlation between CAI and P_{NI} increases with every percentile of ribosomal density (inset). (**h**) Polysome profiles link slower initiation, transcript accumulation, ribosome sequestration and physiological cost. Data were first binned by deciles of P_{NI} and then by deciles of W_{NI} to produce a coarse-grained version of **a**. In each bin, white lines and black shadings show the interquartile ranges of polysome fractions ($n \in$ [2,264; 2,272] strains for each bin). Red lines and points mark mean W_{NI} and corresponding row-wise optima. Background is color-coded by the apparent ribosome utilization (ARU).

RNA_{SS} when combined with strong $STR_{+31:+90}$ and preferably weak $STR_{+01:+60}$ (**Fig. 5b**, top). This arrangement of strong, non-overlapping structures also hinders translation initiation and consequent structural unwinding by ribosomes, and might be particularly effective at shielding the transcript (**Supplementary Fig. 11f**).

Translating ribosomes have long been suspected to sterically protect transcripts against RNases[48]. Efficiently translated transcripts are characterized by a weak $STR_{-30:+30}$ and an $STR_{+01:+60}$ that is either weak or competed by a stronger $STR_{+31:+90}$ (**Fig. 3c**). The same pattern drives high RNA_{PTX} (**Fig. 5b**, bottom). Further, RNA_{PTX} is highly correlated with P_{NI} (r = 0.76, **Supplementary Fig. 11f**), which supports the notion of a protection mechanism. But why then is high P_{NI} associated with only a modest increase in RNA_{SS} under exponential growth (**Fig. 5a,e**)? We hypothesize that this discrepancy

results from an artificial increase of the ribosome-to-transcript ratio during the degradation assay. In these conditions, fast-initiating transcripts become more coated by ribosomes than they would in a cell with steady amounts of competing transcripts (**Fig. 5a**, inset). This active ribosome repartitioning to intact transcripts confounds the interpretation of RNA_{HL} in our data (**Supplementary Fig. 11f–h**). It is also likely to confound all stability assays relying on transcriptional arrest, artificially inflating the reported half-life of highly translated transcripts.

Poorly initiated, overly stable mRNAs can trap ribosomes

The dependence of RNA_{SS} on P_{NI} shows non-monotonicity comparable to that of growth (**Fig. 5e**). For a given protein production, higher RNA_{SS} is always associated with larger growth defects, though the strength of this association decreases with P_{NI} (**Fig. 5c**,**d**). These interdependencies delineate a complex phenotypic envelope (**Fig. 5f**, **Supplementary Video 1** and **Supplementary Code 22**) wherein growth is modulated by the sheer cost of protein biogenesis and a cost linked to transcript abundance.

This mRNA burden is unlikely to be the direct cost of making or maintaining reporter transcripts. As variations in RNA_{SS} are largely ascribable to degradation rates, there is no need to invoke differential transcription rates and ensuing costs among strains. Furthermore, maximal growth is always achieved at intermediate rather than minimal levels of P_{NI} when strains with similar level of RNA_{SS} are compared (**Fig. 5c**, colored regression lines). This observation argues against deleterious effects of nucleotide sequestration *per se* and strongly suggests that the defect lies in the quality rather than the quantity of some transcripts.

Strikingly, the highest RNA_{SS} corresponded to the population of underachievers (**Fig. 5c**). High RNA_{SS} requires limitation of initiation by strong $STR_{-30:+30}$ (**Fig. 4g**, middle), transcript stabilization by noncompeting interactions with $STR_{+31:+90}$ (**Fig. 5b**, top) and limited ribosome availability (**Fig. 4g**, bottom). We hypothesize that highly structured, abundant (but untranslated) transcripts cause toxicity by sequestering single initiating—rather than several elongating—ribosomes. This would be the pathological manifestation of a general mechanism whereby initiation accounts for a non-negligible part of the translation cost. Alternatively, translation of some very short open reading frames has been shown to trigger growth defects through a termination defect[49]. We verified that the early terminated leader sequence determined by normal conditions of reporter initiation does not adversely affect growth (**Supplementary Fig. 5d**).

To directly measure ribosome sequestration, we used high-throughput targeted polysome profiling and obtained data for 240,403 strains (Online Methods, **Supplementary Fig. 12a** and **Supplementary Code 23**). Only one replicate was used in this experiment. We observed good correlation between the measured mean ribosome density per transcript (MRD) and P_{NI} ($r = 0.73$, **Fig. 5g** and **Supplementary Fig. 12b–f**).

To characterize polysome profiles, we binned the library into equally sized subpopulations within the protein-growth phenotypic space (**Fig. 5h** and **Supplementary Fig. 12g**). We found that gradual increases in P_{NI} reflect a progressive shift from low (initiation-limited) to high (elongation-limited) polysomal fractions (left–right rows). Strikingly, flat intermediate profiles marking initiation – elongation equilibrium always correspond to local fitness optima (red dots). For any given P_{NI}, increasing RNA_{SS} values skew profiles toward lower fractions (top-down columns), showing that lower initiation and loading of individual transcripts is associated with higher cost. The best growth improvements occur when balancing initiation and elongation in subpopulations with the highest RNA_{SS}, even if that leads to larger and thus more costly increases in protein production (bottom rows).

We calculated the apparent ribosome utilization (ARU = RNA_{SS} × MRD) as a static measure of ribosome sequestration. The ARU showed an inverse correlation with growth ($r = -0.46$) that was maximal in underachievers (**Fig. 5h**, colored background). This suggests that slow-initiating ribosomes remain unproductively bound to transcripts for at least minutes. During that time, they are sequestered from the transcriptome more effectively than they would be when queuing on better initiated but poorly elongated transcripts (**Supplementary Fig. 12h**).

DISCUSSION

We applied a formal, factorial design of experiments to systematically test the contributions of eight separate molecular properties to the efficacy of translation of bacterial transcripts. Large-scale DNA synthesis allowed us to design multiple replication levels, ranging from a few mutations in sequences of a given combination of properties to full design replication in 56 regions of sequence space. This enabled us to quantify effects on translation in as unbiased a way as is currently possible (**Supplementary Fig. 2**). An epigenetic structure controller in our expression system enabled us to modulate one of the dominant parameters—the transcript secondary structure in the translation initiation region (TIR)—experimentally, so that the effects of other factors could be better estimated. To our knowledge, this designed, factorial library and associated datasets represent the most comprehensive exploration of bacterial translation to date.

Several of our main findings are negative. Over our entire dataset, we were able to explain just over half (53%) of the total variance in protein production via our design parameters and errors in measurements or designs themselves. While this fraction varied among mutational series (36–70%), the lack of explanation points to fundamental mechanisms yet to be discovered. Though there are known unaccounted-for properties in these designs—e.g., dicodon usage, which is difficult to manipulate exhaustively—these seem insufficient to explain all of the inability to predict variance. A few exceptions aside—for example, a double proline in one series (**Supplementary Fig. 9**)—we found little evidence for mechanisms to explain the systematic phenotypic differences observed between series (**Supplementary Fig. 13**). Despite using advanced statistical techniques to link genetic variations to particular phenotypes[50], we could not identify generic properties that increased our explanatory power.

We did find that properties such as the codon ramp and amino acid charge or hydropathy that have been documented in studies of natural transcripts[10,18,19] had very small effects in our study (**Supplementary Figs. 7** and **12**). The functional importance of these signals might depend on features not present in our synthetic transcripts or be confounded by other parameters not well controlled in natural transcripts.

Transcript structure and codon usage both had significant effects that were detected by our large-scale measurements of protein production, mRNA abundance and stability; the distribution of ribosomes per transcript; and bacterial growth rate. Each of these measurements shed light on different aspects of translation cost and efficiency that are mediated by these features (**Fig. 6a**). In common with other studies, we found that secondary structure, over the start codon ($STR_{-30:+30}$) as well as in the gene ($STR_{+01:+60}$, $STR_{+31:+90}$), can affect total protein production[14,21,36], abundance and behavior of transcripts[40], as well as the ensuing physiological costs[5]. The effect of codon usage is mediated mainly by the configuration of these

structures. From combinations of these parameters, we identified six sequence classes that led to different efficiency–cost outcomes (**Fig. 6** and **Supplementary Data 35**). We marked certain members as representative of each of these scenarios (**Supplementary Data 15** and **Supplementary Code 24**) and hypothesize that these, and related sequences, are good targets for further studies of ribosomal recruitment, elongation and release.

Sequences with mild TIR structure are useful as references for translation efficiency (**Fig. 6b**). They can produce ample protein if ribosomes are numerous enough to support the next initiation before a TIR structure opened in the preceding event can refold (ribosome cooperativity). Absence of downstream structures and regular ribosome flow ensure basal transcript decay rates. The physiological burden corresponds to the material cost of moderate protein biosynthesis, which decreases when nutrients are limited and fewer proteins overall are produced (C_M in **Fig. 6b** and **Fig. 4g**, bottom).

When weak structures in the transcript are combined with high CAI (**Fig. 6c**, top), fast initiation and unobstructed elongation lead to increased protein production, with dense but smooth ribosome trafficking (medium MRD) and consequent transcript protection (medium RNA_{SS}). Fast turnover ensures ribosome redistribution toward other transcripts, leading to moderate utilization cost despite sizable ARU. When CAI is poor (**Fig. 6c**, bottom), elongation can be limited, leading to greater ribosome retention and cost to the cell since more essential transcripts are thereby less translated[14]. We found only a small positive association between CAI and growth (**Fig. 4c**,**f** and **Supplementary Fig. 10e**). Although improving CAI slightly decreases MRD (**Fig. 5g**) and associated cost of ribosomal utilization, it also augments RNA_{PTX}, RNA_{SS} and $P_{NI} - P_{FI}$ (**Supplementary Figs. 12b**,**c**, **6g** and **7f**)—which ultimately results in costs for increased protein production and, perhaps, slower initiation of individual transcript (lower cooperation). The signal relating CAI to fitness might thus be blurred by nearly compensatory transactions between the costs associated to these mechanisms (**Fig. 6c**, right).

Structure within transcripts can affect initiation and elongation, and potential structural regions within a transcript can interact with each other, creating complex subpopulations of transcripts with different translational properties. Initiation is generally regarded as the rate-limiting step in translation[2,51–53], and our use of an epigenetic structure controller proves this. Notwithstanding an invariant SD_R sequence (**Fig. 1d**), secondary structures in the TIR ($STR_{-30:+30}$ and to a lower extent $STR_{+01:+60}$) have by far the largest effect on translation efficiency (**Figs. 2** and **3d**). However, the precise effect of TIR structure on productivity (**Fig. 3a–c**), transcript stability (**Fig. 5b**) and cost (**Supplementary Fig. 10f**,**g**) depends on complex interactions with other structures in the transcript.

Pre-initiating ribosomes bound to transcripts with a stably folded TIR must actively accommodate TIR structure to initiate translation, in a process that may be favored[54] or stalled[55] by other regulators. Transcripts in this state are likely protected from RNases around the TIR and exposed to degradation further downstream if not folded there (**Fig. 6d**). This configuration determines moderate transcript accumulation and medium ARU, which may hinder timely ribosome cooperation. Considering weak protein production, initiation-based ribosome sequestration determines most of the small fitness cost measured under normal growth condition. Indeed, that cost is markedly aggravated when ribosomes are more limiting (C_M; **Fig. 4g**).

Accommodation can be thermodynamically facilitated in the presence of stronger overlapping structures downstream (**Figs. 3b**,**c** and **6e**). Subsequent initiation improvements may, for example, be mitigated by slower elongation through distal structure(s). Such

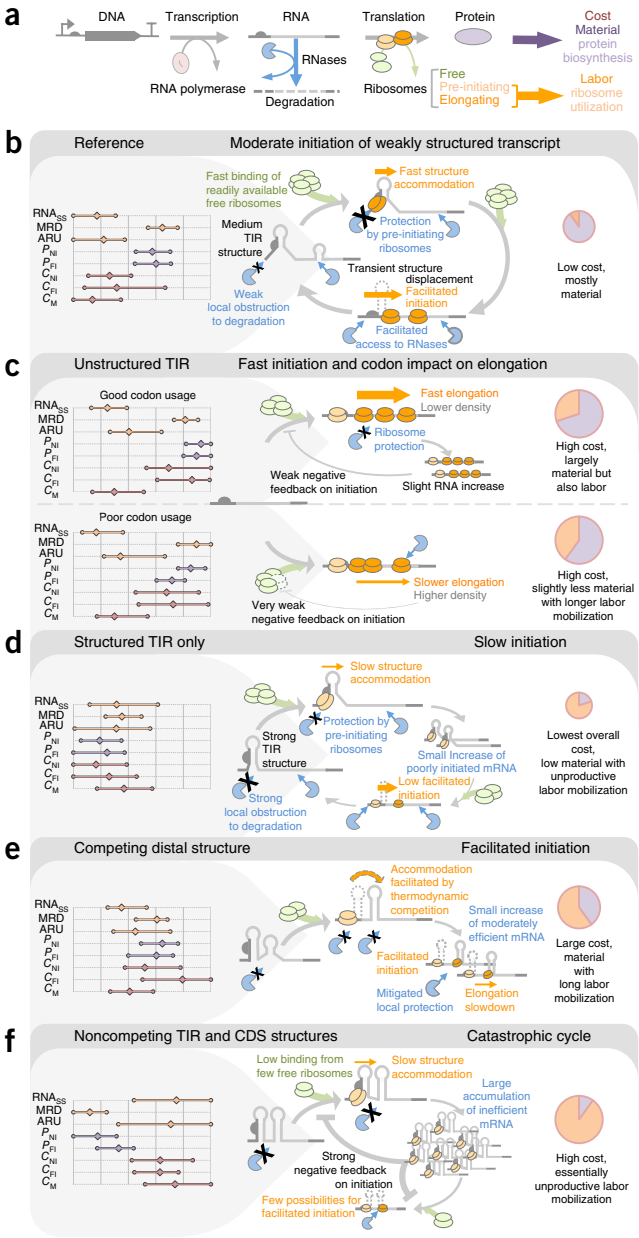


Figure 6 Impact of archetypal sequences on translation efficiency and physiological cost. (**a**) Key processes in the central dogma of molecular biology. This diagram introduces graphical elements used in the schematics below. (**b**–**f**) Combinations of sequence properties and consequent variations in translation efficiency determine diverse physiological costs. Left charts show quartiles of various measured phenotypic averages across 56 factorial series; the costs C_{NC}, C_C and C_M are the opposite of W_{NI}, W_{FI} and W_M, respectively. Data are subsets by combinations of categorical properties used for design as follows. (**b**) Medium $STR_{-30:+30} \cap$ weak $STR_{+01:+60} \cap$ weak $STR_{+31:+90}$ (n = 9,039 strains). (**c**) Top: weak $STR_{-30:+30} \cap$ weak $STR_{+01:+60} \cap$ high CAI (n = 9,037 strains); bottom: weak $STR_{-30:+30} \cap$ weak $STR_{+01:+60} \cap$ weak CAI (n = 9,036 strains). (**d**) Strong $STR_{-30:+30} \cap$ medium $STR_{+01:+60} \cap$ weak $STR_{+31:+90}$ (n = 9,039 strains). (**e**) Medium $STR_{-30:+30} \cap$ medium $STR_{+01:+60} \cap$ strong $STR_{+31:+90}$ (n = 9,040 strains). (**f**) Strong $STR_{-30:+30} \cap$ weak $STR_{+01:+60} \cap$ strong $STR_{+31:+90}$ (n = 9,039 strains). Schematic models in the middle provide a mechanistic framework to explain these data. Physiological consequences are rendered as schematic pie charts on the right, the sizes of which convey the magnitude of the overall cost, while the slices differentiate material (violet) and labor (orange) contributions. All aspects of these models are supported by experimental measurements.

structured transcripts accumulate and collectively yield sizeable protein production at a relatively high cost due to high ARU.

The most detrimental architecture we observed was a combination of strong TIR structure and strong non-overlapping distal structure (**Fig. 6f**). Individually, these properties result in slow initiation (**Fig. 2**) and increased transcript stability (**Supplementary Fig. 11h**). Combined, their effects define a self-reinforcing catastrophic cycle. The rarity of translation events (lowest MRD) minimizes structure unfolding, thus ensuring maximal transcript stabilization (highest RNA_{SS}). Accumulating transcripts then sequester pre-initiating ribosomes (largest ARU), further limiting ribosome cooperation (**Fig. 5h** and **Supplementary Fig. 12d–f**). This leads to a futile cycle of increased unproductive ribosome sequestration and decreased translational capacity for not only this transcript, but all transcripts in the cell. Subtle parameter variations in this extreme scenario account for the continuum of phenotypes observed in our experiments. In fact, the depth of our design allowed us to uncover considerable flexibility regarding the position and size of structures influencing initiation in the larger TIR (**Fig. 3d**), though SD accessibility is a major factor (**Fig. 3e**).

We could not identify structural descriptors capable of improving our explanations of the whole library (**Fig. 3f**). In fact, we found that unrelated sequences with largely similar structural profiles *in silico* can yield vastly different phenotypic responses (**Supplementary Fig. 14**). We speculate that this is due to a poor ability to predict nucleic acid structures and their dynamics accurately *in vivo*[34] because they may form hard-to-compute tertiary arrangements and interact with other *cis* structures and *trans*-acting molecules (for example, ribosomes[56]). Although technologies such as ribosome profiling are yielding insights into procession of ribosomes on transcripts, and techniques such as SHAPE-SEQ[57] and Structure-SEQ[58] are beginning to yield insight into average *in vivo* transcript structure, an improved ability to monitor single RNA structures *in vivo* would result in breakthroughs in understanding the mechanisms underlying much unexplained variance in expression and consequent physiological behaviors. Methods to account for the sequestration of translational resources, resulting decrease in translation of adaptive genes, potential lock-up of nucleic acids and amino acids in useless molecules and associated stress responses are also needed to fully interpret our data.

Our study used a precisely-designed library to test specific sequence-function hypotheses. There are costs and benefits of such highly-constrained libraries which were evident in our design around factors affecting translational efficiency and its fitness impact. Designed synthetic libraries are well suited to testing known possible mechanisms and enabling robust quantification of effect sizes. However, this framework—with its sparse and non-uniformly distributed mutants in sequence space—provides limited opportunities to uncover novel mechanisms. It is possible that unknown genetic signals are more readily detected in a vast collection of natural or randomized sequences, which are diverse enough to break genetic linkage between functional motifs and surrounding sequences. However, sample sizes would need to be huge (tens of thousands of individuals) to minimize statistical biases. Although single genomes might not yield large enough datasets for robust analyses, combining results from many species or taxa might be plagued by species-specific idiosyncrasies. Such combinations should be carefully chosen.

Ideally, both designed and natural sequence sets should be used to define a virtuous knowledge building cycle. As sequence repositories continue to grow, comparative analyses may lead to better models for how to build controlled libraries that are more representative of natural systems and more informative by design.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank V. Mutalik, C. Liu, L. Jacob, M. Price, A. Deutschbauer, M. Samoilov, P. Shah, J. Plotkin, J. Savitskaya and L. Ciandrini for discussions. We are grateful to the Agilent Laboratories and the Synthetic Biology Institute (SBI) for providing the OLS array. We thank J. Sampson, P. Anderson and S. Laderman from Agilent Laboratories for discussing OLS setup and processing. G.C. was funded by the Human Frontier Science Program (LT000873/2011-l), J.C.G. by the Portuguese Fundação para a Ciência e Tecnologia (SFRH/BD/47819/2008). We acknowledge financial support by the Synthetic Biology Engineering Research Center (SynBERC under National Science Foundation grant 04-570/0540879). This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley (NIH S10 Instrumentation Grants S10RR029668 and S10RR027303).

AUTHOR CONTRIBUTIONS

G.C. and A.P.A. conceived the work; G.C. and J.C.G. designed sequences; G.C. performed experiments and processed data; G.C. and A.P.A. analyzed the data and J.C.G. contributed *post hoc* secondary structure analyses; G.C. and A.P.A. wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

1. Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J.S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157**, 624–635 (2014).
2. Andersson, S.G. & Kurland, C.G. Codon preferences in free-living microorganisms. *Microbiol. Rev.* **54**, 198–210 (1990).
3. Scott, M., Gunderson, C.W., Mateescu, E.M., Zhang, Z. & Hwa, T. Interdependence of cell growth and gene expression: origins and consequences. *Science* **330**, 1099–1102 (2010).
4. Ceroni, F., Algar, R., Stan, G.-B. & Ellis, T. Quantifying cellular capacity identifies gene expression designs with reduced burden. *Nat. Methods* **12**, 415–418 (2015).
5. Frumkin, I. *et al.* Gene architectures that minimize cost of gene expression. *Mol. Cell* **65**, 142–153 (2017).
6. Ikemura, T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.* **151**, 389–409 (1981).
7. Sharp, P.M. & Li, W.H. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
8. Cannarozzi, G.M. & Schneider, A. *Codon Evolution* (Oxford Univ. Press, 2012).
9. Mitarai, N., Sneppen, K. & Pedersen, S. Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization. *J. Mol. Biol.* **382**, 236–245 (2008).
10. Charneski, C.A. & Hurst, L.D. Positively charged residues are the major determinants of ribosomal velocity. *PLoS Biol.* **11**, e1001508 (2013).
11. Pop, C. *et al.* Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol. Syst. Biol.* **10**, 770 (2014).
12. Del Campo, C., Bartholomäus, A., Fedyunin, I. & Ignatova, Z. Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function. *PLoS Genet.* **11**, e1005613 (2015).
13. Adhin, M.R. & van Duin, J. Scanning model for translational reinitiation in eubacteria. *J. Mol. Biol.* **213**, 811–818 (1990).
14. Kudla, G., Murray, A.W., Tollervey, D. & Plotkin, J.B. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255–258 (2009).
15. Mutalik, V.K. *et al.* Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10**, 354–360 (2013).
16. Espah Borujeni, A., Channarasappa, A.S. & Salis, H.M. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res.* **42**, 2646–2659 (2014).
17. Tuller, T. & Zur, H. Multiple roles of the coding sequence 5′ end in gene expression regulation. *Nucleic Acids Res.* **43**, 13–28 (2015).
18. Tuller, T. *et al.* An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344–354 (2010).

19. Tuller, T. *et al.* Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol.* **12**, R110 (2011).
20. Charneski, C.A. & Hurst, L.D. Positive charge loading at protein termini is due to membrane protein topology, not a translational ramp. *Mol. Biol. Evol.* **31**, 70–84 (2014).
21. Goodman, D.B., Church, G.M. & Kosuri, S. Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342**, 475–479 (2013).
22. Allert, M., Cox, J.C. & Hellinga, H.W. Multifactorial determinants of protein expression in prokaryotic open reading frames. *J. Mol. Biol.* **402**, 905–918 (2010).
23. Ilzarbe, L., Álvarez, M.J., Viles, E. & Tanco, M. Practical applications of design of experiments in the field of engineering: a bibliographical review. *Qual. Reliab. Eng. Int.* **24**, 417–428 (2008).
24. Montgomery, D.C. *Design and Analysis of Experiments* (Wiley, 2017).
25. Zhou, H., Vonk, B., Roubos, J.A., Bovenberg, R.A.L. & Voigt, C.A. Algorithmic co-optimization of genetic constructs and growth conditions: application to 6-ACA, a potential nylon-6 precursor. *Nucleic Acids Res.* **43**, 10560–10570 (2015).
26. Zhang, C., Zou, R., Chen, X., Stephanopoulos, G. & Too, H.-P. Experimental design-aided systematic pathway optimization of glucose uptake and deoxyxylulose phosphate pathway for improved amorphadiene production. *Appl. Microbiol. Biotechnol.* **99**, 3825–3837 (2015).
27. Mutalik, V.K. *et al.* Quantitative estimation of activity and quality for collections of functional genetic elements. *Nat. Methods* **10**, 347–353 (2013).
28. Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **110**, 14024–14029 (2013).
29. Taniguchi, Y. *et al.* Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).
30. Guimaraes, J.C., Rocha, M., Arkin, A.P. & Cambray, G. D-Tailor: automated analysis and design of DNA sequences. *Bioinformatics* **30**, 1087–1094 (2014).
31. Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T.C. & Waldo, G.S. Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* **24**, 79–88 (2006).
32. Young, T.S., Ahmad, I., Yin, J.A. & Schultz, P.G. An enhanced system for unnatural amino acid mutagenesis in *E. coli*. *J. Mol. Biol.* **395**, 361–374 (2010).
33. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).
34. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J.S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705 (2014).
35. Yoo, J.-H. & RajBhandary, U.L. Requirements for translation re-initiation in *Escherichia coli*: roles of initiator tRNA and initiation factors IF2 and IF3. *Mol. Microbiol.* **67**, 1012–1026 (2008).
36. Kelsic, E.D. *et al.* RNA structural determinants of optimal codons revealed by MAGE-seq. *Cell Syst.* **3**, 563–571.e6 (2016).
37. 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).
38. Hilterbrand, A., Saelens, J. & Putonti, C. CBDB: the codon bias database. *BMC Bioinformatics* **13**, 62 (2012).
39. Dana, A. & Tuller, T. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res.* **42**, 9171–9181 (2014).
40. Boël, G. *et al.* Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature* **529**, 358–363 (2016).
41. Chevance, F.F.V., Le Guyon, S. & Hughes, K.T. The effects of codon context on in vivo translation speed. *PLoS Genet.* **10**, e1004392 (2014).
42. van Opijnen, T. & Camilli, A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat. Rev. Microbiol.* **11**, 435–442 (2013).
43. Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**, 588–592 (2005).
44. Schaechter, M., MaalOe, O. & Kjeldgaard, N.O. Dependency on medium and temperature of cell size and chemical composition during balanced growth of *Salmonellatyphimurium*. *J. Gen. Microbiol.* **19**, 592–606 (1958).
45. Li, G.-W.Howdo bacteria tune translation efficiency? *Curr. Opin. Microbiol.* **24**, 66–71 (2015).
46. Deana, A. & Belasco, J.G. Lost in translation: the influence of ribosomes on bacterial mRNA decay. *Genes Dev.* **19**, 2526–2533 (2005).
47. Presnyak, V. *et al.* Codon optimality is a major determinant of mRNA stability. *Cell* **160**, 1111–1124 (2015).
48. Hui, M.P., Foley, P.L. & Belasco, J.G. Messenger RNA degradation in bacterial cells. *Annu. Rev. Genet.* **48**, 537–559 (2014).
49. Dinçbas, V. & Heurgué-Hamard, V. Shutdown in protein synthesis due to the expression of mini-genes in bacteria. *J. Mol. Biol.* **291**, 745–759 (1999).
50. Jaillard, M. *et al.* A fast and agnostic method for bacterial genome-wide association studies: bridging the gap between kmers and genetic events. Preprint at *bioRxiv* https://doi.org/10.1101/297754 (2018).
51. Bulmer, M. The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**, 897–907 (1991).
52. Shah, P., Ding, Y., Niemczyk, M., Kudla, G. & Plotkin, J.B. Rate-limiting steps in yeast protein translation. *Cell* **153**, 1589–1601 (2013).
53. Ciandrini, L., Stansfield, I. & Romano, M.C. Ribosome traffic on mRNAs maps to gene ontology: genome-wide quantification of translation initiation rates and polysome size regulation. *PLoS Comput. Biol.* **9**, e1002866 (2013).
54. Duval, M. *et al. Escherichia coli* ribosomal protein S1 unfolds structured mRNAs onto the ribosome for active translation initiation. *PLoS Biol.* **11**, e1001731 (2013).
55. Marzi, S. *et al.* Structured mRNAs regulate translation initiation by binding to the platform of the ribosome. *Cell* **130**, 1019–1031 (2007).
56. Qu, X. *et al.* The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature* **475**, 118–121 (2011).
57. Takahashi, M.K. *et al.* Using in-cell SHAPE-Seq and simulations to probe structure-function design principles of RNA transcriptional regulators. *RNA* **22**, 920–933 (2016).
58. Ding, Y., Kwok, C.K., Tang, Y., Bevilacqua, P.C. & Assmann, S.M. Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq. *Nat. Protoc.* **10**, 1050–1066 (2015).

ONLINE METHODS

Genomic analysis. We used the genome of *Escherichia coli* str. K-12 substr. MG1655 (GenBank U00096.2) as a reference for all design analyses. Annotated CDSs and the preceding 99 nt were extracted as a multifasta file using Biopython (**Supplementary Code 25**). Only CDSs longer than 99 nt were considered. This file was used as input for the batch sequence analysis mode of D-Tailor[30], for which we developed extensions to explore and calculate the sequence properties of interests (**Supplementary Code 26**).

Ten shuffled versions of each wild-type gene were generated under constraints to maintain UTR nucleotide composition and amino acid sequence (**Supplementary Code 27** and **Supplementary Data 36**). These were used as a null reference set for comparison purposes when appropriate (for example, **Supplementary Fig. 1k**).

We used functional genomic data from Taniguchi *et al.*[29] for preliminary correlation analyses between calculated sequence properties and mRNA and protein abundances (**Supplementary Fig. 1**).

Coding sequence design. We used the design mode of D-Tailor[30] to derive factorial series from input seed sequences (**Supplementary Code 26**). 56 seed sequences of 96 nt were randomly generated and evolved using a simple Monte Carlo simulation to maximize the pairwise Hamming distance between sequences within the set (**Supplementary Code 28**). Seed sequences were then considered in the context of our expression system and subjected to mutation to generate all combinations of properties. Mutational processes used to generate variant sequences were first constrained to favor synonymous mutations. To ease completion of the full factorial designs, that constraint was relaxed when the discovery rate of new property combinations became too low and at least two-thirds of a series was completed. For some seeds, the discovery rate stalled rapidly during the design process, so that deriving a full factorial set was impractical. In such case, the seed and derived set of sequences was discarded. A new random seed was generated, evolved for maximal distance to other seeds (**Supplementary Code 27**) and submitted to diversification by D-Tailor.

For a given seed, the design process stopped once all combinations of properties (1,458) were obtained at least once. Multiple sequence solutions were found for some combinations of properties. We used a simple heuristic procedure to select a full factorial set minimizing the average pairwise sequence distances within the series and thus to obtain a very local cluster in sequence space (**Supplementary Fig. 2c** and **Supplementary Code 26**).

For each sequence in a given series, we then derived two variants by introducing between one and four mutations at random while maintaining the original combination of sequence properties. We used this replication level to calculate the design error (**Fig. 2a** and **Supplementary Code 7**). Due to space limitations on the synthesis chip, we did not derive a third variant for 944 sequences of seed no. 136, which thus comprises only 3,430 sequences. Each factorial series but one thus contains 4,374 sequences. Altogether, the final library comprises exactly 244,000 sequences.

Throughout the design process, we instructed D-Tailor to rejected sequences containing useful restriction sites, potential promoters, terminators or internal ribosome binding sites.

D-Tailor modules implemented to analyze and evolve the sequence properties of interest are listed in **Supplementary Code 26** and available on the project's GitHub repository (https://github.com/jcg/d-tailor).

Oligonucleotide design and synthesis. Designed sequences (96 nt) were prefixed and suffixed by 29- and 26-nt-long appendices, respectively. These contain constant sequence context including restriction sites and 24-nt priming sites necessary for PCR amplification upon synthesis. We used three orthogonal pairs of priming sites so that three subsets of the library could be amplified independently. Two of these subsets (15,811 and 15,831 sequences, respectively) were defined so that each 20-mer in the synthetic sequences could be uniquely mapped to a particular construct within each subset. This was originally intended to enable ribosome profiling in these subsets, although we have not pursued this application in this work. The third subset contains the remaining sequences (212,358).

The 151-nt-long sequences resulting from this procedure have the following generic form:

FFFFFFFFFFFFFFFFGGTACCATAATGNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGGTGGA TCCRRRRRRRRRRRRRRRRR where N correspond to the design sequence and F and R to the variable part of the forward and reverse priming sequence, respectively. Restriction sites for cloning of the synthetic sequences (KpnI and BamHI) are shown in bold. The sequence targeted by PCR primers is underlined.

Priming sequences are listed in **Supplementary Table 1**.

Sequences were named with unique identifiers resulting from concatenating series number, properties level string and replicate number (series_level_replicate). Sequences were then written in a multifasta file and sent to Agilent Laboratories for synthesis. The 244,000 sequences were synthesized as single-stranded DNA oligonucleotides on a single high-density array using the OLS technology[59]. The procedure generated ~90 ng of full-length synthetic DNA as quantified on a 2100 Bioanalyzer (Agilent) and sent to us as pool in a single tube (~3 ng/μL).

Plasmid description. *Accessory plasmid.* We modified pEVOLPROBE-aaRS (a gift from C. Liu, University of California, Irvine), which comprises a pBBR1 origin of replication and kanamycin resistance cassette derived from pBROBE[60] and an orthogonal tRNA/aminoacyl-tRNA synthetase system derived from pEVOL[32]. pEVOL permits one to express an optimized $tRNA_{CUA}$ and two copies of a cognate aminoacyl-tRNA synthetase (aaRS) gene evolved to selectively discriminate the unnatural amino acid *p*-acetylphenylalanine (pAcF). One of the aaRS copies is controlled by a pBAD promoter, which is repressed by the product of the *araC* gene, also encoded on the plasmid. Upon addition of 0.1% arabinose (Sigma) and 10 mM pAcF (Synchem), $tRNA_{CUA}$ are effectively charged with pAcF and outcompete RF1 for the decoding of the amber stop codon (UAG), resulting in the incorporation of the unnatural amino acid into the polypeptide chain and continuation of elongation[32].

We amplified a chloramphenicol resistance cassette from pBbE1c-RFP[61] and cloned it in place of the original kanamycin resistance cassette (NotI/XbaI) to yield plasmid pGC4510.

We amplified the *tev* protease gene from pRK603[62] and cloned it into pBbA2k[61] (EcoRI/BamHI), yielding pGC2109.

A fragment containing oppositely oriented *tev* and *tetR* genes driven by a bidirectional pTet promoter was amplified from pGC2109 and cloned in pGC4510 (NotI), to yield the accessory plasmid pGC4593 (Addgene https://www.addgene.org/110206/110206). Addition of 200 nM anhydrotetracycline (aTc) permits the expression of the TEV protease, which recognizes and cuts a specific polypeptide motif[63].

Reporter plasmid. The reporter plasmid was built from a pFAB512 backbone[64] that comprises a standard *sfGFP* expression cassette, a p15a origin of replication and a kanamycin resistance cassette. To increase the sensitivity of fluorescence measurements at weak protein production, p15a was replaced origin of replication with a higher copy number (ColE1, ~30 plasmids per cell) from plasmid pBbE1c-RFP[61] (AvrII/EcoRI). To ensure strong transcriptional repression of the reporter, we inserted the *lacI* gene amplified from *E. coli* BW25113 into the plasmid (BglII/XbaI) and modified its promoter in the process to reconstruct the highly expressed $LacI^Q$ variant[65]. We amplified *mRFP1* followed by the rnaT1 terminator from pFAB809[64] and cloned it as a transcriptional fusion to the $lacI^Q$ gene (EcoRI/XbaI). We then randomly mutated the RBS associated with *mRFP1* using inverse PCR (iPCR) to tune its expression to a low but clearly detectable level. We used this fluorescent reporter to control for extrinsic noise during flow cytometry applications[28,66,67]. This whole procedure yielded pGC4724.

We iteratively engineered the reporter system using pFAB512 as an initial template:

1. We replaced the original *sfGFP* promoter with a strong synthetic promoter tightly repressed by a lac operator[15]. To that end, we introduced two BsaI site by iPCR of the whole backbone (excluding the native promoter) and cloned the repressible promoter in the form of annealed oligonucleotides with matching overhangs.

2. We introduced a linker sequence at the very beginning of the reporter through another iPCR. That sequence comprises a unique BamHI restriction site, a flexible 3xGGS linker, a 6×His tag, and a specific cleavage motif for the TeV protease[63].

doi:10.1038/nbt.4238

3. We used iPCR to insert a 57-nt-long leader sequence, the stop codon of which overlaps the start codon of the reporter and further introduces a perfect SD sequence in the reporter translation initiation region[15], as well as a unique KpnI restriction site.

4. We used site-directed mutagenesis to introduce amber stop codons at various locations in the leader sequence. On the basis of experimental results, we selected the sequence with TAG at the fifth codon position as the one with the most efficient coupling properties (highest reporter protein production in coupling conditions, with low readthrough in noncoupling conditions; **Supplementary Fig. 5a**).

5. We modified the codon usage of *sfGFP*. The goal of these modifications was to introduce a translation bottleneck near the end of the gene's coding sequence. A bottleneck is a region where the elongation rate is expected to be slower than that of the rest of the coding sequence, as estimated using a tAI-based profile[18]. The strongest bottleneck is located at the beginning of the original *sfGFP* nucleotide sequence (**Supplementary Fig. 1i**, gray line), which did not allow us to test the effect of a bottleneck at the end of the gene, as intended with the Btl_P property. We thus used three iPCRs to remove two putative bottlenecks at the beginning of the gene and strengthen a bottleneck at the end (position 232; **Supplementary Fig. 1i**, green line). This iterative process introduced a total of 22 codons change, and we checked that the successive changes did not dramatically affect expression. The strength of the C-terminal bottleneck was set to a moderately high value so that both stronger and weaker bottlenecks could be easily encoded in the N-terminal variable region, thereby enabling implementation of the Btl_P design (**Supplementary Fig. 1j**).

6. We amplified a *ccdB* expression cassette from the gateway plasmid pDONR221 (Invitrogen) and cloned it between KpnI and BamHI.

This whole procedure yielded pGC4742, a counterselectable acceptor plasmid for the reporter system. To obtain the final reporter plasmid pGC4750, we subcloned the engineered reporter system from pGC4742 into pGC4691 (BglII/AvrII). pGC4750 is propagated in the *E. coli* strain *ccdB* Survival 2 (Invitrogen). Cloned versions of pGC4750 isolated from the library have been made available: pGC4780 expresses high GFP (Addgene 113408) and pGC4785 expresses low GFP but is highly susceptible to facilitated initiation through translation coupling (Addgene 113409).

Library construction. Aliquots of 1.5 ng of the oligonucleotide pool were amplified separately with each of three primer pairs using Phusion DNA polymerase (NEB) (15 cycles, four tubes of 100 µL final volume for a given reaction). Amplicons were separated by electrophoresis in a 4% agarose gel (NuSieve GTG, Lonza). We excised the bands at the expected size (151 bp), as well as larger, fuzzier bands located just below the 200 bp marker; recovered the DNA on a column (Zymoclean Gel DNA Recovery Kit, Zymo Research); and eluted in water. Cleaned PCR products were digested overnight with KpnI-HF and BamHI-HF (NEB). Upon digestion, DNA from the low and high bands had the same size. The reporter plasmid pGC4750 was miniprepped (Qiagen) and digested overnight with KpnI-HF and BamHI-HF (NEB). The cut vector was resolved from the *ccdB* insert by electrophoresis in a 0.8% agarose gel and cleaned as above. Sanger sequencing of clones obtained from initial cloning tests shown that inserts from both isolated bands were equally good. For each subset of the library, we thus pooled the two extracts. For the library cloning, about 50 nmol of digested inserts and vector were mixed at a 1:1 molar ratio, as quantified using the Qubit dsDNA HS fluorometric assay (Life Technologies), and ligated overnight with T4 DNA ligase (NEB).

Ligation products were dialyzed and electroporated into the *E. coli* MDS42 *recA* strain (Scarab Genomics) containing the accessory plasmid pGC4593 (see "*E. coli* host strain" below). Upon electroporation, cells were recovered at 37 °C for 1 h without shaking, plated on large square LB agar dishes (245 × 245 mm) supplemented with kanamycin and chloramphenicol for plasmid selection, and grown overnight at 37 °C. In parallel, serial dilutions of the electroporated cells were plated on small round agar plates (60 mm) to estimate transformation efficiency. Transformants were scraped, resuspended in rich MOPS medium supplemented with kanamycin and chloramphenicol, and homogenized by shaking at 900 r.p.m. (37 °C) for 1 h. Aliquots of 50 µL were then mixed with glycerol (15% final) and frozen at −80 °C. We estimated from plating of dilution series that the three library subsets contained 1.25×10^7, 0.51×10^7 and 1.21×10^7 individual clones, representing 791-, 322- and 57-fold coverage of the respective library subset (15,811, 15,831 and 212,358 sequences, respectively). Accordingly, we mixed the three homogenized cultures at 10:10:134 volumetric ratios and saved aliquots of this final library.

The frequency of individual clones varied over a 30-fold range within the library. Such a bias is not unusual in this kind of synthetic library and might originate from several causes that are difficult to fully disentangle. In the first place, synthesis might be more successful for some sequences than others. About a third of the sequence reads obtained upon high-throughput sequencing of the final library contained mutations, with a majority of them being small deletions that are typical of failed coupling steps during synthetis[17]. Such mutants are excluded from all analyses presented in this work. Further analysis might pinpoint particular sequence features that are more prone to such error. Notwithstanding synthesis errors, we estimated on the basis of the total quantity of DNA recovered upon synthesis that no more than 5 million copies (7.6 amol) of each sequence were produced. The PCR step necessary to amplify that pool and produce dsDNA may introduce further biases. As PCR is mandatory to the preparation of sequencing libraries, it is difficult to obtain a clear quantification of potential PCR-related biases. The cloning procedure may also introduce biases, as some sequence features might lead to less efficient annealing and ligation. Cursory analysis did not reveal any obvious such determinants. Finally, biases could arise from differential strain growth upon transformation. In our case, coverage bias most likely reflects various construction biases, rather than deleterious effect of the cloned sequence, because transcription of the reporter was never induced by addition of IPTG at this stage.

E. coli host strain. Early tests of our reporter system were carried out in *E. coli* strain BW25113. At this stage, we repeatedly obtained an IS insertion in the $tRNA_{CUA}$ gene on the accessory plasmid pGC4593. In fact, tRNAs are known insertional hotspots for various mobile elements. That this particular event was repeatedly selected reflects the pressure set on the cells by forcing high translation of our reporter. This prompted us to carry out all our experiments in MDS42, a genome-reduced strain that has been stripped of insertion sequences, pseudogenes and other dispensable regions. In all, 14.30% of the genome has been deleted from the parent MG1655 without adverse effect on fitness[68,69]. As far as we are aware, the strain does not show any peculiar behavior in term of expression. Importantly, tRNA and ribosomal genes have not been altered. Most of the removed genes are part of the accessory *E. coli* genome, which tends to be particularly mobile and consequently shows a disparate nucleotide and codon usage. The codon composition of the genome-reduced strain is thus minimally altered. MDS42 achieves high transformation rates by electroporation and features dramatic improvement of genetic stability. These two desirable traits explain why the strain has been widely adopted for biotechnological applications. We used a *recA*-deleted version of the strain to avoid any unwanted recombination and further improve the stability of the library.

Measurement of protein production. *Growth conditions.* Cells were grown overnight at 37 °C in 5 mL Rich MOPS medium (Teknova) supplemented with kanamycin and chloramphenicol (for plasmid maintenance), IPTG (for induction of reporter transcription) and aTc (for induction of *tev* transcription). In coupling condition, the growth medium was further complemented with arabinose (for induction of unnatural aminoacyl-tRNA synthetase transcription) and the unnatural amino acid pAcF (2.5 nM). Since addition of pAcF results in acidification of the medium, the pH was adjusted to its initial value of 7.2 by addition of NaOH.

Low-throughput measurement of a reference panel. During the construction of the library, individual colonies were randomly picked, independently cultured in 96-well plates and Sanger sequenced. This procedure yielded a small subpanel of the design library comprising 310 clearly identified separate strains. Upon growth as described above, single-cell green and red fluorescence intensities were measured using an automated Guava EasyCyte flow cytometer (EMD Millipore, Hayward, CA, USA) and processed using previously described workflows[64]. All strains were measured at least in triplicate. Although derived from a lower-precision instrument, these data were used to

benchmark the results from the high-throughput procedure described below (**Supplementary Figs. 6a** and **7a**).

High-throughput measurement of the whole library by fluorescence-activated cell sorting (FACS)-seq. High-throughput measurements involve sorting cells into different bins of fluorescence efficiency. Subpopulations thus obtained are barcoded and submitted to targeted sequencing. For each strain in the library, processing of the resulting reads permits reconstruction of a coarse-grained fluorescence distribution, the resolution of which depends on the number of sorting bins (here 16). We replicated this procedure four times with fully independent biological samples of the library. Details of the procedure are presented below.

Sorting of the library into fluorescence bins. In both coupling and noncoupling conditions, replicates 1 and 2 were sorted with a BD INFLUX and replicates 3 and 4 were sorted with a BD FACSARIA II. In both cases, events were tightly gated around mean RFP fluorescence to control for some extrinsic noise in protein production (the *mRFP* gene is located on the same plasmid as the reporter gene and constitutively expressed). The fluorescence range of the library in the green channel was divided into 16 equally sized contiguous bins in $\log_{10}$ space. The population was sorted through each of the 16 resulting gates in different tubes (4 sorting rounds of 4 tubes each). To ensure that the number of sorted cell in each bin was proportionate to their phenotypic density in the initial population, the sort rate was manually maintained constant throughout the procedure and each of the bins was sorted for an equal amount of time. Given the size of the library, the actual sort rates were a limiting factor. Collection times varied between 12 (replicates 1 and 2) and 24 h (replicates 3 and 4) to permit collection of enough cells (at least 100-fold the library size). To avoid phenotypic evolution of the sample between collection times, several cultures of the same replicate were seeded with delays matching their collection times. The order in which the 16 bins were collected was randomly selected for each replicate.

To amplify the populations collected after sorting, we added an equal volume of LB medium to the sorted cells in PBS and supplemented it with appropriate concentrations of kanamycin and chloramphenicol for plasmid maintenance. After growth saturation was reached, we miniprepped (Qiagen) the plasmid DNA from 2-mL aliquots for each subpopulation and quantified concentrations using a Nanodrop (Thermo Scientific).

Preparation of sequencing libraries. For each bin, we amplified 5 ng of extracted plasmids by PCR for 15 cycles. We used long oligonucleotides wherein the priming region is preceded by Illumina sequencing adapters followed by defined 8-nt-long barcodes. Spacers of varying sizes were introduced between barcodes and the standard priming site of the adaptor to introduce complexity at the beginning of the sequence, as required by the clustering step of the Illumina sequencing procedure. Primer sequences are detailed in **Supplementary Table 2**.

Combinations of two barcodes introduced by the forward and reverse primers were used to uniquely identify sample origin upon multiplexing. Amplicons were cleaned and size-selected using magnetic SPRI beads (Agencourt). The quantity of purified DNA was measured using Qbit (Life Technologies). We mixed samples originating from each of the binned population in amounts proportionate to their expected diversities (i.e., the frequencies of the bins in the whole population, as estimated from the sorted cell numbers).

Compatible replicates (i.e., those tagged with different barcode combinations) were pooled in pairs in equal proportions (replicate 1 with 3 and replicate 2 with 4). Each pool was sequenced using the rapid mode of a HiSeq 2500 lane (Illumina, 150PE).

Processing of FACS-seq data. Sequencing reads were de-multiplexed using custom Python scripts (**Supplementary Code 2**). Sequences located in the design region were then mapped to the designed sequences using BWA[70] (**Supplementary Code 3**). Data were compiled to produce a summary dataset that counts the number of perfect read counts observed in each bin for each of the design sequences (**Supplementary Code 4** and **5** and **Supplementary Data 9**). Reads with mutations were saved in a separate file and excluded from further analyses. For each replicate experiment, the read frequencies observed in each bin were adjusted to match those expected on the basis of the cell densities observed during sorting (**Supplementary Data 14**), as follows:

$$f_{s,i,r}^{\text{corrected}} = f_{s,i,r} \times \frac{R_{i,r}}{\sum_{i=1}^{16} R_{i,r}} \times S_{i,r} \qquad (1)$$

where, for replicate r, $f_{s,i,r}^{\text{corrected}}$ is the corrected frequency for sequence s in bin i, $f_{s,i,r}$ is the observed frequency, $R_{i,r}$ is the total read count in bin i across all sequences, and $S_{i,r}$ is the observed frequency of cell sorted in bin i during the FACS procedure for that replicate. This correction is necessary to correct for unavoidable DNA quantification and loading error when multiplexed samples are pooled at specific ratios for sequencing.

From these reconstructed fluorescence profiles, we calculated a mean fluorescence as the weighted mean of the four adjacent bins summing maximal frequencies:

$$\mu_{s,r_{\log}} = \sum_{i=j_{s,r}}^{j_{s,r}+4} (f_{s,i,r} \times i) \Bigg/ \sum_{i=j_{s,r}}^{j_{s,r}+4} f_{s,i,r} \qquad (2)$$

where $j_{s,r} \in [1,12]$ is the index of the first bin in the four adjacent bins selected for each sequence in each replicate. This simple filtering procedure permitted us to partly resolve bimodal profiles, wherein highly fluorescent strains often show weak signal at low fluorescence. This pattern probably results from occasional mutations outside the sequenced region, which tend to be selected among high-reporter producers.

To control for systematic differences between replicates, we rescaled each replicate linearly to minimize the between-replicate error. Again, we observed that high-reporter-producing strains occasionally show much reduced performances in one replicate. We excluded such low values on the basis that they represent a mutant that took over the original strain in a given replicate. We then quantile-normalized the data[71] to obtain corrected values of the mean in log space ($\mu^{*}_{s,r_{\log}}$).

We applied an exponential transformation to map the cleaned data on a linear space:

$$\mu_{s,r\,\text{lin}} = e^{\mu^{*}_{s,r\log/C}} \qquad (3)$$

where C is a constant chosen to maximize the kurtosis of the observed distribution using the nlm package in R. Final data were rescaled between 1 and 100 to obtain the protein production metric used in all analyses:

$$P_{s,r} = \frac{\mu_{s,r_{\text{lin}}} - \min\left(\mu_{s,r_{\text{lin}}}\right)}{\max\left(\mu_{s,r_{\text{lin}}}\right) - \min\left(\mu_{s,r_{\text{lin}}}\right)} \times 99 + 1 \qquad (4)$$

In all graphs, data points show the mean protein production across the four replicates (P_{NI}). Pairwise comparisons of the replicates after processing are shown in **Supplementary Figures 5b** and **6b**.

The whole processing procedure was implemented in R (**Supplementary Code 6**).

Measurement of growth. *Growth conditions.* A library aliquot was initially grown overnight at 37 °C in 5 mL Rich MOPS medium (Teknova) supplemented with kanamycin and chloramphenicol for plasmid maintenance and aTc inducing *tev* transcription. Plasmid DNA was extracted from a 2-mL aliquot of that culture (ZymoPure Plasmid Miniprep Kit, Zymo Research). We then serially propagated three replicate lines of this initial culture concurrently in each of four different growth conditions:

1. *Transcriptional repression.* Rich MOPS medium (Teknova) supplemented with kanamycin and chloramphenicol (plasmid maintenance) and aTc (induction of *tev* transcription). The reporter plasmid contains a strongly expressed *lacI* gene. In the absence of IPTG or lactose, transcription of the reporter is strongly repressed by LacI. Under these conditions, no fluorescence above background could be detected in exponentially growing cells. Three days after plating on LB agar supplemented with kanamycin, chloramphenicol and aTc, we noticed few (~0.1%) faintly bright overgrown colonies presumably corresponding to the highest expressers, which confirms very low amount of transcriptional leakage. After data measurement and processing, this condition yielded measures of W_{UTX}.

2. *Regular initiation.* Same as in 1 with the addition of IPTG to induce transcription of the reporter system. After data measurement and processing, this condition yielded measures of W_{NI}.

3. *Facilitated initiation.* Same as in 2 with the addition of arabinose to induce the expression of the unnatural aminoacyl-tRNA synthetase and the unnatural

doi:10.1038/nbt.4238

amino acid pAcF (2.5 nM). Since the addition of pAcF acidifies the medium, the pH was adjusted to its original value of 7.2 by addition of NaOH. After data measurement and processing, this condition yielded measures of W_{FI}.

4. *Minimal medium*. Same as in 2 but using minimal MOPS medium (Teknova) supplemented by glucose (0.2%). After data measurement and processing, this condition yielded measures of W_M.

The culture propagation protocol was as follows: 100 µL of cells from a saturated culture were diluted in 10 mL fresh medium (1:100 dilution). Inoculated cultures were grown to saturation for 12 h at 37 °C in a thermostatic shaker (250 r.p.m.). We repeated this operation nine times for conditions 1 and 2 and five times for conditions 3 and 4. We then extracted plasmid DNA from 4-mL aliquots of the final saturated culture for each replicate in all conditions (ZymoPure Plasmid Miniprep Kit, Zymo Research).

After each growth cycle, we monitored the fluorescence profile of the population by flow cytometry. We observed that a population of nonfluorescent cells, in both the red and green channels, steadily increased over time in each culture. We isolated this cell fraction by FACS and plated it on LB agar supplemented with kanamycin and chloramphenicol. Plasmid DNA was extracted from a 48 randomly picked colonies and retransformed into naive cells. This confirmed that the nonfluorescent phenotype was specified by the plasmids. We could not obtain PCR products when trying to amplify the design region from these plasmids using the standard primers used for sequencing. These observations suggest that the double-negative fluorescent phenotype results from major genetic rearrangement(s). Since these mutants cannot be amplified during the preparation of the sequencing library, it does not interfere with the deep sequencing assay described below. Nonetheless, these nonfunctional mutants appeared to take over the population over time, probably as a result of decreased cost of protein production. This process was faster in conditions 3 and 4 (facilitated initiation and minimal medium), wherein expression of the reporter is expected to be particularly costly. We set the length of the competition experiments so that the fraction of double negative mutants remained <10% in all populations, as estimated by flow cytometry—hence the difference of competition time between conditions 1 and 2 and conditions 3 and 4. The number of generations required for reaching saturation upon 100-fold dilution is $\log_2(100) \approx 6.64$. Populations were thus propagated for ~60 and ~33 generations for conditions 1 and 2 and conditions 3 and 4, respectively.

In follow up experiments, we sequenced populations evolved for ~13 and ~28 generations under normal initiation conditions (condition 2) and ~26 generations in facilitated initiation conditions (condition 3).

Preparation of sequencing libraries. Sequencing libraries were prepared from plasmid DNA using targeted PCR and sequenced as presented for the measurement of protein production. To ensure constant concentration of template DNA, the quantity of input plasmid was adjusted in each reaction to account for the fraction of double negative mutants measured by flow cytometry. Libraries for the initial population and the 12 evolved samples were mixed in equal quantities and sequenced in multiplex (HiSeq2500 Rapid mode, 150PE, 1.5 lanes). Barcodes were used to keep track of replicate identity and growth conditions upon multiplexing (**Supplementary Table 2**).

Processing of growth data. We mapped sequencing reads to design sequences using the same procedure as described above for protein production (**Supplementary Code 1–5** and **Supplementary Data 9**). Strains with less than 5 reads in any final sample were discarded from the data. Strains with less than 15 reads in the initial conditions were also discarded.

We quantified relative growth rates using the following equation[42,72]:

$$W_{s,r} = \frac{\log\left(2^{g_r} \times \frac{f_{s,r}}{f_{s,i}}\right)}{\log\left(2^{g_r} \times \frac{1-f_{s,r}}{1-f_{s,i}}\right)} \quad (5)$$

where $f_{s,r}$ is the observed frequency of sequence s in replicate r, $f_{s,i}$ is the observed frequency of the same sequence in the initial population, and g_r is the number of generations undergone by replicate r (2^{g_r} gives the fold expansion of the total population during the course of the experiment). We then used the arithmetic mean over replicate as a measure of growth rates.

This procedure straightforwardly yielded W_{UTX} and W_M for conditions 1 and 4, respectively. In conditions 2 and 3, however, many strains went practically extinct by the end of the initial competition experiment. For example, 85,094 strains yielded ≤5 reads in each of the three replicates grown for 60 generations in condition 2. As a result, we could only derive average growth data for 150,172 strains in this condition. Although these data indicated low experimental error for this type of batch competition assay (5% of the total variance), the differences between replicates were larger for lower growth rates (**Supplementary Fig. 10a**). This is because the most severely outcompeted strains produced very discrete read numbers, resulting in misleading fluctuations after data processing. Similarly, we could initially only derive average growth data for 180,855 strains in condition 3. Altogether, these experiments were lacking important information regarding the most physiologically affected strains.

We reasoned that shorter competition times should improve the detection of slowest growing strains. We therefore repeated the assay with a single independent population sampled after ~13 and ~28 generations in condition 2 and ~26 generations in condition 3. To correct for small systematic biases between time points in a given condition, we rescaled the data to minimize the fraction of total variance accounted for by variations between time points within strains. Earlier time points were highly correlated with later ones, while showing larger library coverage and wider dynamic range (**Supplementary Fig. 10b**,**c**). Yet later time points afford increased sensitivity to minute growth differences among strains in the bulk of the library (**Supplementary Fig. 10d**). To combine the benefits of different samples and obtain a broader measure of growth rate, we considered the geometric mean of rates estimated across time points. We thus derived a consolidated growth measurement for 233,846 strains in condition 2 (W_{NI}) and 229,224 strains in condition 3 (W_{FI}).

The entire processing procedure was implemented in R (**Supplementary Code 18**).

Measurement of RNA decay. *Growth conditions.* A frozen aliquot of the library was used to inoculate two replicate cultures in 5 mL Rich MOPS medium (Teknova) supplemented with kanamycin and chloramphenicol for plasmid maintenance. Cultures were grown overnight in a thermostatic shaker at 37 °C (250 r.p.m.). The two cultures were then diluted 1:100 in 50 mL Rich MOPS medium (Teknova) supplemented with kanamycin, chloramphenicol, IPTG (induction of reporter transcription) and aTc (induction of *tev* transcription). Replicate cultures were grown to steady state (OD_{600} = 0.5), at which point two 4-mL samples were taken and the culture was put back to grow after the antibiotic rifampicin was added to a final concentration of 50 µg/mL to stop transcription. Aliquots of 4 mL were then taken after 2.5, 5, 10, 20 and 40 min of growth in these conditions (**Supplementary Fig. 11a**).

Upon sampling, each sample was immediately mixed with 8 mL of RNAprotect Bacteria reagent (Qiagen), incubated for 5 min at room temperature, and centrifuged at 5,000*g* and 4 °C for 5 min. The supernatant was discarded and the pellets were flash frozen in liquid nitrogen and stored at −80 °C.

We grew eight standard strains to OD_{600} = 0.5 following the same procedure. We then mixed these cultures in 1:5:50:500 ratios for standards 3 and 7; 1 and 2; 4 and 5; and 6 and 8, respectively. We processed a 4-mL sample of the resulting population as described above.

Preparation of sequencing libraries. Frozen pellets resulting from the above procedure were thawed on ice and treated with lysozyme for 10 min at room temperature. At this point the standard sample was further diluted 1:100 and 17 µL of the resulting solution was spiked into all library samples. Total RNAs were then extracted from these supplemented samples using the RNeasy kit with DNase treatment following the manufacturer's instruction (Qiagen).

We quantified and checked the integrity of the RNA preparation using the Bioanalyzer RNA 6000 Pico kit (Agilent). From each sample, 5 µg was then treated with RiboZero for Gram-Negative Bacteria (Epibio) to remove ribosomal RNAs. We quantified and checked the integrity of extracted mRNAs using the Bioanalyzer RNA 6000 Nano kit (Agilent). We then used 500 ng mRNA for first-strand cDNA synthesis using SuperScript III Reverse Transcriptase (Invitrogen) with appropriate standard primers used for library sequencing (20-µL reactions, 55 °C for 60 min). Two microliters of the resulting reactions were used as templates for further PCR amplification, as described earlier. One of the two culture samples collected at time 0 was used to extract plasmid DNA and prepare a sequencing library by PCR, as described above.

Samples from the same replicates were pooled in equal quantities, and each replicate was sequenced on a lane of an Illumina HiSeq 2500 (rapid mode, 150PE). Barcodes in the primers were used to keep track of sample identity upon multiplexing, as described above (**Supplementary Table 2**).

Processing of mRNA data. For each replicate, the sequencing reads were de-multiplexed and mapped to the designed and standard sequences using the pipeline described above. Barcodes were used to assign reads to the different samples. Measurements with less than ten reads at the initial time point ($t = 0$) in either the DNA or RNA sample were excluded from the dataset (n = 20,028 and 9,333 in replicates 1 and 2, respectively). We also excluded n = 128 strains with extreme outlying measurements between the two replicates (residuals >10 times the s.d. among residuals after linear regression between replicates at any time points).

Assuming unbiased sampling, we can write the following equality for the read frequencies observed at each sample time t within each replicate r:

$$f_{r,t}^{s} = \frac{n_{r,t}^{s}}{n_{r,t}^{\text{std}} + n_{r,t}^{\text{lib}}} = \frac{N_{r,t}^{s}}{N^{\text{std}} + N_{r,t}^{\text{lib}}} \tag{6}$$

where $n_{r,t}^{s}$ and $N_{r,t}^{s}$ are the read count and actual number of transcript molecules in the population for sequence s, $n_{r,t}^{\text{lib}}$ and $N_{r,t}^{\text{lib}}$ are the total read count and actual number of molecules for library sequences, and $n_{r,t}^{\text{std}}$ and N^{std} are the total read count and actual number of molecules for all standard sequences. Importantly, the number of spiked-in standard molecules N^{std} is assumed to be constant across sample and replicates (as expected for a constant volume of cell lysate spiked before RNA extraction; see above). Likewise, we can write the following equality for the pooled standard frequency:

$$f_{r,t}^{\text{std}} = \frac{n_{r,t}^{\text{std}}}{n_{r,t}^{\text{std}} + n_{r,t}^{\text{lib}}} = \frac{N^{\text{std}}}{N^{\text{std}} + N_{r,t}^{\text{lib}}} \tag{7}$$

As expected, the frequencies of pooled and individual standards increase over time (**Supplementary Fig. 11a**, middle). Rearranging this equation exposes the global decay of library mRNAs over time:

$$N_{r,t}^{\text{lib}} = N^{\text{std}} \times \left(\frac{1}{f_{r,t}^{\text{std}}} - 1 \right) \tag{8}$$

Plotting these quantities, we find a better agreement between the decay profiles of the two replicates if an extra 1 min is systematically added to the sampling times of the second replicate. This slight increase in decay probably reflects the time delay with which the second replicate was handled during the experiment. We thus corrected sampling times for replicate 2 by 1 min.

Rearranging equation (5) and inserting equation (7) allows us to express the actual number of transcript molecules in each sample as a function of the observed frequencies:

$$N_{r,t}^{s} = N^{\text{std}} \times \frac{f_{r,t}^{s}}{f_{r,t}^{\text{std}}} \tag{9}$$

This quantity does not directly correspond to the amount of RNA per cell because strains are present at different frequencies in the library. To account for this, we assumed constant plasmid copy number between cells and, for each sequence, normalized the frequency of RNA by the corresponding DNA frequency measured at $t = 0$:

$$R_{r,t}^{s} = C \times \frac{f_{r,t}^{s}}{f_{r,t}^{\text{std}} \times f_{\text{DNA}}^{s}} \tag{10}$$

For this calculation, we used the mean DNA frequencies over the two replicates (f_{DNA}^{s}) instead of the cognate replicate frequencies. We found poorer replicability for the DNA measurements (correlation between replicates r = 0.82) as compared to the RNA measurements ($r = 0.88$ on average across time samples), and we thus avoided propagation of this error. C is an arbitrary constant equal to the unknown N^{std} divided by the average RNA/DNA ratio over the two replicates at $t = 0$ (i.e., such that $\sum_{r,s} R_{r,0}^{s} / N = 1$, with N being the total number of sequences).

For each strain, we then fitted the time series using the following nonlinear model of exponential decay:

$$\text{RNA}_{r}^{s}(t) = \frac{a}{b + e^{-c/t}} \tag{11}$$

where $t \in \{10^{-300}, 2.5, 5, 10, 20, 40\}$ for the first replicate, $t \in \{10^{-300}, 3.5, 6, 11, 21, 41\}$ for the second replicate and $c > 0$. We rejected fits in which $R^2 < 0.9$ and could derive estimates of a, b and c for 233,487 strains in the library. We used these estimates to derive steady-state mRNA abundance RNA_{SS}; transcript half-life RNA_{HL}; and the transcript protection index RNA_{PTX}, representing the fraction of steady-state mRNA asymptotically resistant to degradation:

$$\text{RNA}_{\text{SS}} = \frac{a}{b} \tag{12}$$

$$\text{RNA}_{\text{HL}} = -\frac{c}{\log(b)} \tag{13}$$

$$\text{RNA}_{\text{PTX}} = \frac{b}{b+1} \tag{14}$$

The processing of all RNA measurements was implemented in R (**Supplementary Code 21**).

Polysome profiling. *Polysome extraction.* A library aliquot was used to start an overnight culture in 5 mL Rich MOPS medium (Teknova) supplemented with kanamycin and chloramphenicol (plasmid maintenance), IPTG (induction of reporter transcription) and aTc (induction of *tev* transcription). The next day, the culture was diluted 1:100 in 200 mL of the same medium and grown to steady state ($\text{OD}_{600} = 0.5$) in a thermostatic shaker (37 °C, 250 r.p.m.). Cells were rapidly collected on a filter using a disposable vacuum trap (Corning). The filter and the cells were immediately detached from the trap, flash frozen in liquid nitrogen and ground using a mortar filled with liquid nitrogen after adding 800 µL of cell lysis solution (as described by Oh *et al.*[73]). The crushed material was collected and stored at −80 °C. Half of the material was further treated in a frozen state with a refrigerated bead beater (two pulses of 45 s at 4,800 r.p.m.). The resulting material was melted, incubated on ice for 20 min and then centrifuged at 4 °C and 17,000*g*. The supernatant was loaded onto a 10–50% sucrose gradient and ultracentrifuged at 35,000 r.p.m. (SW41 rotor) and 4 °C for 3 h to separate polysome fractions from the cell extract[74]. The first five fractions (monosome to pentasome) were clearly separated and could be collected separately (**Supplementary Fig. 12a**). Collected samples were immediately flash frozen in liquid nitrogen.

Preparation of sequencing libraries. RNA was extracted from each collected fraction using TRIzol (Life Technologies). The integrity and quantity of the extract were measured using a Bioanalyzer RNA 6000 Nano kit (Agilent). As expected, these profiles showed little contamination by ribosomal RNAs. We directly used 500 ng of that RNA extract as template for first-strand cDNA synthesis using 1.5 µL (300 U) of SuperScript III Reverse Transcriptase (Invitrogen) with appropriate standard sequencing primers (20 µL reaction, 55 °C for 60 min). Two microliters of the resulting reactions were used as template for further PCR amplification, as described above. Barcodes were used to keep track of fraction identity. The library was sequenced on half a lane of a HiSeq 2500 (rapid mode, 150PE).

Processing of polysome profile data. We calculated read frequencies of each strain within each ribosomal fraction. We quantile-normalized these frequencies across fractions to remove systematic bias between samples. For strain s, this provided a set of $f_{s,i}$ values forming a polysome profile over fractions i. To correct for differences in strain abundance and steady-state RNA abundance between strains, we further mean-normalized the observed frequencies across ribosomal fraction within each strain:

$$f_{s,i}^{\text{norm}} = \frac{5 f_{s,i}}{\sum_{i=1}^{5} f_{s,i}} \tag{15}$$

This measure provides relative polysome loading profiles that are comparable across strains. For each transcript, we then calculated the mean polysome density MRD_{s} as follows:

doi:10.1038/nbt.4238

$$\mathrm{MRD}_s = \frac{\sum_{i=1}^{5}(i \times f_{s,i})}{\sum_{i=1}^{5} f_{s,i}} \quad (16)$$

The processing of polysome profiling data was implemented in R (**Supplementary Code 23**).

Statistics. We used R for all statistical analyses[75]. Given our large sample sizes, we have refrained from using *P*-values as a measure of statistical significance[76]. In ANOVAs, for example, *P*-values diminish greatly with sample size, which improves our ability to detect the existence of effects regardless of their magnitude. Many highly significant effects are in fact so small that they might not be relevant (for example, $\eta^2 < 0.01$; **Supplementary Table 3**) Thus, we have generally based our analyses and interpretations of the data on effect sizes (η^2 for ANOVAs and R^2 for linear regressions).

Data processing, management and analysis. All data were consolidated into a single dataset and analyzed using R[75] (**Supplementary Data 15**). A dataset description is provided as **Supplementary Note 1**. Along with **Supplementary Code 1–28** and **Supplementary Data 1–36**, scripts to generate all figures in this manuscript are available from the Open Science Framework website (https://osf.io/a56vu/?view_only=0d5b05fb08d84b76b21f399e832808b6; https://doi.org/10.17605/OSF.IO/A56VU).

Material availability. Reporter (Addgene 113408 and 113409) and accessory (Addgene 110206) plasmids are deposited at Addgene (https://www.addgene.org/). The entire library is available upon request to G.C. or A.P.A.

Code availability. Code is available in **Supplementary Code 1–28**. Modules used to generate the synthetic sequences have been added to the D-Tailor project (https://github.com/jcg/d-tailor, v1.0). All scripts used to process raw sequencing data, resulting measurement data and additional pieces of data, as well as all R scripts used to generate main and supplementary figures, are available for unrestricted download from the Open Science Framework (https://osf.io/a56vu/?view_only=0d5b05fb08d84b76b21f399e832808b6; https://doi.org/10.17605/OSF.IO/A56VU).

Data availability. Data are available in **Supplementary Data 1–36.** De-multiplexed sequencing reads are deposited in the Sequence Read Archive under project accession code SRP086076.

59. Kosuri, S. & Church, G.M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11**, 499–507 (2014).
60. Miller, W.G., Leveau, J.H.J. & Lindow, S.E. Improved *gfp* and *inaZ* broad-host-range promoter-probe vectors. *Mol. Plant Microbe Interact.* **13**, 1243–1250 (2000).
61. Lee, T.S. *et al.* BglBrick vectors and datasheets: a synthetic biology platform for gene expression. *J. Biol. Eng.* **5**, 12 (2011).
62. Kapust, R.B. & Waugh, D.S. Controlled intracellular processing of fusion proteins by TEV protease. *Protein Expr. Purif.* **19**, 312–318 (2000).
63. Kapust, R.B., Tözsér, J., Copeland, T.D. & Waugh, D.S. The P1′ specificity of tobacco etch virus protease. *Biochem. Biophys. Res. Commun.* **294**, 949–955 (2002).
64. Cambray, G. *et al.* Measurement and modeling of intrinsic transcription terminators. *Nucleic Acids Res.* **41**, 5139–5148 (2013).
65. Glascock, C.B. & Weickert, M.J. Using chromosomal lacIQ1 to control expression of genes on high-copy-number plasmids in *Escherichia coli*. *Gene* **223**, 221–231 (1998).
66. Elowitz, M.B., Levine, A.J., Siggia, E.D. & Swain, P.S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
67. Liang, J.C., Chang, A.L., Kennedy, A.B. & Smolke, C.D. A high-throughput, quantitative cell-based screen for efficient tailoring of RNA device activity. *Nucleic Acids Res.* **40**, e154 (2012).
68. Pósfai, G. *et al.* Emergent properties of reduced-genome *Escherichia coli*. *Science* **312**, 1044–1046 (2006).
69. Csörgo, B., Fehér, T., Tímár, E., Blattner, F.R. & Pósfai, G. Low-mutation-rate, reduced-genome *Escherichia coli*: an improved host for faithful maintenance of engineered genetic constructs. *Microb. Cell Fact.* **11**, 11 (2012).
70. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
71. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
72. van Opijnen, T., Bodi, K.L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods* **6**, 767–772 (2009).
73. Oh, E. *et al.* Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell* **147**, 1295–1308 (2011).
74. Qin, D. & Fredrick, K. Analysis of polysomes from bacteria. *Methods Enzymol.* **530**, 159–172 (2013).
75. R Core Team. R: a language and environment for statistical computing https://www.R-project.org/ (2017).
76. Sullivan, G.M. & Feinn, R. Using effect size—or why the *P* value is not enough. *J. Grad. Med. Educ.* **4**, 279–282 (2012).

nature research

Corresponding author(s): Guillaume Cambray and Adam Arkin

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.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

Describe how sample size was determined.

We used a full factorial Design of Experiment comprising 1,458 combinations of sequence properties. We chose to replicate that design in 56 distinct regions of sequence space, so that we could further design 2 additional local replicates for each synthetic sequences. These numbers were dictated by the maximal number of sequences that could be synthesized on a single OLS oligoship (244,000). In all, we thus designed, synthesized and cloned sequences to obtain a theoretical library of 244,000 strains.
Data were derived from high-throughput sequencing read counts. Although we always sequenced at a depth enabling sufficient coverage of the library (at least 100x), some strains occasionally fell below detection level. According to the specificity of each measurements, we determined empirically the minimal number of reads required to give consistent results. As detailed below, we excluded data that did not meet these thresholds.

2. Data exclusions

Describe any data exclusions.

In any measurements, sequencing reads that did not provided an exact match to a design sequence were not considered. Sequencing read showing mixed index were also discarded.

For the measurement of protein production, some strains showed bimodal fluorescence distribution, generally with very high and very low signals. We considered that the very low signal originated from mutant breaking high reporter expression. As detailed in online methods, we used a simple filtering procedure to calculate protein production from the main mode of the distribution. We only considered strains for which one bin in the mode comprised at least 5 sequencing reads. For the calculation of average protein production from the four biological replicates, we excluded replicate values that were more 5 units of fluorescence (on a scale from 1 to 16) away from the mode of the replicates.

For mRNA measurements, we discarded strains with less than 10 sequencing reads (either DNA or RNA-Seq) at the initial time point. Data showing extreme variations between two biological replicates were discarded, as we had no way to determine which was the most sensible. To determine extreme variations for each time points, we regressed reads counts for the two replicates against one another, calculated the standard deviation of the residuals and excluded strains for which the absolute value of residuals were superior to 10 times that standard deviation. For each replicate, we estimated the quality of the fit to an exponential decay model using R-squared. We excluded replicate with R-squared below 0.9 and those with R-squared more than 0.05 below the other replicate.

For growth measurements, we excluded strains with less than 15 read counts in the initial population and less than 5 counts in the evolved populations.

For polysome profiling, we attempted to isolate and sequence polysome fractions 6 and 7, even though they were barely visible on the gradient profile. Upon sequencing, the data were not consistent with those obtained from lower fractions and were thus not considered in the analysis.

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

For protein production, we performed 4 measurements on independent biological replicates and two FACS-Machine. The repetability was excellent in regular condition of initiation (r~0.99), a bit less so under facilitated initiation (r~0.90, except for one replicate which had an obvious problem but was kept for analysis). We also compared the high-throughput measurement with lower throughput ones established on isolated strains (r~0.95 and 0.90 for regular and facilitated initiation).

Nature Biotechnology: doi:10.1038/nbt.4238

For growth measurements, we performed parallel triplicate experiments from the same initial culture for the main measurement (regular and facilitated initiation). The repoducibility was good for this type of experiment (r=0.89-0.92).

For mRNA abundance and stability measurements, we could only afford 2 replicates (each involving 7 sequencing libraries). We observed good replicability for RNA measurements (r=0.90, 0.89, 0.92, 0.88, 0.87, 0.82 at t=0, 2.5, 5, 10, 20 and 40 min, respectively), but lower replicability for DNA measurements at t=0 (0.82). Because we suspected a problem with one of the DNA sample, we averaged DNA measurements before calculating RNA/DNA ratios and thus partly avoid error propagation. Replicability between ratioed replicate was nonetheless affected (r=0.82, 0.84, 0.89, 0.83, 0.76, 0.66).

For polysome profiling we performed only one replicate. As mentioned in the text, we used correlation with protein production (r=0.73) to ensure the consistency of these measurements.

In general, the smooth relationships between measurements is also proof of the overall consistency of the results.

We constructed and analyzed multiple level of replication of our molecular Design of Experiment, which allowed us to estimate various sources of error and avoid over-generalization of idiosyncratic findings.

4. Randomization

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

The concept of randomization does not fully apply to our study.

We used a design software based on a monte-carlo process to generate our sequence subjects. Seed sequences were chosen randomly and evolved to satisfy particular sequence properties to satisfy a Full-Factorial design of experiments. his constrained random sequence generation certainly provided the ability to apply our factorial analysis and we validated that most of the data met the criteria for this generalized linear model. However, the physical constraints and random seeding certainly led to some biasing in the sampling though we do not believe this affects the conclusions of our analysis.

5. Blinding

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

Data were acquired through diverse high-throughput screening and subsequent sequencing performed on a single library containing all subjects. The identity of the individuals became apparent only after sequencing. Sequencing data were processed in bulk without paying attention to individuals identity. Likewise, data analysis were performed in bulk.

Note: all in vivo studies must report how sample size was determined and 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 *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes 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 in all relevant figure captions (with explicit mention of central tendency and variation)

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

We used D-Tailor for sequence analysis and design. We used python to process sequencing

Nature Biotechnology: doi:10.1038/nbt.4238

Describe the software used to analyze the data in this study.	data. We used BWA for read mapping. We used R for all analysis and figure plotting. All scripts have been made publicly available on the Open Science Framework (or on Github for D-Tailor).

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 third party.	Both reporter and accessory vectors have been deposited at Addgene. All the material will be made available upon request. Only 300+ library clone have been individualized. All others are mixed in a complex library.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	Not applicable

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	Not applicable
b. Describe the method of cell line authentication used.	Not applicable
c. Report whether the cell lines were tested for mycoplasma contamination.	Not applicable
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.	Not applicable

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 all relevant details on animals and/or animal-derived materials used in the study.	Not applicable

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.	Not applicable

Nature Biotechnology: doi:10.1038/nbt.4238