Absolute quantification of translational regulation and burden using combined sequencing approaches

Thomas E. Gorochowski[1,2,*], Irina Chelysheva[3], Mette Eriksen[4], Priyanka Nair[3], Steen Pedersen[4], Zoya Ignatova[3,*]

[1] BrisSynBio, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol, UK

[2] School of Biological Sciences, University of Bristol, Tyndall Avenue, Bristol, UK

[3] Biochemistry and Molecular Biology, Department of Chemistry, University of Hamburg, Hamburg, Germany

[4] Biomolecular Sciences, Department of Biology, University of Copenhagen, Copenhagen, Denmark

* Correspondence should be addressed to T.E.G (thomas.gorochowski@bristol.ac.uk) and Z.I. (zoya.ignatova@uni-hamburg.de)

Keywords: translation; transcription; Ribo-seq; RNA-seq; genetic circuits; biometrology; systems biology; synthetic biology

Abstract

Translation of mRNAs into proteins is a key cellular process. Ribosome binding sites and stop codons provide signals to initiate and terminate translation, while stable secondary mRNA structures can induce translational recoding events. Fluorescent proteins are commonly used to characterize such elements but require the modification of a part's natural context and allow only a few parameters to be monitored concurrently. Here, we develop an approach that combines ribosome profiling (Ribo-seq) with quantitative RNA sequencing (RNA-seq) to enable the high-throughput characterization of genetic parts controlling translation in absolute units. We simultaneously measure 743 translation initiation rates and 746 termination efficiencies across the *Escherichia coli* transcriptome, in addition to translational frameshifting induced at a stable RNA pseudoknot structure. By analyzing the transcriptional and translational response, we discover that sequestered ribosomes at the pseudoknot contribute to a σ^{32}-mediated stress response, codon-specific pausing, and a drop in translation initiation rates across the cell. Our work demonstrates the power of integrating global approaches towards a comprehensive and quantitative understanding of gene regulation and burden in living cells.

Introduction

Gene expression is a multi-step process involving the transcription of DNA into messenger RNA (mRNA) and the translation of mRNAs into proteins. To fully understand how a cell functions and adapts to changing environments and adverse conditions (e.g., disease or chronic stress), quantitative methods to precisely observe these processes are required (Belliveau *et al*, 2018). Gene regulatory networks (also known as genetic circuits) control where and when these processes take place and underpin many important cellular phenotypes. Recently, there has been growing interest in building synthetic genetic circuits to understand the function of natural gene regulatory networks through precise perturbations and/or creating systems *de novo* (Wang *et al*, 2016; Smanski *et al*, 2016).

In synthetic biology, genetic circuits are designed to control gene expression in a desired way (Brophy & Voigt, 2014). Circuits have been built to implement a range of digital (Fernandez-Rodriguez *et al*, 2015; Moon *et al*, 2012) and analog functions (Daniel *et al*, 2013), and have been integrated with endogenous pathways to control cellular behaviors (Tan *et al*, 2016; Nielsen & Voigt, 2014). The construction of a genetic circuit requires the assembly of many DNA-encoded parts that control the initiation and termination of transcription and translation. A major challenge is predicting how a part will behave when assembled with many others (Cardinale *et al*, 2013). The sequences of surrounding parts (Poole *et al*, 2000), interactions with other circuit components or the host cell (Ceroni *et al*, 2015; Gyorgy *et al*, 2015; Cardinale *et al*, 2013; Gorochowski *et al*, 2016), and the general physiological state of the cell (Gorochowski *et al*, 2014; Wohlgemuth *et al*, 2013) can all alter a part's behavior. Although biophysical models have been refined to capture some contextual effects (Salis *et al*, 2009; Espah Borujeni *et al*, 2013; Seo *et al*, 2013), and new types of part created to insulate against these factors (Shendure *et al*, 2017; Yang *et al*, 2014; Daniel *et al*, 2013; Moon *et al*, 2012; Siuti *et al*, 2013; Mutalik *et al*, 2013), we have yet to reach a point where large and robust genetic circuits can be reliably built on our first attempt.

Fluorescent proteins and probes are commonly used to characterize the function of genetic parts (Hecht *et al*, 2017; Jones *et al*, 2014) and debug the failure of genetic circuits (Nielsen *et al*, 2016). When used for characterization, the part of interest is usually placed into a new genetic backbone (often a plasmid) and its behavior is directly linked to the expression of one or more fluorescent proteins (Cambray *et al*, 2013). When debugging a circuit failure, it is not possible to extract the part of interest as the context of the circuit is important. For circuits that use transcription rate (i.e. RNAP flux) as a common signal between components (Canton *et al*, 2008), debugging plasmids containing a promoter responsive to the signal of interest have been used to track the propagation of signals and reveal the root cause of failures (Nielsen *et al*, 2016). Alternatively, any

genes whose expression is controlled by the part of interest can be tagged by a fluorescent protein (Snapp, 2005). Such modifications allow for a readout of protein level but come at the cost of alterations to the circuit. This is problematic as there is no guarantee the fluorescent tag itself will not affect a part's function (Baens *et al*, 2006; Margolin, 2012).

The past decade has seen tremendous advances in sequencing technologies. This has resulted in continuously falling costs and a growing range of information that can be captured (Goodwin *et al*, 2016). Sequencing methods exist to measure chromosomal architecture (Lieberman-aiden *et al*, 2009), RNA secondary structure (Lucks *et al*, 2011), DNA and RNA abundance (Conesa *et al*, 2016), and translation efficiency (Ingolia, 2014). New developments have expanded the capabilities even further towards more quantitve measurements of transcription and protein synthesis rates with native elongating transcript sequencing (NET-seq) (Mayer *et al*, 2015) and ribosome profiling (Ribo-seq) (Li *et al*, 2014; Ingolia *et al*, 2009). Ribo-seq provides position-specific information on the translating ribosomes through sequencing of ribosome-protected fragments (RPFs; approximately 25–28 nt) which allows genome-wide protein synthesis rates to be inferred with accuracy similar to quantitative proteomics (Li *et al*, 2014).

Sequencing technologies offer several advantages over fluorescent probes for characterization and debugging genetic parts and circuits. First, they do not require any modification of the circuit DNA. Second, they provide a more direct measurement of the processes being controlled (e.g. monitoring transcription of specific RNAs), and third, they capture information regarding the host response and consequently their indirect effects on a part's function. Furthermore, for large multi-component circuits or synthetic genomes, sequencing is the only way of gaining a comprehensive view of the system's behavior, offering a scalable approach which goes beyond the limited numbers of fluorescent probes that can be measured simultaneously. Recently, RNA-seq has been used to characterize every transcriptional component in a large logic circuit composed of 46 genetic parts (Gorochowski *et al*, 2017). While successful in demonstrating the ability to characterize genetic part function, observe internal transcriptional states, and find the root cause of circuit failures, the use of RNA-seq alone restricts the method to purely transcriptional elements and does not allow for quantification of this process in physically meaningful units.

Here, we address these limitations by combining Ribo-seq with a modified version of RNA-seq to quantitatively characterize genetic parts controlling transcription and translation at a nucleotide resolution. By supplementing the sequencing data with other experimentally measured cell parameters, we are able to generate transcription and translation profiles that capture the flux of RNA polymerases (RNAPs) and ribosomes governing these processes in absolute units. We apply our method to *Escherichia coli* and demonstrate how local changes in these profiles can be

interpreted using biophysical models to measure the performance of five different types of genetic part in absolute units. Finally, we demonstrate how genome-wide shifts in transcription and translation can be used to dissect the burden that synthetic genetic constructs place on the host cell and the role that competition for shared cellular resources, such as ribosomes, plays.

Results

Generating transcription and translation profiles in absolute units

To enable quantification of both transcription and translation in absolute units, we modified the RNA-seq protocol and extended the Ribo-seq protocol with quantitative measurements of cellular properties (red elements in **Figure 1A**). For RNA-seq, we introduced a set of RNA spike-ins to our samples at known molar concentrations before the random alkaline fragmentation of the RNA (left panel, **Figure 1A**). The RNA spike-ins span a wide range of lengths (250–2000 nt) and concentrations and share no homology with the transcriptome of the host cell (**Supplementary Figure S1**). Using the known concentrations of the RNA spike-ins, the mapped reads can be converted to absolute molecule counts and then normalized by cell counts give absolute transcript copy numbers per cell (Bartholomäus *et al*, 2016; Mortazavi *et al*, 2008) (**Materials and Methods**). The total number of transcripts per cell was ~8200 which well correlates with earlier measurements of ~7800 mRNA copies/per cell using a single spike-in (Bartholomäus *et al*, 2016). Similar overall copy numbers have been theoretically predicted (Bremer *et al*, 2003) and experimentally determined for another *E. coli* strain (Taniguchi *et al*, 2010). For Ribo-seq, we directly ligated adaptors to the extracted ribosome-protected fragments (RPFs) (Guo *et al*, 2010) to capture low-abundance transcripts (Del Campo *et al*, 2015). Sequencing was also complemented with additional measurements of cell growth rate, count, and protein mass (right panel, **Figure 1A**).

A previous method was employed to generate the transcription profiles that capture the number of RNAPs passing each nucleotide per unit time across the entire genome (i.e. the RNAP flux). This assumes that RNA levels within the cells have reached a steady-state (Gorochowski *et al*, 2017) and that all RNAs have a fixed degradation rate (0.0067 s^{-1}) (Chen *et al*, 2015) so that RNA-seq data, which captures a snapshot of relative abundances of RNAs, can be used to estimate relative RNA synthesis rates (Gorochowski *et al*, 2017). Because each RNA is synthesized by an RNAP these values are equivalent to the relative RNAP flux. By using the known molar concentrations of the RNA spike-ins and their corresponding RNA-seq reads from our modified protocol (**Supplementary Figure S1**), we are able to convert the transcription profiles into RNAP/s units. Existing biophysical models of promoters and terminators were then used to interpret changes

in the transcription profiles and infer the performance of these parts in absolute units, similar to previous work (Gorochowski *et al*, 2017).

To generate the translation profiles that capture the ribosome flux per transcript, we first took each uniquely mapped RPF read from the Ribo-seq data and considering the architecture of a translating ribosome we estimated the central nucleotide of each codon in the ribosomal P site (e.g. the peptidyl-tRNA site) (**Materials and Methods**) (Mohammad *et al*, 2016). By summing these positions for all reads at each nucleotide *x*, we computed the RPF coverage *N*(*x*). If we assume that each ribosome translates at a relatively constant speed, which holds true in most cases (Gorochowski *et al*, 2015; Li *et al*, 2014), then the RPF coverage is proportional to the number of ribosomes at each nucleotide at a point in time and thus captures relative differences in ribosome flux; more heavily translated regions will have a larger number of ribosomes present and so accrue a larger number of RPF reads in the Ribo-seq snapshot.

We next needed to convert the RPF coverage into a translation profile whose height corresponds directly to the ribosome flux in ribosomes/s units. By assuming that each RPF read corresponds to an actively translating ribosome which synthesizes a full-length protein product and that the cellular proteome is at steady-state, then the protein copy number for gene *i* is given by $n_i = \frac{f_i m_t}{f_t m_i}$. Here, f_t is the total number of mapped RPF reads, m_t is the total protein mass per cell, and f_i and m_i are the number of mapped RPF reads and the protein mass of gene *i*, respectively. We measured m_t directly (**Figure 1A**) and calculated m_i from the amino acid sequence of gene *i* (**Materials and Methods**). Because proteins are synthesized by incorporating individual amino acids during the translocation cycle (i.e. by ribosome translocating from the A to P site), the replication of the entire proteome requires $r_t = \sum_i n_i a_i$ ribosome translocations, where a_i is the number of amino acids in the protein encoded by gene *i*. Assuming that cells are growing at a constant rate with doubling time t_d, then the total ribosome flux across the entire transcriptome per unit time is given by $q = 3r_t/t_d$. The factor of three accounts for ribosomes translocating at three-nucleotide registers (i.e. 1 codon/s = 3 nt/s). Finally, the translation profile for nucleotide *x* is calculated by multiplying the total ribosome flux *q* by the fraction of active ribosomes $N(x)/f_t$ at that position and normalizing by the number of transcripts per cell of the gene being translated m_x, computed from the RNA-seq data (**Figure 1A**). This gives,

$$R(x) = \frac{q \cdot N(x)}{m_x f_t}. \quad (1)$$

Importantly, because both the transcription and translation profiles are given in absolute units (RNAP/s and ribosomes/s, respectively), they can be directly compared across samples without any further normalization.

Characterizing genetic parts controlling translation

Genetic parts controlling translation alter ribosome flux along a transcript and these changes are captured by the translation profiles. We developed biophysical models to interpret these signals and quantify the performance of RBSs, stop codons and translational recoding (e.g. ribosome frameshifting) in open reading frames (ORFs) at stable secondary structures.

In prokaryotes, RBSs facilitate translation initiation and cause a jump in the translation profile after the start codon of the associated gene due to an increase in ribosome flux originating at that location (**Figure 1B**). If initiation is rate limiting (Li *et al*, 2014), then the translation initiation rate of an RBS (in ribosomes/s units) is given by the increase in ribosome flux downstream of the RBS,

$$\delta R = \sum_{i=x_s}^{x_e} \frac{R(i)-C(i)}{(x_e-x_s)} - \sum_{i=x_0-n}^{x_0} \frac{R(i)-C^- -C^+}{n}. \quad (2)$$

Here, x_0 is the start point of the RBS, and x_s and x_e are the start and end point of the protein coding region associated with the RBS, respectively (**Figure 1B**). A window of n = 30 nt (10 codons) was used to average fluctuations in the translation profile upstream of the RBS; the averaging window is equal to the approximate length of a ribosome footprint. If the transcription start site (TSS) of the promoter expressing this transcript fell in the upstream window, then the start point ($x_0 - n$) was adjusted to the TSS to ensure that the incoming ribosome flux is not underestimated. A similar change was made if the coding sequence was within an operon and the end of an upstream gene falls in this window. In this case, the start point was adjusted to 9 nt (3 codons) downstream of the stop codon of the overlapping gene. We also included correction factors to remove the effect of translating ribosomes upstream of the RBS that are not in the same reading frame as the RBS-controlled ORF, and therefore may not fully traverse the coding sequence due to out-of-frame stop codons. These are given by,

$$c^- = \sum_{i=0}^{(x_0-n)/3} \frac{R(x_0-n+3i+2)}{(x_0-n)/3}, \quad (3)$$

$$c^+ = \sum_{i=0}^{(x_0-n)/3} \frac{R(x_0-n+3i+1)}{(x_0-n)/3}, \quad (4)$$

$$C(x) = \begin{cases} c^- + c^+, & x < s^- \wedge x < s^+ \\ c^-, & x < s^- \wedge x \geq s^+ \\ c^+, & x \geq s^- \wedge x < s^+ \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where s^- and s^+ are the positions of the first out-of-frame stop codon downstream of $x_0 - n$ in the –1 and +1 reading frame, respectively. c^- and c^+ capture the average out-of-frame ribosome flux in the region upstream of the RBS in the –1 and +1 reading frame, respectively, and $C(x)$ calculates the total sum of these ribosome fluxes that would reach nucleotide x downstream of the RBS.

In eukaryotes, genes are generally monocistronic and translation initiation occurs through scanning of the 5' untranslated region (5'-UTR) by the 43S preinitiation complex until a start codon is reached. This allows a translation-competent 80S ribosome to assemble and translation elongation to begin (Jackson *et al*, 2010). In this case, no ribosome flux is generated by upstream genes. Therefore, when calculating the initiation rate of a 5'-UTR, the second term in Equation 2 and the correction factors are set to zero (i.e. $\delta R = \sum_{i=x_s}^{x_e} \frac{R(i)}{(x_e - x_s)}$).

Ribosomes terminate translation and disassociate from a transcript when a stop codon (TAA, TAG or TGA) is encountered. This leads to a drop in the translation profile at these points (**Figure 1C**). Although this process is typically efficient, there is a rare chance that some ribosomes may read through a stop codon and continue translating downstream (Arribere *et al*, 2016). Assuming that all ribosomes translating the protein coding region are in-frame with the associated stop codon and do not frameshift prior to it, then the termination efficiency of the stop codon (i.e. the fraction of ribosomes terminating) is given by,

$$T_e = 1 - \frac{\sum_{i=x_1}^{x_1+n} R(i)/n}{\sum_{i=x_s}^{x_0} R(i)/(x_o - x_s)}\,, \qquad (6)$$

where x_0 and x_1 are the start and end nucleotide of the stop codon, respectively, x_s is the start of the coding region associated to this stop codon, and n = 30 nt (10 codons) is the window, with the same width as described above, used to average fluctuations in the translation profile downstream of the stop codon (**Figure 1C**). If additional stop codons are present in the downstream window, the end point of this window (x_1 + n) was adjusted to ensure that the termination efficiency of only the (first) stop codon was measured. A similar adjustment was made if the end of a transcript generated by an upstream promoter ends within this window.

Translation converts the information encoded in mRNA into protein whereby each triplet of nucleotides (a codon) is translated into a proteinogenic amino acid. Because of the three-nucleotide periodicity in the decoding, each nucleotide could be either in the first, second or third position of a codon, thus defining three reading frames for every transcript. Consequently, a single mRNA sequence can encode three different proteins. Although synthetic biology rarely use multiple reading frames, natural systems exploit this feature in many different ways (Giedroc & Cornish, 2009; Condron *et al*, 1991a; Tsuchihashi & Kornberg, 1990; Bordeau & Felden, 2014). In our workflow, the RPFs used to generate the translation profiles were aligned to the middle nucleotide of the codon residing in the ribosomal P site, providing the frame of translation. To characterize genetic parts that cause translational recoding through ribosomal frameshifting, we compared regions directly before and after the part. Strong frameshifting will cause the fraction of RPFs to shift from the original frame to a new one when comparing these regions with the frameshifting efficiency given by,

$$F_e = 1 - \frac{\sum_{i=x_1}^{x_e} R(i)/(x_e - x_1)}{\sum_{i=x_s}^{x_0} R(i)/(x_o - x_s)}. \quad (7)$$

Here, x_0 is the nucleotide at the start of the region where frameshifting occurs, and x_1 is the end nucleotide of the stop codon for the first coding sequence (**Figure 1D**).

Measuring genome-wide translation initiation and termination in Escherichia coli

We applied our approach to *Escherichia coli* cells harboring a *lacZ* gene whose expression is induced using isopropyl β-D-1-thiogalactopyranoside (IPTG) (**Figure 2A**). After induction for 10 min, *lacZ* expression reached 14% of the total cellular protein mass (**Supplementary Table S1**). Samples from non-induced and induced cells were subjected to the combined sequencing workflow (**Figure 1A**). Sequencing yielded between 41–199 million reads per sample (**Supplementary Table S2**) with no measurable bias across RNA lengths and concentrations (**Supplementary Figure S1**), and a high correlation in endogenous gene expression between biological replicates (R^2 > 0.96; **Supplementary Figure S2**).

Transcription and translation profiles were generated from this data and used to measure translation initiation rates of RBSs and termination efficiencies of stop codons across the genome. To remove the bias due to the RPF enrichment in the 5'-end of coding regions (Ingolia *et al*, 2009) (**Figure 2B**), x_s was adjusted to 51 bp (17 codons) downstream of the start codon when estimating average ribosome flux across a coding region in Equations 2 and 6. To determine whether translation rates were constant across each gene, we compared the number of RPFs mapping to the first and second half of each coding region. This is a necessary condition for our models to ensure that changes in the height of a translation profile between two different points is purely a result of initiating or terminating ribosomes. If the speed of a translating ribosome varies along a transcript, then regions of slower movement would be enriched in RPFs, resulting in an increase in the translation profile at those points. This would make it is impossible from the translation profile alone to distinguish between changes in ribosome speed and the rate of initiation/termination events. If the ribosomes traverse the coding sequence at a constant speed, then the two halves of a transcript should have a near identical RPF coverage. We found a high correlation between both halves for non-induced and induced cells suggesting a constant speed of the ribosomes across the coding sequences (**Supplementary Figure S3**).

We characterized chromosomal RBSs in *E. coli* by assuming that each covered a region spanning 15 bp upstream of the start codon. The translation initiation rates of the 761 RBSs we measured varied over two orders of magnitude with a median initiation rate of 0.1 ribosome/s (**Figure 2C**; **Supplementary Data S1**). This closely matches previously measured rates for single

genes (Kennell & Riezman, 1977). A few RBSs mostly related to stress response functions (*tabA*, *hdeA*, *uspA*, *uspG*), the ribosomal subunit protein L31 (*rpmE*), and some unknown genes (*ydiH*, *yjdM*, *yjfN*, *ybeD*), reached much higher rates of up to 2.45 ribosomes/s.

To estimate termination efficiency at chromosomal stop codons, we considered that they spanned 9 nt up and downstream of the stop codon (**Figure 2B**). We also excluded overlapping genes and those bearing internal sites that promote frameshifting, both of which break the assumptions of our model (Baggett *et al*, 2017). In total, the termination efficiency of 746 stop codons was measured and their median termination efficiency across the genome was found to be 0.987, with 336 of them (45% of all measured) having termination efficiencies >0.99 (**Figure 2D**; **Supplementary Data S2**). Similar performance for both RBSs (R^2 = 0.84) and terminators (R^2 = 0.52) was found between non-induced and induced conditions (**Figures 2E** and **2F**; **Supplementary Data S1** and **S2**).

Quantifying differences in transcription and translation of endogenous and synthetic genes

The quantitative measurements produced by our methodology allow both transcription and translation to be monitored simultaneously. To demonstrate this capability, we first focused on differences in the contributions of transcription and translation to overall protein synthesis rates of endogenous genes in *E. coli*. For each gene we calculated the protein synthesis rate by multiplying the transcript copy number by the RBS-mediated translation initiation rate per transcript. We found a strong correlation with previously measured synthesis rates (Li *et al*, 2014) (**Figure 3A**). We also extracted the transcription and translation profiles of three genes (*uspA*, *ompA* and *gapA*) whose protein synthesis rate was similar, but whose expression was controlled differently at the levels of transcription and translation (**Figures 3B**). Quantification of the promoters and RBSs for these three genes showed more than an order of magnitude difference in their transcription and translation initiation rates; *uspA* was weakly transcribed and highly translated, *ompA* was highly transcribed and weakly translated, and *gapA* was moderately transcribed and translated (**Figure 3C**).

Because we measure transcription and translation initiation rates in absolute units, it was also possible to determine their relative contribution to the final synthesis rate by calculating the ratio of transcription and translation initiation rates, giving RNAPs/ribosomes. High RNAP/ribosome values relate to genes whose expression is mostly controlled by transcription, while low values correspond to a greater contribution by translation. This analysis revealed a non-uniform split with a trend for weakly expressed genes to be mostly governed by translation, while strongly expressed genes were mostly controlled by transcription (**Figure 3A**).

These different modes of gene expression have a major influence on the efficiency of protein synthesis (Ceroni *et al*, 2015) and can influence the variability in protein levels between cells (Raser & O'Shea, 2005). For example, the most metabolically efficient way to strongly express a protein of interest is by producing very high numbers of transcripts (high transcription initiation rate and stable transcript) with a relatively weak RBS (low translation initiation rate). This ensures that each ribosome initiating on a transcript has a very low probability of colliding with others, guaranteeing efficient translation elongation. Indeed, we observe this efficient expression strategy is enriched for strongly expressed endogenous genes (**Figure 3A**).

We next sought to demonstrate the ability to measure dynamic changes in the function of regulatory parts using the LacZ construct. We quantified the inducible promoter and terminator controlling transcription, and the RBS and stop codon controlling translation when the inducer IPTG was absent and present. The transcription and translation profiles clearly showed the beginning and end of both the transcript and protein coding region, with sharp increases and decreases at transcriptional/translational start and stop sites (**Figure 3D**). Induction caused a large increase in the number of *lacZ* transcripts from 0.18 to 110 copies per cell, which was directly observed in the transcription profiles. In contrast, the translation profiles remained stable across conditions. The P_{tac} promoter initiated transcription at a rate of 0.0009 RNAP/s in the absence and 0.73 RNAP/s in the presence of IPTG (1 mM) (**Figure 3E**). The RBS for the *lacZ* gene had consistent translation initiation rates of between 0.21 and 0.35 ribosomes/s, respectively (**Figure 3E**). It may seem counterintuitive to observe translation without IPTG induction because very few transcripts will be present. However, leaky expression from the P_{tac} promoter was sufficient to capture enough RPFs during sequencing to generate a translation profile. It should be noted that the translation profile represents the ribosome flux per transcript, thus its shape was nearly identical to that when the P_{tac} promoter was induced. Like the RBS, both the transcriptional terminator and stop codon showed similar efficiencies of 0.93–0.95 and 0.9–0.93, respectively (**Figure 3E**).

Characterizing a synthetic pseudoknot that induces translational recoding

Pseudoknots (PKs) are stable tertiary structures that regulate gene expression. They are frequently combined with slippery sequences in compact viral genomes to stimulate translational recoding and produce multiple protein products from a single gene (Giedroc & Cornish, 2009; Brierley *et al*, 2007; Sharma *et al*, 2014; Tsuchihashi & Kornberg, 1990). The percentage of recoding events generally reflects the stoichiometry of the translated proteins (e.g. capsule proteins for virus assembly), and helps overcome problems where the stochastic nature of transcription and translation make maintenance of specific ratios difficult (Condron *et al*, 1991a). PKs are the most common type of

structure used to facilitate mostly –1 frameshifting (Atkins *et al*, 2016) and in much rarer cases +1 frameshifting (e.g., in eukaryotic antizyme genes) (Ivanov *et al*, 2004). PKs consist of a hairpin with an additional loop that folds back to stabilize the hairpin via extra base pairing (**Figure 4A**). In addition to stimulating recoding events, PKs regulate translational initiation, where they interfere with an RBS through antisense sequences that base pair with the RBS (Unoson & Wagner, 2007; Bordeau & Felden, 2014). They also act as an evolutionary tool, reducing the length of sequence needed to encode multiple protein coding regions and therefore act as a form of genome compression.

Two elements signal and stimulate frameshifting. The first is a slippery site consisting of a heptanucleotide sequence of the form XXXYYYZ which enables out-of-zero-frame paring in the A or P site of the ribosome, facilitating recoding events. The second is a PK situated 6–8 nt downstream of the slippery site. In bacteria, the distance between the slippery site and the 5'-end of the PK positions mRNA in the entry channel of the 30S ribosomal subunit, enabling contact with the PK which pauses translation and provides an extended time window for frameshifting to occur (Giedroc & Cornish, 2009).

To demonstrate our ability to characterize this process, we created an inducible genetic construct (referred to as PK-LacZ) that incorporated a virus-inspired PK structure within its natural context (*gene10*) fused to *lacZ* in a –1 frame (**Figure 4A**) (Tholstrup *et al*, 2012). A slippery site UUUAAAG preceded the PK. *Gene10* of bacteriophage T7 produces two proteins, one through translation in the zero frame and one through a –1 frameshift; both protein products constitute the bacteriophage capsid (Condron *et al*, 1991b). We generated translation profiles to assess ribosome flux along the entire construct (**Figure 4B**). These showed high-levels of translation up to the PK with a major drop of 80–90% at the PK to the end of the *gene10* coding region, and a further drop of ~97% after this region (**Figure 4B**). To analyze frameshifting within *gene10*, we divided the construct into three regions: (1) the *gene10* gene up to the slippery site, (2) the middle region, which covers the slippery site along with the PK up to the *gene10* stop codon, and (3) the downstream *lacZ* gene in a –1 frame. For each region, we calculated the fraction of RPFs in each frame as a total of all three possible frames. We found that the zero and –1 frames dominate the *gene10* and *lacZ* regions, respectively, with >46% of all RPFs being found in these frames (top row, **Figure 4C**). The middle region saw a greater mix of all three, and the zero-frame further dropped in the *lacZ* region. This is likely due to a combination of ribosomes that have passed the PK successfully and terminated in zero-frame at the end of *gene10* and those that have frameshifted. Similar results were found with and without induction by IPTG (**Figure 4C**). An identical analysis of the reading frames from the RNA-seq data revealed that no specific frame was preferred with equal fractions of

each (bottom row, **Figure 4C**). This suggests that the reading frames recovered for the RPFs were not influenced by any sequencing bias. We further tested if the major translation frame could be recovered by analyzing the entire genome and measured the fraction of each frame across every gene. The correct zero-frame dominated in most cases (**Figure 4D**).

Finally, to calculate the efficiency of frameshifting by the PK, we compared the density of RPFs per nucleotide for the middle and *lacZ* regions. Because the PK causes ribosome stalling, the assumption of constant ribosome speed is broken for the *gene10* region upstream of the PK. Therefore, when calculating the frameshifting efficiency using Equation 7, x_s and x_0 were set to the start and end nucleotide of the middle region, directly downstream of the PK where pausing was not expected to occur. We found that the PK caused 2–3% of ribosomes to frameshift, ~3-fold less than the 10% reported for the PK in its natural context (Condron *et al*, 1991a).

Cellular response to a strong synthetic pseudoknot

Expression of strong PKs can severely impact cell growth, but the reason for this remains unclear (Tholstrup *et al*, 2012). We observed a large number of RPF reads within the *gene10* region (**Figure 4B**) and many of these reads capture stalled ribosomes. Stalling increases the abundance of partially synthesized protein products but also limits the availability of translational resources, raising the question as to whether expression of the PK-LacZ construct elicits cellular stress by sequestering ribosomes. To better understand the burden that expression of both *lacZ* and *PK-lacZ* exhibited on the cell, we compared shifts in transcription (i.e. mRNA counts) and translation efficiency (i.e. density of ribosome footprints per mRNA) of endogenous genes following induction with IPTG (**Figure 5A**; **Supplementary Data S3**). No major changes were observed for the LacZ construct (**Figure 5A**). In contrast, the PK-LacZ construct caused significant shifts in the expression of 491 genes (**Supplementary Data S4**). Of these, 341 were transcriptionally (i.e. significant changes in mRNA counts) and 204 translationally regulated (i.e. significant changes in translational efficiency), with little overlap (54 genes) between the two types of regulation (**Figure 5B**). Of the transcriptionally regulated genes, most saw a drop in mRNA counts, while translationally regulated genes were split between increasing and decreasing translational efficiencies. Gene ontology (GO) analysis revealed a clustering of transcriptionally downregulated genes in categories mostly linked to translation, e.g. ribosomal proteins, amino acid biosynthesis, amino acid activation (aminoacyl synthetases), and genes involved in respiration and catabolism (**Supplementary Data S5**). Transcriptionally upregulated genes were associated with ATP binding, chaperones (*ftsH*, *lon*, *clpB*, *dnaJK*, *groLS*, *htpG*), ion binding, proteolytic activities (*ftsH*, *prlC*, *htpX*), and an endoribonuclase (*ybeY*). Interestingly, the expression of all of these are under σ^{32} regulation which is the most

common regulatory mode to counteract heat stress. σ^{32} upregulation is often observed by expressing synthetic constructs, although the precise mechanism of σ^{32} activation is not known (Ceroni *et al*, 2018). In our case, the incompletely synthesized polypeptides from the stalled ribosomes on the *PK-LacZ* mRNA are most likely partially folded or misfolded and generate misfolding stress similar to the heat shock response. Binding of the major *E. coli* chaperone systems, DnaK/DnaJ and GroEL/S, to the misfolded proteins negatively regulates σ^{32}. The shift of the chaperones to misfolded proteins releases σ^{32}, which then binds to the RNA polymerases and induces expression of heat shock genes (Mogk *et al*, 2011; Guisbert *et al*, 2004). This notion is supported by the fact that *dnaJ*, *groL/S*, and *grpE* were transcriptionally upregulated during PK induction as well as *ftsh*, which encodes the protease that degrades σ^{32}.

To test whether *PK-lacZ* expression caused changes in translation dynamics (e.g. ribosome pausing at particular codons), we next computed the dwell time of ribosomes at each codon (also known as codon occupancy) across the genome and compared it to that without inducing *PK-lacZ* expression (Lareau *et al*, 2014). Notable increases in occupancy were found for the codons AGA, CTA, CCC, TCC, which encode for arginine, leucine, proline and serine, respectively (**Figure 5C**). All of these codons are rarely used in the genome for their cognate amino acid but were found in higher proportions across *gene10*. For example, the CTA codon that codes for leucine is only used by 4% of codons in the genome, while accounting for 8% of the *gene10* region. Coupled with the strong expression of *gene10*, the stress induced by this abnormal demand on cellular resources would be amplified.

The broad shifts in regulation at a cellular-scale and changes in codon occupancy suggest that *PK-lacZ* expression may significantly limit the availability of shared cellular resources. From a translational perspective, this would manifest as a cell-wide drop in translation initiation rates as the pool of free ribosomes would be reduced (Gorochowski *et al*, 2016). To test this hypothesis, we compared the RBS initiation rates of endogenous genes before and after induction of *lacZ* and *PK-lacZ* expression and found a consistent reduction across all genes for both synthetic constructs (**Figure 5D**; **Supplementary Data S1**). While relatively small for the LacZ construct (18%) where no notable stress response was detected, the PK-LacZ construct triggered a large (43%) drop in translation initiation rates across the cell (**Figure 5D**). Analysis of the transcriptome composition and distribution of engaged ribosomes across cellular transcripts further revealed that the PK-LacZ construct made up 40% of all mRNAs and captured 47% of the shared ribosome pool engaged in translation (**Figure 5E**). This would account for the global drop in translation initiation rates and misfolding stress induced by the partially translated proteins from *gene10* transcripts, explaining the strong σ^{32}-mediated response.

We also observed a large difference in the number of transcripts for each construct after induction; the *lacZ* transcripts were 43-fold lower than those for *PK-lacZ* (81 vs. 3504 transcripts/cell, respectively). Such a difference is unlikely to occur solely through an increased transcription initiation rate at the P_{tac} promoter. Previous studies have shown that the decay rate of the *lacZ* transcript is highly dependent on the interplay between transcription and translation rates (Yarchuk *et al*, 1992; Makarova *et al*, 1995; Iost & Dreyfus, 1995). RNase E sites within the coding region become accessible to cleavage by RNase E when translation initiation rates are low because fewer translating ribosomes are present to sterically shield these sites and prevent degradation (Yarchuk *et al*, 1992). This mechanism could account for the lower *lacZ* transcript numbers, which in turn would reduce the number of sequestered ribosomes translating *lacZ* mRNAs and explain the lack of a stress response for this construct.

Discussion

In this work, we present new approach to quantify transcription and translation in living cells at a nucleotide resolution. This is based on a deep-sequencing workflow that combines a modified version of RNA-seq and Ribo-seq with measures of key cellular parameters and uses biophysical models to interpret this data (**Figure 1**). We show that our high-throughput approach is able to simultaneously characterize the translation initiation rate of 743 RBSs and termination efficiency of 746 stop codons across the *E. coli* transcriptome (**Figure 2**), in addition to measuring the precise behavior of the genetic parts controlling transcription and translation of several endogenous genes and a synthetic genetic construct that expresses *lacZ* (**Figure 3**). Because our methodology is based on sequencing, it can scale beyond the number of simultaneous measurements that are possible with common fluorescence-based approaches, and through the use of spike-in standards we are able to extract part parameters in absolute units (i.e. transcription and translation rates in RNAP/s and ribosomes/s units, respectively).

To demonstrate the ability to quantitatively assess various translational processes that have been difficult to measure, we studied the behavior of a genetic construct that contains a strong virus-inspired PK structure that induces a translational frameshift (**Figure 4**). Following expression of *PK-lacZ*, the main reading frame shifts, but the efficiency is ~3-fold lower than the PKs native viral context. In contrast to *lacZ* expression, *PK-lacZ* also causes a major burden to the cell, sequestering a large proportion of the shared gene expression machinery, e.g. ribosomes (**Figure 5**). We observe transcriptome-wide increases in ribosome dwell times at codons rare for the *E. coli* endogenous genes, but more frequent in the synthetic construct, suggesting that the strong expression of this gene places significant demands on the translational resources of the cell. This burden also results

in significant changes in gene regulation (both transcriptional and translational), which was mediated by the alternative polymerase subunit, σ^{32} that remodels the bacterial proteome following thermal stress (Guo & Gross, 2014a). The likely cause of σ^{32} activation is a combination of strong overexpression of *gene10* and misfolding stress triggered by partial unfolding of incompletely synthesized polypeptides (Giedroc & Cornish, 2009; Guo & Gross, 2014b). To our knowledge the stress response induced by a strong pseudoknot has not been reported before making this work a valuable data set for future studies.

Previous studies have used sequencing to investigate translational regulation. Ribo-seq was employed by Li *et al.* (Li *et al*, 2014) to measure the protein synthesis rate of 3,041 genes and by Baggett *et al.* (Baggett *et al*, 2017) to analyze translation termination at 1200 stop codons. However, unlike our approach, which is calibrated by external RNA spike-in standards, these previous studies had no means of assessing the sensitivity of their measurements. Measuring the variability of several different RNA spike-in molecules at similar known molar concentrations allows us to accurately calculate a detection limit, emphasizing the benefit of including external standards in sequencing experiments.

A limitation of our approach is that the models underpinning the generation and interpretation of the transcription and translation profiles rely on some key assumptions that may not always hold true. For the transcription profiles to accurately capture RNAP flux it is essential that the system has reached a steady-state because RNA-seq only measures RNA abundance at a single point in time and not directly the rate of RNA production (Gorochowski *et al*, 2017). While this assumption is valid for cells that have been exponentially dividing for several generations, rapidly changing RNA production or degradation rates (e.g. through increased expression of degradation machinery or a change in growth phase) may cause issues. Furthermore, for quantification of absolute transcript numbers, while the RNA spike-ins will undergo the same depletion during sequencing library preparation, it is necessary to assume that the total RNA from the cells is efficiently extracted prior to this step. Incomplete cell lysis or low-efficiency RNA extraction would require a further correction during the quantification process.

For the translation profiles, the key assumptions are that every ribosome footprint gives rise to a full-length protein and that translation elongation globally proceeds at a near uniform speed along all transcripts. Translation is a complex multi-step process and can be affected by ribosome pausing (e.g. due to amino acid charge) (Charneski & Hurst, 2013), premature termination (Freistroffer *et al*, 2000), and environmental conditions that alter cell physiology (Bartholomäus *et al*, 2016) or the global availability of cellular resources (e.g. ribosomes, tRNAs, amino acids) (Dong *et al*, 1996; Wohlgemuth *et al*, 2013; Gorochowski *et al*, 2016). Although these factors normally have

only a small effect (Ingolia *et al*, 2009; Li *et al*, 2014), significant genome-wide shifts induced by long-term chronic stress can increase their occurrence and potentially alter translation elongation speed and processivity in a non-uniform way (Bartholomäus *et al*, 2016). Our calculation of absolute protein synthesis rates also relies on the assumption that proteins are stable with dilution by cell division dominating their degradation rate (Li *et al*, 2014). This holds for most proteins, but care should be taken under stress conditions or for synthetic constructs where the proteome is heavily modified (e.g. by overexpressing proteases).

Being able to measure RNAP and ribosome flux across multi-component genetic circuits offers synthetic biologists a powerful tool for designing and testing new living systems (Nielsen *et al*, 2016; Gorochowski *et al*, 2017). These capabilities are particularly useful for large genetic circuits where many parts must function together to generate a required phenotype. Ideally, complex circuits are built by readily connecting simpler parts together. In electronics this is made possible by using the flow of electrons as a common signal that captures the state at every point in a circuit. This signal can be easily routed between components using conductive wires to create more complex functionalities. In genetic circuits, RNAP and ribosome fluxes can serve a similar role acting as common carrier signals (Canton *et al*, 2008; Brophy & Voigt, 2014). Promoters and RBSs guide these signals to particular points in a circuit's DNA/RNA and allow them to propagate and be transformed.

The ability to easily connect large numbers of genetic parts allows for the implementation of more complex functionalities, but can also lead to fragile circuits that break easily (Nielsen *et al*, 2016). This is particularly common for those that use components with sharp switch-like transitions (e.g. repressors with high cooperativity) (Nielsen *et al*, 2016). These types of part can lead to situations where although the output of the circuit behaves as desired, it becomes highly sensitive to changes in growth conditions or the inclusion of other genetic components (Gorochowski *et al*, 2017). This problem arises because the state of these parts can fall close to their sharp transition point allowing for minor perturbations to cause large deviations in expression that then propagate to the output of the circuit. The only way to ensure the robustness of such systems is to measure every internal state (Gorochowski *et al*, 2017) or to implement feedback control within the circuit itself to enable self-regulation (Ceroni *et al*, 2018). The ability to monitor every element in a circuit also makes our approach valuable when elucidating the root cause of failures. Instead of time-consuming tinkering with a circuit until the problem is found, our method allows doe targeted modifications that precisely correct malfunctioning parts, accelerating developments in the field (Gorochowski *et al*, 2017).

Mature engineering fields rely on predictive models to efficiently develop complex systems by reducing the need to physically construct and test each design. To date, the accuracy of models in synthetic biology have been hampered by a lack of reliable, quantitative and high-throughput measurements of genetic parts and devices, as well as their effects on the host cell. Attempts have been made to improve this situation by using standard calibrants to increase reproducibility across labs and equipment (Castillo-Hair *et al*, 2016; Davidsohn *et al*, 2015; Beal *et al*, 2016) and by including synthetic RNA spike-ins to enable absolute quantification of transcription (Owens *et al*, 2016). Our methodology complements these efforts by combining RNA-seq and Ribo-seq with RNA spike-in standards to quantify the regulation of transcription and translation by genetic circuits. The importance of pushing biology towards measurements in absolute units has also seen growing interest (Justman, 2018) and is becoming widely recognized as essential for developing mechanistic models that can support reliable predicative design (Jones *et al*, 2014; Belliveau *et al*, 2018; Endy *et al*, 2000). To demonstrate why, it is important to realize that many behaviors are intrinsically linked to their absolute scale. For example, the stochastic nature of biochemical reactions means that the inherent noise when only a few molecules are present will be far greater than when there are many. Therefore, knowing if one arbitrary unit corresponds to one or 10,000 molecules is essential if the models are to hold any predictive power as to the expected variability. The use of absolute measurements in mechanistic models of biological parts (Belliveau *et al*, 2018; Jones *et al*, 2014) and entire genetic systems (Endy *et al*, 2000) has already seen some success.

As we attempt to implement ever more complex functionalities in living cells (Nielsen *et al*, 2016) and push towards a deeper understanding of the processes sustaining life, scalable and comprehensive methodologies for quantitative measurement of fundamental processes become paramount. Such capabilities will move us beyond a surface-level view of living cells to one that allows the exploration of their inner most regulation and homeostasis.

Materials and Methods

Strains, media, and inducers. The *E. coli* K12 strain, [K-12, *recA1* Δ*(pro-lac) thi ara F':lacIq1 lacZ::Tn5 proAB*+], harbours a pBR322-derived plasmid containing either *lacZ* with a fragment insert that contains a truncated *lac* operon with the P_{tac} promoter and the wildtype *lacZ* under *lacI* control, or a pseudoknot-lacZ (*PK-lacZ*) consisting of *gene10*, a virus-derived RNA pseudoknot (Tholstrup *et al*, 2012), 22/6a, fused upstream of the *lacZ*. Bacteria were grown in MOPS minimal medium supplemented with 0.4% glycerol, 2.5 µg/ml vitamin B1, 100 µg/ml ampicillin, 20 µg/ml kanamycin and additionally 50 µg/ml arginine for the *lacZ* expressing strain. The cells were grown for at least 10 generations at 37°C to ensure stable exponential growth before induction.

Gene expression and preparation of the sequencing libraries. LacZ and *PK-lacZ* expression were induced with isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 1 mM at OD_{600} ≈ 0.4 for 10 min and 15 min, respectively. One aliquot of each culture was used to isolate RPFs and prepare the cDNA library for Ribo-seq as described in Bartholomäus *et al*. (Bartholomäus *et al*, 2016) In parallel, from another aliquot, total RNA was isolated with TRIzol (Invitrogen) and subjected to random alkaline fragmentation for RNA-seq as described in Bartholomäus *et al*. (Bartholomäus *et al*, 2016) Different than the previous protocol, prior to alkaline fragmentation, the total RNA was spiked in with RNA standards (ERCC RNA Spike-In Mix; Ambion) which were used to (a) determine the detection limit in each data set and (b) calculate the copy numbers per cell. The RNA standards consist of 92 different transcripts, covering lengths of 250-2000 nt and approximately a 106-fold concentration range. Detection threshold (RPKM) has been set at values with a linear dependence between the reads from the spike-in controls and concentration in each RNA-Seq data set. Spike-ins with linear correlation were used in the copy number analysis (**Supplementary Figure S1**). Total protein concentration (grams of wet mass per ml culture) were determined by the Bradford assay using serial dilutions of the exponentially growing cells at different time points (e.g. prior the induction time at OD 0.4 and following induction with 1 mM IPTG). Using the cell number and the volume of *E. coli* as 1 femtoliter, the protein mass was recalculated as grams of wet protein mass per cell.

Processing of sequencing data. Sequenced reads were quality trimmed using fastx-toolkit version 0.0.13.2 (quality threshold: 20), sequencing adapters were cut using cutadapt version 1.8.3 (minimal overlap: 1 nt) and the reads were uniquely mapped to the genome of *E. coli* K-12 MG1655 strain using Bowtie version 1.1.2 allowing for a maximum of two mismatches. *LacZ* and other similar parts of the plasmids were masked in the genome. Reads aligning to more than one sequence including tRNA and rRNA were excluded from the data. The raw reads were used to generate gene read counts by counting the number of reads whose middle nucleotide (for reads with an even length the nucleotide 5' of the mid-position) fell in the coding sequence (CDS). Gene read counts were normalized by the length of the unique CDS per kilobase (RPKM units) and the total mapped reads per million (RPM units) (Mortazavi *et al*, 2008). Biological replicates were performed for all sequencing reactions. Based on the high correlation between replicates (**Supplementary Figure S2**), reads from both biological replicates were merged into metagene sets (Ingolia *et al*, 2009). Differential gene expression was performed using DESeq2 version 1.20. Firstly, transcripts with $P < 0.01$ for both translational efficiency and mRNA expression were selected. *P*-values were adjusted

for multiple testing using false-discovery rate (FDR) according to Benjamini and Hochberg. Since the RNA-Seq data sets have very high reproducibility between replicates (**Supplementary Figure S1**), we decided to apply more restrictive threshold $P < 0.001$ and additionally selected the 25th percentile. The GO terms with significant enrichment ($P < 0.01$) were calculated using GO.db version 2.10.

Calculating absolute transcript numbers. To calculate the transcript copy number, we used a method previously described by Bartholomäus *et al*. and Mortazavi *et al*. (Bartholomäus *et al*, 2016; Mortazavi *et al*, 2008). Briefly, the mapped reads for a transcript were related to the total reads and the length of the transcriptome. The latter was determined using the molecules of all spike-in standards above the detection limit (**Supplementary Figure S1**) and was normalized by cell number.

Calibration of ribosome profiling reads. RPFs were binned in groups of equal read length, and each group was aligned to the stop codons as described previously by Mohammad *et al*. (Mohammad *et al*, 2016) For each read length we calculated the distance between the point a transcript leaves the ribosome and the middle nucleotide in the P site. This distance was used to determine the center of each P site codon along each mRNA. As expected, the majority of our sequence reads were 23–28 nt and these read lengths were used for the further analysis. The ribosome dwelling occupancy per codon over the whole transcriptome was calculated as described by Lareau *et al*. (Lareau *et al*, 2014), where the reads over each position within a gene were normalized to the average number of footprints across this gene. Metagene analysis of the ribosome occupancies within the start and stop codon regions was performed as described by Baggett *et al*. (Baggett *et al*, 2017) Thereby, only genes with at least 5 RPFs in the chosen window were considered. Overlapping genes were excluded from the analysis.

Data analysis and visualization. Data analysis was performed using custom scripts run with R version 3.4.4 and Python version 3.6.3. Plots was generated using matplotlib version 2.1.2 and genetic constructs were visualized using DNAplotlib version 1.0 (Der *et al*, 2017) with Synthetic Biology Open Language Visual (SBOLv) notation (Myers *et al*, 2017).

Data availability. Sequencing data from RNA-Seq and Ribo-Seq were deposited in the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra/) under accession number SRP144594.

Acknowledgements

We thank Alexander Bartholomäus for the initial mapping and earlier data analysis. This work was supported by BrisSynBio, a BBSRC/EPSRC Synthetic Biology Research Centre (grant BB/L01386X/1), a Royal Society University Research Fellowship (grant UF160357 to T.E.G.), the MOLPHYSX program of the University of Copenhagen (S.P.), and by the European Union (grants NICHE ITN and SynCrop ETN to Z.I.)

Author Contributions

Z.I. and T.E.G. conceived of the study. M.E. performed the sequencing experiments, P.N. performed the quantitative determination of cellular parameter. S.P. provided the LacZ and PK-LacZ constructs and advised the experimental acquisition of sequencing data. T.E.G. developed the biophysical models. I.C. processed the sequencing data. Z.I., T.E.G. and I.C. analyzed the data. Z.I., T.E.G. and I.C. wrote the manuscript.

Conflict of Interest

The authors declare no competing financial interest.

References

Arribere JA, Cenik ES, Jain N, Hess GT, Lee CH, Bassik MC & Fire AZ (2016) Translation readthrough mitigation. *Nature* **534:** 719–723

Atkins JF, Loughran G, Bhatt PR, Firth AE & Baranov P V. (2016) Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. *Nucleic Acids Res.* **44:** 7007–7078

Baens M, Noels H, Broeckx V, Hagens S, Fevery S, Billiau AD, Vankelecom H & Marynen P (2006) The dark side of EGFP: Defective polyubiquitination. *PLoS One* **1:** 1–6

Baggett NE, Zhang Y & Gross C (2017) Global analysis of translation termination in E. coli. *PLoS Genet.* **13:** e1006676

Bartholomäus A, Fedyunin I, Feist P, Sin C, Zhang G, Valleriani A & Ignatova Z (2016) Bacteria differently regulate mRNA abundance to specifically respond to various stresses. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **374:** 20150069

Beal J, Haddock-Angelli T, Gershater M, De Mora K, Lizarazo M, Hollenhorst J, Rettberg R, Demling P, Hanke R, Osthege M, Schechtel A, Sudarsan S, Zimmermann A, Gabryelczyk B, Ikonen M, Salmela M, Acar M, Aktas MF, Bestepe F, Ceylan FS, et al (2016) Reproducibility of fluorescent expression from engineered biological constructs in E. coli. *PLoS One* **11:** 1–22

Belliveau NM, Barnes SL, Ireland WT, Jones DL, Sweredoski MJ, Moradian A, Hess S, Kinney JB & Phillips R (2018) Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proc. Natl. Acad. Sci.* **115:** E4796–E4805

Bordeau V & Felden B (2014) Curli synthesis and biofilm formation in enteric bacteria are controlled by a dynamic small RNA module made up of a pseudoknot assisted by an RNA chaperone. *Nucleic Acids Res.* **42:** 4682–4696

Bremer H, Dennis P & Ehrenberg M (2003) Free RNA polymerase and modeling global transcription in Escherichia coli. *Biochimie* **85:** 597–609

Brierley I, Pennell S & Gilbert RJC (2007) Viral RNA pseudoknots: Versatile motifs in gene expression and replication. *Nat. Rev. Microbiol.* **5:** 598–610

Brophy JAN & Voigt CA (2014) Principles of genetic circuit design. *Nat Meth* **11:** 508–520

Cambray G, Guimaraes JC, Mutalik VK, Lam C, Mai QA, Thimmaiah T, Carothers JM, Arkin AP & Endy D (2013) Measurement and modeling of intrinsic transcription terminators. *Nucleic Acids Res.* **41:** 5139–5148

Del Campo C, Bartholomäus A, Fedyunin I & Ignatova Z (2015) Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. *PLoS Genet.* **11:** 1–23

668 Canton B, Labno A & Endy D (2008) Refinement and standardization of synthetic biological parts
669 and devices. *Nat. Biotechnol.* **26:** 787–93
670 Cardinale S, Joachimiak MP & Arkin AP (2013) Effects of genetic variation on the E. coli host-circuit
671 interface. *Cell Rep.* **4:** 231–237
672 Castillo-Hair SM, Sexton JT, Landry BP, Olson EJ, Igoshin OA & Tabor JJ (2016) FlowCal: A User-
673 Friendly, Open Source Software Tool for Automatically Converting Flow Cytometry Data from
674 Arbitrary to Calibrated Units. *ACS Synth. Biol.* **5:** 774–780
675 Ceroni F, Algar R, Stan GB & Ellis T (2015) Quantifying cellular capacity identifies gene expression
676 designs with reduced burden. *Nat. Methods* **12:** 415–418
677 Ceroni F, Furini S, Gorochowski TE, Boo A, Borkowski O, Ladak YN, Awan AR, Gilbert C, Stan G-B
678 & Ellis T (2018) Burden-driven feedback control of gene expression. *Nat. Methods***:** 177030
679 Charneski C a & Hurst LD (2013) Positively charged residues are the major determinants of
680 ribosomal velocity. *PLoS Biol.* **11:** e1001508
681 Chen H, Shiroguchi K, Ge H & Xie XS (2015) Genome-wide study of mRNA degradation and
682 transcript elongation in Escherichia coli. *Mol. Syst. Biol.* **11:** 781–781
683 Condron BG, Atkins JF & Gesteland RF (1991a) Frameshifting in gene 10 of bacteriophage T7. *J.*
684 *Bacteriol.* **173:** 6998–7003
685 Condron BG, Gesteland RF & Atkins JF (1991b) An analysis of sequences stimulating frameshifting
686 in the decoding of gene 10 of bacteriophage T7. *Nucleic Acids Res.* **19:** 5607–5612
687 Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW,
688 Gaffney DJ, Elo LL, Zhang X & Mortazavi A (2016) A survey of best practices for RNA-seq data
689 analysis. *Genome Biol.* **17:** 1–19
690 Daniel R, Rubens JR, Sarpeshkar R & Lu TK (2013) Synthetic analog computation in living cells.
691 *Nature* **497:** 619–623
692 Davidsohn N, Beal J, Kiani S, Adler A, Yaman F, Li Y, Xie Z & Weiss R (2015) Accurate Predictions
693 of Genetic Circuit Behavior from Part Characterization and Modular Composition. *ACS Synth.*
694 *Biol.* **4:** 673–681
695 Der BS, Glassey E, Bartley BA, Enghuus C, Goodman DB, Gordon DB, Voigt CA & Gorochowski TE
696 (2017) DNAplotlib: Programmable Visualization of Genetic Designs and Associated Data. *ACS*
697 *Synth. Biol.* **6:**
698 Dong H, Nilsson L & Kurland CG (1996) Co-variation of tRNA abundance and codon usage in
699 Escherichia coli at different growth rates. *J. Mol. Biol.* **260:** 649–63
700 Endy D, You L, Yin J & Molineux IJ (2000) Computation, prediction, and experimental tests of fitness
701 for bacteriophage T7 mutants with permuted genomes. *Proc. Natl. Acad. Sci.* **97:** 5375–5380

Espah Borujeni A, Channarasappa AS & Salis HM (2013) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res.*: gkt1139

Fernandez-Rodriguez J, Yang L, Gorochowski TE, Gordon DB & Voigt CA (2015) Memory and Combinatorial Logic Based on DNA Inversions: Dynamics and Evolutionary Stability. *ACS Synth. Biol.* **4:**

Freistroffer D V., Kwiatkowski M, Buckingham RH & Ehrenberg M (2000) The accuracy of codon recognition by polypeptide release factors. *Proc. Natl. Acad. Sci.* **97:** 2046–2051

Giedroc DP & Cornish P V. (2009) Frameshifting RNA pseudoknots: Structure and mechanism. *Virus Res.* **139:** 193–208

Goodwin S, McPherson JD & McCombie WR (2016) Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17:** 333–351

Gorochowski TE, Avcilar-Kucukgoze I, Bovenberg RAL, Roubos JA & Ignatova Z (2016) A Minimal Model of Ribosome Allocation Dynamics Captures Trade-offs in Expression between Endogenous and Synthetic Genes. *ACS Synth. Biol.* **5:**

Gorochowski TE, Van Den Berg E, Kerkman R, Roubos JA & Bovenberg RAL (2014) Using synthetic biological parts and microbioreactors to explore the protein expression characteristics of escherichia coli. *ACS Synth. Biol.* **3:** 129–139

Gorochowski TE, Espah Borujeni A, Park Y, Nielsen AAK, Zhang J, Der BS, Gordon DB & Voigt CA (2017) Genetic circuit characterization and debugging using RNA-seq. *Mol. Syst. Biol.* **13:**

Gorochowski TE, Ignatova Z, Bovenberg RAL & Roubos JA (2015) Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic Acids Res.* **43:** 3022–3032

Guisbert E, Herman C, Lu CZ & Gross CA (2004) A chaperone network controls the heat shock response in E. coli. *Genes Dev.* **18:** 2812–2821

Guo H, Ingolia NT, Weissman JS & Bartel DP (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466:** 835–840

Guo MS & Gross CA (2014a) Stress-induced remodeling of the bacterial proteome. *Curr. Biol.* **24:** R424–R434

Guo MS & Gross CA (2014b) Stress-induced remodeling of the bacterial proteome. *Curr. Biol.* **24:** R424–R434

Gyorgy A, Jiménez JI, Yazbek J, Huang HH, Chung H, Weiss R & Del Vecchio D (2015) Isocost Lines Describe the Cellular Economy of Genetic Circuits. *Biophys. J.* **109:** 639–646

Hecht A, Glasgow J, Jaschke PR, Bawazer LA, Munson MS, Cochran JR, Endy D & Salit M (2017)

Measurements of translation initiation from all 64 codons in E. coli. *Nucleic Acids Res.* **45:** 3615–3626

Ingolia NT (2014) Ribosome profiling: New views of translation, from single codons to genome scale. *Nat. Rev. Genet.* **15:** 205–213

Ingolia NT, Ghaemmaghami S, Newman JRS & Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (80-.).* **324:** 218–23

Iost I & Dreyfus M (1995) The stability of Escherichia coli lacZ mRNA depends upon the simultaneity of its synthesis and translation. *EMBO J.* **14:** 3252–61

Ivanov IP, Anderson CB, Gesteland RF & Atkins JF (2004) Identification of a new antizyme mRNA +1 frameshifting stimulatory pseudoknot in a subset of diverse invertebrates and its apparent absence in intermediate species. *J. Mol. Biol.* **339:** 495–504

Jackson RJ, Hellen CUT & Pestova T V. (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.* **11:** 113–127

Jones DL, Brewster RC & Phillips R (2014) Promoter architecture dictates cell-to-cell variability in gene expression. *Science (80-.).* **346:** 1533–1536

Justman Q (2018) Splitting the World with Absolute Measurements: A Call for Collaborations in Physical Biology. *Cell Syst.* **6:** 395–396

Kennell D & Riezman H (1977) Transcription and translation initiation frequencies of the Escherichia coli lac operon. *J. Mol. Biol.* **114:** 1–21

Lareau LF, Hite DH, Hogan GJ & Brown PO (2014) Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *Elife* **2014:** 1–16

Li GW, Burkhardt D, Gross C & Weissman JS (2014) Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157:** 624–635

Lieberman-aiden E, Berkum NL Van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J & Mirny LA (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science (80-.).* **33292:** 289–294

Lucks JB, Mortimer SA, Trapnell C, Luo S, Aviran S, Schroth GP, Pachter L, Doudna JA & Arkin AP (2011) Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci.* **108:** 11063–11068

Makarova O V, Makarov EM, Sousa R & Dreyfus M (1995) Transcribing of Escherichia coli genes with mutant T7 RNA polymerases: stability of lacZ mRNA inversely correlates with polymerase speed. *Proc. Natl. Acad. Sci. U. S. A.* **92:** 12250–4

Margolin W (2012) The price of tags in protein localization studies. *J. Bacteriol.* **194:** 6369–6371

770 Mayer A, Di Iulio J, Maleri S, Eser U, Vierstra J, Reynolds A, Sandstrom R, Stamatoyannopoulos JA
771 & Churchman LS (2015) Native elongating transcript sequencing reveals human transcriptional
772 activity at nucleotide resolution. *Cell* **161:** 541–544
773 Mogk A, Huber D & Bukau B (2011) Integrating protein homeostasis strategies in prokaryotes. *Cold*
774 *Spring Harb. Perspect. Biol.* **3:** 1–19
775 Mohammad F, Woolstenhulme CJ, Green R & Buskirk AR (2016) Clarifying the Translational
776 Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Rep.* **14:** 686–694
777 Moon TS, Lou C, Tamsir A, Stanton BC & Voigt CA (2012) Genetic programs constructed from
778 layered logic gates in single cells. *Nature* **491:** 249–253
779 Mortazavi A, Williams BA, McCue K, Schaeffer L & Wold B (2008) Mapping and quantifying
780 mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5:** 621–628
781 Mutalik VK, Guimaraes JC, Cambray G, Lam C, Christoffersen MJ, Mai QA, Tran AB, Paull M,
782 Keasling JD, Arkin AP & Endy D (2013) Precise and reliable gene expression via standard
783 transcription and translation initiation elements. *Nat. Methods* **10:** 354–360
784 Myers CJ, Beal J, Gorochowski TE, Kuwahara H, Madsen C, McLaughlin JA, Misirli G, Nguyen T,
785 Oberortner E, Samineni M, Wipat A, Zhang M & Zundel Z (2017) A standard-enabled workflow
786 for synthetic biology. *Biochem. Soc. Trans.* **45:**
787 Nielsen AA & Voigt CA (2014) Multi-input CRISPR/Cas genetic circuits that interface host regulatory
788 networks. *Mol. Syst. Biol.* **10:** 763–763
789 Nielsen AAK, Der BS, Shin J, Vaidyanathan P, Paralanov V, Strychalski EA, Ross D, Densmore D &
790 Voigt CA (2016) Genetic circuit design automation. *Science (80-.).* **352:**
791 Owens NDL, Blitz IL, Lane MA, Patrushev I, Overton JD, Gilchrist MJ, Cho KWY & Khokha MK
792 (2016) Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals
793 Transcriptome Kinetics in Development. *Cell Rep.* **14:** 632–647
794 Poole ES, Brown CM & Tate WP (2000) The identity of the base following the stop codon
795 determines the efficiency of in vivo translational termination in Escherichia cobl. *EMBO J.* **14:**
796 151–158
797 Raser JM & O'Shea EK (2005) Noise in Gene Expression: Origins, Consequences, and Control.
798 *Science (80-.).* **309:** 2010–2014
799 Salis HM, Mirsky E a & Voigt C a (2009) Automated design of synthetic ribosome binding sites to
800 control protein expression. *Nat. Biotechnol.* **27:** 946–50
801 Seo SW, Yang JS, Kim I, Yang J, Min BE, Kim S & Jung GY (2013) Predictive design of mRNA
802 translation initiation region to control prokaryotic translation efficiency. *Metab. Eng.* **15:** 67–74
803 Sharma V, Prère MF, Canal I, Firth AE, Atkins JF, Baranov P V. & Fayet O (2014) Analysis of tetra-

804 and hepta-nucleotides motifs promoting-1 ribosomal frameshifting in Escherichia coli. *Nucleic*
805 *Acids Res.* **42:** 7210–7225
806 Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA & Waterston RH
807 (2017) DNA sequencing at 40: Past, present and future. *Nature* **550:** 345–353
808 Siuti P, Yazbek J & Lu TK (2013) Synthetic circuits integrating logic and memory in living cells. *Nat.*
809 *Biotechnol.* **31:** 448–452
810 Smanski MJ, Zhou H, Claesen J, Shen B, Fischbach MA & Voigt CA (2016) Synthetic biology to
811 access and expand nature's chemical diversity. *Nat. Rev. Microbiol.* **14:** 135–149
812 Snapp E (2005) Design and Use of Fluorescent Fusion Proteins in Cell Biology. *Curr. Protoc. Cell*
813 *Biol.*
814 Tan SZ, Manchester S & Prather KLJ (2016) Controlling Central Carbon Metabolism for Improved
815 Pathway Yields in Saccharomyces cerevisiae. *ACS Synth. Biol.* **5:** 116–124
816 Taniguchi Y, Choi PJ, Li G-W, Chen H, Babu M, Hearn J, Emili A & Xie XS (2010) Quantifying E.
817 coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science (80-.*
818 *).* **329:** 533 LP-538
819 Tholstrup J, Oddershede LB & Sørensen MA (2012) MRNA pseudoknot structures can act as
820 ribosomal roadblocks. *Nucleic Acids Res.* **40:** 303–313
821 Tsuchihashi Z & Kornberg A (1990) Translational frameshifting generates the gamma subunit of
822 DNA polymerase III holoenzyme. *Proc. Natl. Acad. Sci.* **87:** 2516–2520
823 Unoson C & Wagner EGH (2007) Dealing with stable structures at ribosome binding sites: Bacterial
824 translation and ribosome standby. *RNA Biol.* **4:** 113–117
825 Wang L-Z, Wu F, Flores K, Lai Y-C & Wang X (2016) Build to understand: synthetic approaches to
826 biology. *Integr. Biol.* **8:** 394–408
827 Wohlgemuth SE, Gorochowski TE & Roubos JA (2013) Translational sensitivity of the Escherichia
828 coli genome to fluctuating tRNA availability. *Nucleic Acids Res.* **41:** 8021–8033
829 Yang L, Nielsen AAK, Fernandez-Rodriguez J, McClune CJ, Laub MT, Lu TK & Voigt CA (2014)
830 Permanent genetic memory with >1-byte capacity. *Nat. Methods* **11:** 1261–1266
831 Yarchuk O, Jacques N, Guillerez J & Dreyfus M (1992) Interdependence of translation, transcription
832 and mRNA degradation in the lacZ gene. *J. Mol. Biol.* **226:** 581–596
833

Figures and Captions

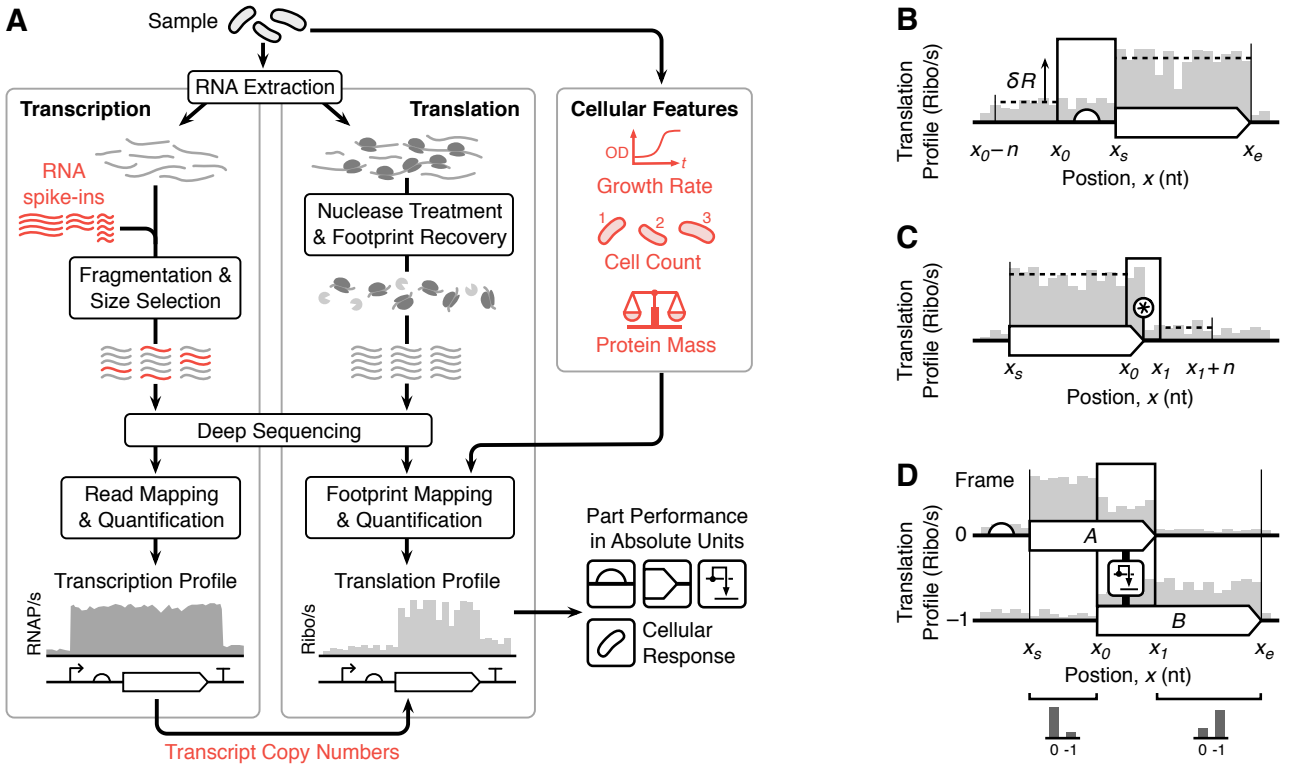

Figure 1: Overview of the workflow. (**A**) Major steps involved when quantifying transcription (RNA-seq) and translation (Ribo-seq) and the additional cellular features measured. Elements required for quantification in absolute units are highlighted in red. (**B**) Model for calculating the translation initiation rate of a ribosome binding site (Eq. 2). (**C**) Model for calculating the termination efficiency of a stop codon (Eq. 3). Star denotes the location of the stop codon. (**D**) Model for calculating translational frameshifting efficiency between two coding regions 'A' and 'B' in zero and –1 reading frames, respectively (Eq. 4).

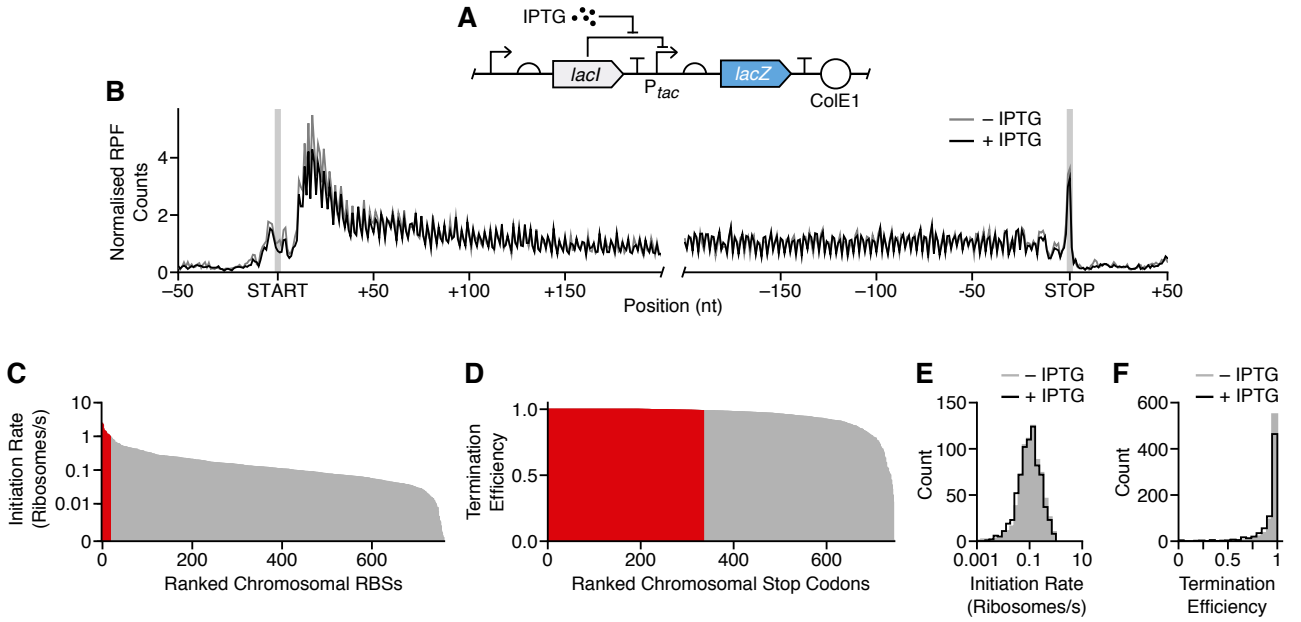

Figure 2: Measuring translation initiation and termination signals across the *E. coli* transcriptome. (**A**) Genetic design of the LacZ reporter construct whose expression is activated by the inducer IPTG. (**B**) Normalized RPF count profile averaged for all *E. coli* transcripts. Profiles generated for cells grown in the absence and presence of IPTG (1 mM). Start and stop codons are shaded. (**C**) Bar chart of all measured RBS initiation rates ranked by their strength. Strong RBSs with initiation rates >1 ribosome/s are highlighted in red. (**D**) Bar chart of all measured stop codon termination efficiencies ranked by their strength. Stop codons with termination efficiency >0.99 are highlighted in red. (**E**) Distribution of initiation rates for cells grown in the absence and presence of IPTG (1 mM). (**F**) Distribution of stop codon termination efficiencies for cells grown in the absence and presence of IPTG (1 mM).

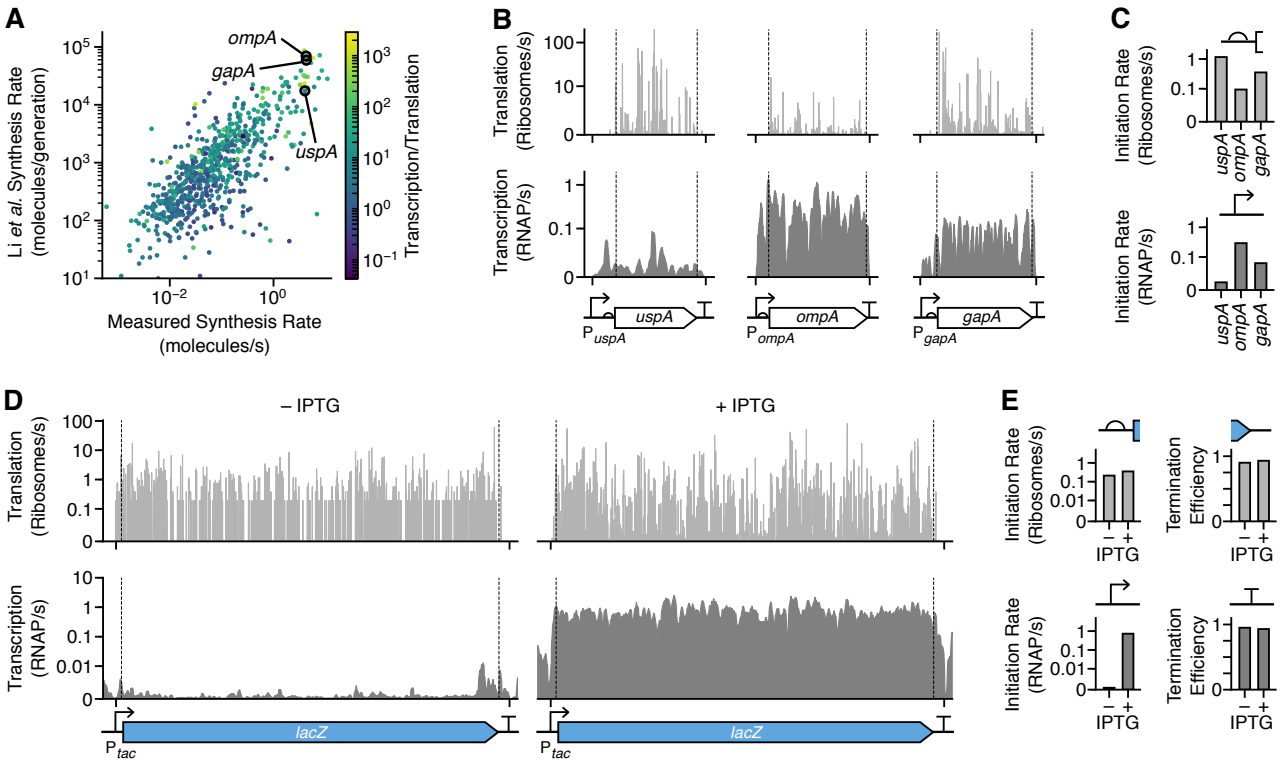

Figure 3: Simultaneous quantification of transcription and translation of endogenous genes and a synthetic genetic construct. (**A**) Comparison of protein synthesis rate of endogenous *E. coli* genes measured using Ribo-seq from this study (in molecules/s units) and from that by Li *et al.* (Li *et al*, 2014) (in molecules/generation units). Each point corresponds to a single gene and color denotes the ratio of transcription and translation rate capturing whether transcription (light yellow) or translation (dark blue) is the dominant factor. (**B**) Transcription (bottom) and translation (top) profiles for *yggN*, *rpoH* and *greA*, computed from the RNA-seq and Ribo-seq data without induction. Positions of the genetic parts and gene are shown below the profiles. (**C**) Promoter strengths in RNAP/s units and RBS initiation rates in ribosome/s units. (**D**) Transcription (bottom) and translation (top) profiles for *lacZ*. Profiles are shown for cells in the absence and presence of IPTG (1 mM). Position of genetic parts and gene is shown below the profiles. RBS is omitted from the genetic design due to its size. (**E**) Measured promoter strength in RNAP/s units, RBS initiation rate in ribosomes/s units, and the transcriptional terminator and stop codon termination efficiency for *lacZ*. Data shown for cells in the absence and presence of IPTG (1 mM).

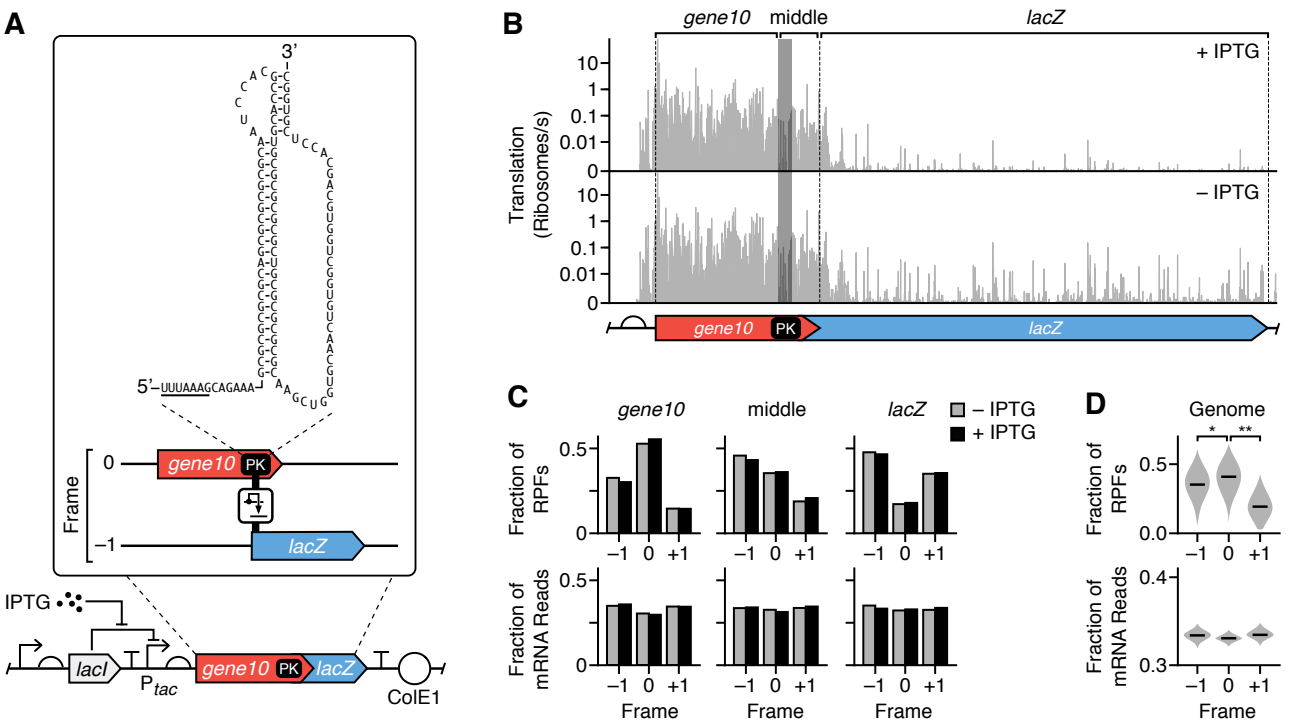

Figure 4: Characterization of a synthetic pseudoknot construct that induces translational frameshifting. (**A**) Genetic design of the PK-LacZ construct. Expanded sequence shows the PK secondary structure with the slippery site underlined, as well as the two genes (*gene10* and *lacZ*) in differing reading frames. (**B**) Translation profiles for the PK-LacZ construct in cells cultured in the absence (bottom) and presence (top) of IPTG (1 mM). The *gene10*, middle, and *lacZ* regions are labelled above the profiles. Shaded region denotes the PK, and dashed lines denote the start codon and stop codons of *gene10* and *LacZ*. (**C**) Fraction of the total RPFs and mRNA reads in each reading frame for the *gene10*, PK or middle, and *lacZ* regions schematically shown below and are of the PK-LacZ construct. Data shown separately for cells cultured in the absence and presence of IPTG (1 mM). (**D**) Violin plots of the distributions of fractions of total RPFs and mRNA reads in each reading frame for all *E. coli* transcripts. Median values shown by horizontal bars. *, $P = 0.049$; **, $P = 1.6 \times 10^{-9}$.

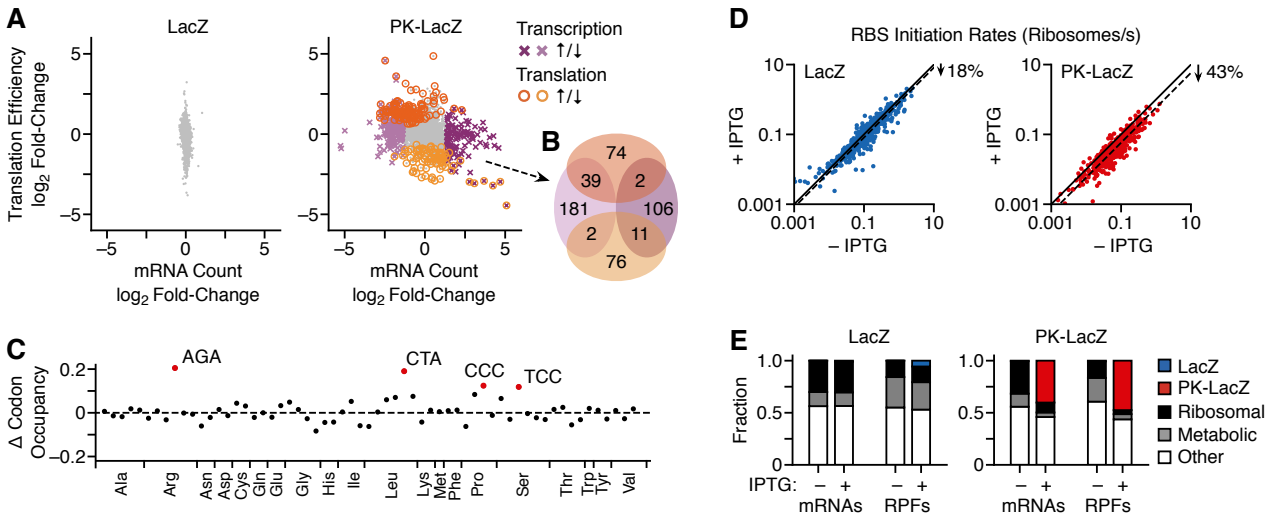

Figure 5: Cellular response to the expression of a synthetic pseudoknot construct. (**A**) Change in expression of chromosomal genes in *E. coli* cells following induction of *PK-lacZ* expression (1 mM IPTG). Each point represents a transcript. Differentially expressed genes (mRNA count: $P < 0.001$ and absolute $\log_2$ fold-change > 1.37; translation efficiency: $P < 0.01$) are highlighted in color and by an alternative point shape (transcriptional regulation: purple cross; translational regulation: orange open circle). (**B**) Venn diagram of genes significantly regulated transcriptionally and translationally after induction of the PK-LacZ construct. Colors match those in panel A. (**C**) Change in codon occupancy for cells harboring the PK-LacZ construct after induction by IPTG (1 mM) calculated from the Ribo-seq data. Each point corresponds to a codon, which are ordered by amino acid identity and then by abundance in the genome (left most abundant, right least abundant). Dashed horizontal line denotes no change. Outliers are labelled and highlighted in red (Tukey test: 1.5 times the interquartile range below the first quartile or above the third quartile). (**D**) Translation initiation rates for all *E. coli* RBSs in cells harboring the LacZ and PK-LacZ constructs in the absence and presence of IPTG (1 mM). Solid line shows the same initiation rate for both conditions. Dotted lines denote linear regressions for the data with no offset. (**E**) Fractions of mRNA reads and RPFs mapping to each synthetic expression construct (LacZ and PK-LacZ) and *E. coli* transcripts, which are divided into three major categories: ribosomal, metabolic, and other functions. Data shown for cells cultured in the absence and presence of IPTG (1 mM).