ORIGINAL ARTICLE
Journal Section

Computational design and interpretation of single-RNA translation experiments

Luis U. Aguilera[1] | William Raymond[1,2] | Zachary R. Fox[2] | Michael May[2] | Elliot Djokic[1,2] | Tatsuya Morisaki[3] | Timothy J. Stasevich[3,4] | Brian Munsky[1,2]

[1]Department of Chemical and Biological Engineering, Colorado State University Fort Collins, CO 80523, USA

[2]School of Biomedical Engineering, Colorado State University Fort Collins, CO 80523, USA

[3]Department of Biochemistry and Molecular Biology and Institute for Genome Architecture and Function, Colorado State University, Fort Collins, CO, USA

[4]Cell Biology Unit, Institute of Innovative Research, Tokyo Institute of Technology, Nagatsuta-cho 4259, Midori-ku, Yokohama, Kanagawa, 226-8503, Japan

Correspondence
Brian Munsky, Department of Chemical and Biological Engineering, Colorado State University Fort Collins, CO 80523, USA
Email: Brian.Munsky@colostate.edu

Funding information
Research was supported by the National Institute of General Medical Sciences of the National Institutes of Health, the WM Keck Foundation and the National Science Foundation.

Advances in fluorescence microscopy have introduced new assays to quantify live-cell translation dynamics at single-RNA resolution. We introduce a detailed, yet efficient sequence-based stochastic model that generates realistic synthetic data for several such assays, including fluorescence correlation spectroscopy (FCS), ribosome runoff assays (ROA) after Harringtonine application, and fluorescence recovery after photo-bleaching (FRAP). We simulate these experiments under multiple imaging conditions and for thousands of human genes, and we evaluate through simulations which experiments are most likely to provide accurate estimates of elongation kinetics. Finding that FCS analyses are optimal for short or long length genes, we integrate our model with experimental FCS data to capture the nascent protein statistics and temporal dynamics for three human genes: KDM5B, β-actin, and H2B. Finally, we introduce a new open-source software package, RNA Sequence to NAscent Protein Simulator (rSNAPsim) to easily simulate the single-molecule dynamics of any gene sequence for any of these assays and for different assumptions regarding synonymous codon us-

Abbreviations: aa, amino acid; CAI, Codon Adaptation Index; FCS, Fluorescence Correlation Spectroscopy; FPS, Frames Per Second; FRAP, Fluorescence Recovery After Photobleaching; MCP, MS2 Coat Protein; POI, Protein of Interest; ROA, Run Off Assays; RMSE, Root Mean Squared Error; rSNAPsim, RNA Sequence to NAscent Protein Simulator; SEM, Standard Error of the Mean; ump, Units of Mature Protein.

age, tRNA level modifications, or ribosome pauses. RSNAPSIM is implemented in Python and is available at: `https://github.com/MunskyGroup/rSNAPsim.git`.

KEYWORDS
Single-molecule, translation, live-cell, stochastic, modeling, FRAP, FCS

1 | INTRODUCTION

The central dogma of molecular biology (i.e., DNA codes are transcribed into messenger RNA, which are then translated to build proteins) has been a foundation of biological understanding since it was stated by Francis Crick in 1958. Yet, despite their overwhelming importance to biological and biomedical science, many of the fundamental steps in the gene expression process are only now becoming observable in living cells through the application of real time single-molecule fluorescence imaging approaches. Imaging single-molecule transcription was first achieved two decades ago through the use of the MS2 system [1], which uses bacteriophage gene sequences to encode for specific and repeated stem-loop secondary structures in the transcribed mRNAs. These stem-loops are subsequently recognized and bound by multiple fluorescently-tagged MS2 Coat Proteins (MCP), which produce bright fluorescent spots that allow for the detection and spatial tracking of single-mRNA [2]. Tracking these labeled RNA has made it possible to observe many aspects of RNA dynamics that were previously obscured using bulk RNA measurements, such as the production of RNA from different alleles [3], the movement of mRNA–protein complexes from nucleus to cytoplasm through nuclear pores [4], and the association of RNA with different regions of the cell [5].

Even more recently, imaging single-molecule translation has also become possible through the discovery of similar approaches [6, 7, 8, 9, 10]. In this case, the mRNA is modified to encode for multiple epitopes in the open reading frame of a protein of interest (POI). As the protein is translated, these epitopes are quickly recognized and bound by fluorescent antibody fragment probes, Fig 1A. By combining the MS2 approach with these epitope recognition sites, the co-localization of mRNA spots and nascent protein spots reveal single-mRNA molecules that are undergoing active translation, as shown in Fig 1B. As was the case for single-RNA tracking, precise spatiotemporal imaging of translation sites within single living cells allows for multiple advances in comparison to bulk or single-cell assays [11]. For example, Morisaki and Stasevich, [12] recently reviewed three different approaches to track the translation dynamics for individual mRNA molecules over time and then use these data to infer translation rates. The first design is related to Fluorescence Correlation Spectroscopy (FCS), in that the nascent protein fluorescence signal is monitored over time and used to compute the autocorrelation function (G(τ), Fig 1C, bottom). The time, τ_{FCS}, at which G(τ) reaches zero denotes the characteristic time for a ribosome to translate the gene from the tag region to the end of the protein of interest [6]. A second approach to measure translation rate is to chemically block translation initiation (e.g., through application of a drug such as Harringtonine, as depicted in Fig 1D, top). In this run off assay (ROA) approach, the time, τ_{ROA}, at which the fluorescence signal disappears corresponds to the time for a single ribosome to translate the entire coding region, including the tag region itself [10]. A third technique, shown in Fig 1E, uses Fluorescence Recovery After Photobleaching (FRAP) to optically eliminate the nascent protein fluorescence associated with a single mRNA and then record the recovery of the signal to its original level. As for the FCS approach, the time of total recovery, τ_{FRAP}, relates to the time required for a single ribosome to complete translation from the tag region to the termination codon [9, 8].

The temporal resolution offered by live single-mRNA, nascent translation imaging makes it possible to directly

visualize and quantify initiation, elongation, and termination processes in live-cells [13]. Single-molecule studies have uncovered previously unknown events and mechanisms taking place during translation, such as the presence of active and inactive transcripts in specific locations in the cell [9, 6], different elongation rates caused by codon optimized sequences [10], the spatiotemporal translation of specific genes in specific cellular compartments [7, 8], ribosomal frameshifting with bursty dynamics [14], and non-canonical forms of translation [15].

As these experimental techniques rapidly evolve, they induce a growing need for precise and flexible computational tools to interpret the resulting data and to design the next wave of single-RNA translation experiments. To help fill this gap, we present a versatile new set of theoretical and computational tools to complement the experimental study of translation at single-molecule resolution. We demonstrate the generality of our analyses by simulating results for several different single-molecule experiments for a large database of human genes. We explore these different combinations of gene and experiment to ask which methodologies are better to measure specific biophysical parameters and for which types of genes. We then constrain our model by fitting it to experimental data for several genes. Finally, we describe and demonstrate the use of a new open-source and user-friendly software package: RNA Sequence to NAscent Protein Simulation (RSNAPSIM), which allows the user to easily simulate the single-molecule dynamics of any gene. Finally, we discuss future directions and the potential limitations of the current form of this new technology.

2 | RESULTS

2.1 | Modeling Single-RNA Translation Dynamics

To simulate translation with single-molecule resolution, we developed a stochastic model consisting of a set of reactions where random variables $\{x_i\}$ represent the fluctuating occupancy of the ribosomes at a specific i^{th} codon along a single mRNA,

$$\varnothing \xrightarrow{w_0(x_1,\ldots,x_{nf})} x_1 \xrightarrow{w_i(x_2,\ldots,x_{nf+i})} x_i \xrightarrow{w_{i+1}(x_{i+1},\ldots,x_{nf+i+1})} \ldots \xrightarrow{w_t} x_L, \tag{1}$$

where L is the length of the gene in codons, n_f is the ribosome footprint, and $\mathbf{x} = [x_1, x_2, \ldots, x_L] \in \mathbb{B}^L$ is a binary vector of zeros and ones, known as the occupancy vector, which represents the presence ($x_i = 1$) or absence ($x_i = 0$) of ribosomes at every i^{th} codon. The initial reaction in the model describes the initiation step, where the ribosomes bind to the mRNA at the rate $w_0(x_1, \ldots, x_{nf})$. Ribosomes are large biomolecules that occupy around 20 to 30 nuclear bases (or seven to 10 codons) once bound to the mRNA [16]. This is captured in the model by specifying the ribosome footprint, $n_f = 9$, which guarantees that initiation cannot occur if another downstream ribosome is already present within the first n_f codons, Fig 2A. This binding restriction can be written simply as:

$$w_0 = k_i \prod_{j=1}^{nf} (1 - x_j), \tag{2}$$

where k_i is the initiation constant, and the product is equal to one if and only if there are no ribosomes within the first n_f codons.

Similarly, one can represent the elongation reactions, where the ribosome moves along the mRNA from codon to

codon in direction 5' to 3' according to:

$$w_i = k_e(i) \cdot x_i \prod_{j=1}^{nf} (1 - x_{i+j}), \quad \text{for } i = 1, ..., L-1; \tag{3}$$

where $k_e(i)$ is the elongation rate, and the product again enforces ribosome exclusion. To implement the effect of codon-usage bias and tRNA availability during protein synthesis, we adopt a similar argument to that presented by Georgoni et al., [17]: rare codons are correlated with low tRNA abundance, which cause a longer waiting time for the ribosome to synthesize the given amino acid at that codon. As tRNA concentrations have been related to codon usage [18], we assume each codon's elongation rate is proportional to its usage in the human genome according to:

$$k_e(i) = \bar{k}_e \cdot (u(i)/\bar{u}), \tag{4}$$

where $u(i)$ denotes the codon usage frequency in the human genome (given in Appendix Table S1 from [19]), $\bar{u}$ represents the average codon usage frequency in the human genome, and the global parameter $\bar{k}_e$ is an average elongation constant, which can be determined through experiments.

Although simple in its specification, the above model allows for many adjustments to explore different experimental circumstances. As a few examples, one could represent translation inhibition analyses such as those performed in [7] by making the initiation rate, k_i, a function of time or external input; one can analyze effects of synonymous codon substitution by replacing codons with their more or less common relatives; one may represent codon depletion, as studied in [17] by reducing the corresponding rates $k_e(i)$ for all i corresponding to the depleted tRNA; or one could explore the effects of pausing or traffic jams at specific codons by reducing $k_e(i)$ at specific codons. We will explore several such circumstances below.

With our simple specification of the translation initiation, elongation and termination reactions, we can now simulate random trajectories, $\mathbf{x}(t)$, which we collect to form binary occupancy trajectory matrices $\mathbf{X} = \left[\mathbf{x}(t_1)^T, \ldots, \mathbf{x}(t_{N_t})^T\right]^T \in \mathbb{B}^{N_t \times L}$, where each row refers to the ith position on the gene, and each column represents a specific time t_j. To visualize ribosome movement trajectories, each random $\mathbf{X}$ can be plotted in two dimensions (position v.s. time) to form kymographs similar to those extensively used to represent organelle movement [20]. For example, Fig 2B shows a visualization of $\mathbf{X}$ for a case study on the $\beta-$actin gene. Each line from left to right on the kymograph corresponds to the movement of a single ribosome from initiation to termination. We note that averaging along the columns of $\mathbf{X}$ (i.e., in the vertical direction of the kymograph) yields the time-averaged loading of the ribosomes at each codon position, and summing across the rows of $\mathbf{X}$ (i.e., in the horizontal direction of the kymograph) yields the number of ribosomes for that mRNA at each instant in time.

To relate our model describing ribosome occupancy to experimental measurements of translation spot fluorescence, we introduce a *fluorescence intensity vector* that converts the instantaneous occupancy vector, $\mathbf{x}(t)$, to the total number of translated epitopes available to bind to fluorescent markers. This intensity vector can be written as:

$$I(t) = \sum_{i=1}^{L} c_i \cdot x_i(t) = \mathbf{c}\mathbf{x}^{\mathsf{T}}, \tag{5}$$

where $\mathbf{c} = [c_1, c_2, \ldots, c_L]$ and each c_i is the cumulative number of fluorescent probes bound to epitopes encoded at positions $(1, \ldots, i)$ along the mRNA. For example, $\mathbf{c}$ would be defined as $\mathbf{c} = [0, 0, 1, 1, 2, 2, 3, 3, ..., 3]$ for an RNA sequence with epitopes encoded at positions [3, 5, 7]. We note that the random occupancy matrices, $\mathbf{X}$, are easily converted to intensity time traces using the simple algebraic operation $\mathbf{I} = \left[I(t_1), \ldots, I(t_{N_t})\right] = \mathbf{c}\mathbf{X}^{\mathsf{T}}$.

2.1.1 | Approximate Model for Statistical Moments of Nascent Transcript Kinetics

When ribosome loading is sparse (e.g., for slow initiation or fast elongation such that $(k_i/ke \ll 1/nf)$), ribosome collisions will become negligible, and the nonlinearities in Eqs. 2-3 have less effect on the overall ribosome dynamics. Under such circumstances, it is possible to derive a simplified linear system model for the elongation dynamics. In the linear model, the propensity of the codon-dependent elongation step (Eq. 2) is simplified to $w_i(x_i) = k_i x_i$ such that the ability of a ribosome to add another amino acid only depends on the current position of the ribosome, and not on the footprint of other ribosomes.

We define the reaction stoichiometry matrix to describe the change in the ribosome loading vector, $\mathbf{x}$, for every reaction as:

$$\mathbf{S}_{i,j} = \begin{cases} 1 \text{ for all } i = j, \\ -1 \text{ for all } i = j - 1, \end{cases} \tag{6}$$

where i corresponds to each codon in the protein of interest. The first column of $\mathbf{S}$ corresponds to the initiation reaction, the next $L - 1$ columns refer to elongation steps when an individual ribosome transitions from the ith to the $i + 1^{\text{th}}$ codon, and the final column corresponds to the final elongation step and termination. Maintaining the same order of reactions, and neglecting ribosome exclusion, the propensities of all reactions can be written in the affine linear form as:

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{W}_1\mathbf{x}, \tag{7}$$

where $\mathbf{w}_0$ is a column vector of zeros with the first entry k_i and $\mathbf{W}_1$ is a matrix defined as:

$$[\mathbf{W}_1]_{i,j} = \begin{cases} k_e(i) \text{ for all } i = j + 1, \\ 0 \text{ otherwise .} \end{cases} \tag{8}$$

Using the definition of the fluorescence intensity from Eq. 5, the first two uncentered moments of the intensity $I(t)$ can be written in terms of the ribosome position vector $\mathbf{x}$ as:

$$\mathbb{E}\{I(t)\} = \mathbb{E}\{\mathbf{c}\mathbf{x}\} = \mathbf{c}\mathbb{E}\{\mathbf{x}\}, \tag{9}$$

$$\Sigma_I(0) = \mathbb{E}\{(I(t) - \mathbb{E}\{I(t)\})^2\} = \mathbf{c}\Sigma_{\mathbf{x}}(0)\mathbf{c}^{\mathsf{T}}, \tag{10}$$

where $\mathbb{E}\{\mathbf{x}\}$ and $\Sigma_{\mathbf{x}}(0)$ are the mean and zero-lag-time variance in the ribosome occupancy vector, respectively. For the approximate linear propensity functions in Eq. 7, the moments of the ribosome position vector are governed by the equations [21]:

$$\frac{d\mathbb{E}\{\mathbf{x}\}}{dt} = \mathbf{S}\mathbf{W}_1\mathbb{E}\{\mathbf{x}\} + \mathbf{S}\mathbf{w}_0 \tag{11}$$

$$\frac{d\Sigma_{\mathbf{x}}}{dt} = \mathbf{S}\mathbf{W}_1\Sigma_{\mathbf{x}} + \Sigma_{\mathbf{x}}\mathbf{W}_1^T\mathbf{S}^T + \mathbf{S}\text{diag}(\mathbf{W}_1\mathbb{E}\{\mathbf{x}\} + \mathbf{w}_0)\mathbf{S}^T. \tag{12}$$

By setting the left hand side of Eq. 11 to zero, the steady state mean ribosome loading vector can be found by solving

the algebraic expression:

$$\mathbf{S}\mathbf{W}_1\mathbb{E}\{\mathbf{x}\} + \mathbf{S}\mathbf{w}_0 = 0. \tag{13}$$

Similarly, the steady-state covariance matrix, $\Sigma_{\mathbf{x}}$, in the ribosome loading vector is given by the solution to the Lyapunov equation (from right hand side of Eq. 12):

$$\mathbf{S}\mathbf{W}_1\Sigma_{\mathbf{x}} + \Sigma_{\mathbf{x}}\mathbf{W}_1^T\mathbf{S}^T + \mathbf{S}\mathrm{diag}(\mathbf{W}_1\mathbb{E}\{\mathbf{x}\} + \mathbf{w}_0)\mathbf{S}^T = 0. \tag{14}$$

The autocorrelation dynamics of the nascent protein fluorescence intensity is defined:

$$G(\tau) = \mathbb{E}\{\mathbf{c}(\mathbf{x}(t) - \mathbb{E}\{\mathbf{x}(t)\})(\mathbf{x}(t) - \mathbb{E}\{\mathbf{x}(t)\})^T\mathbf{c}^T\} = \mathbf{c}\mathbb{E}\{(\mathbf{x}(t) - \mathbb{E}\{\mathbf{x}(t)\})(\mathbf{x}(t) - \mathbb{E}\{\mathbf{x}(t)\})^T\}\mathbf{c}^{\mathsf{T}} \tag{15}$$

$$= \mathbf{c}\Sigma_{\mathbf{x}}(\tau)\mathbf{c}^{\mathsf{T}}, \tag{16}$$

where $\Sigma_{\mathbf{x}}(\tau)$ is the cross-correlation of the ribosome occupancies at a lag time of length τ. Noting that the probe design, $\mathbf{c}$, is constant with respect to τ, it is only necessary to find the cross-correlations of the ribosome occupancy. Following the regression theorem [22], these correlations are given by the solution to the set of ODEs,

$$\frac{d\Sigma_{\mathbf{x}}(\tau)}{d\tau} = \phi\Sigma_{\mathbf{x}}(\tau), \tag{17}$$

where the initial condition is provided by steady-state covariance of the process (i.e., the solution for $\Sigma_{\mathbf{x}}(0)$ in Eq. 14) and the autonomous matrix of the process is given by $\phi = \mathbf{S}\mathbf{W}_1$. Integrating Eq. 17, the autocorrelation of the intensity $G(\tau)$ can be found using Eq. 16.

2.1.2 | Simplified Theoretical Model for Nascent Translation Kinetics

In the limit of low initiation events and long genes, the probe region can be approximated by a single point, and the above model can be simplified even further to allow direct estimation of steady state translation features. First, since the average time for a ribosome to move one codon is $\Delta t_i = 1/k_e(i)$, the total average time it takes a ribosome to complete translation from the start codon to the end of the mRNA is:

$$\tau = \sum_{i=1}^{L} \frac{1}{k_e(i)}, \tag{18}$$

where L is the gene length. Using the codon-dependent translation rates from Eq. 4, we can modify Eq. 18 to

$$\tau = \frac{1}{\bar{k}_e}\sum_{i=1}^{L}\frac{\bar{u}}{u(i)}. \tag{19}$$

If one could experimentally measure τ_{Exp} using one of the techniques described above, then $\bar{k}_e$ could be estimated as:

$$\bar{k}_e \approx \frac{1}{\tau_{\mathrm{Exp}}}\sum_{i=n_p}^{L}\frac{\bar{u}}{u(i)}, \tag{20}$$

where n_p is the effective codon position of the fluorescent tag. In practice, the specification of n_p will vary depending upon the type of experiment (e.g., FCS, FRAM or ROA) used to estimate τ_{Exp}, as will be discussed in more detail below.

Given the apparent association time of a ribosome on the mRNA (τ) and the initiation rate (k_i), the distribution for the number of visible ribosomes on a transcript at steady state can also be estimated using this simplified model. Under the assumption that each initiation event is an independent and exponentially distributed random event, the number of ribosomes downstream from the n_b^{th} codon, and therefore the fluorescence in units of mature proteins, would be approximated by a Poisson distribution with mean (and variance) equal to

$$\mu \approx \sigma^2 \approx k_i \cdot \tau. \tag{21}$$

For a more realistic treatment of the fluorescence intensity, one could assume that the multiple probes are spread uniformly over a finite region, such that the fluorescence will increase linearly as ribosomes pass through the probe region. To approximate this gradual increase in fluorescence, Eq. 21 can be corrected by a multiplicative factor (see Methods) as:

$$\mu_I \approx k_i \cdot \tau \left(1 - \frac{L_{\text{t}}}{2L}\right), \tag{22}$$

$$\sigma_I^2 \approx k_i \cdot \tau \left(1 - 2\frac{L_{\text{t}}}{3L}\right), \tag{23}$$

where L_{t} is the length of the tag region (e.g., $L_{\text{t}} = 318$ aa for the 10X FLAG 'Spaghetti Monster' SM-tag used in [6]).

To demonstrate the close agreement between the full stochastic model, the reduced linear moments model, and the simplified theoretical analysis, Table 1 compares the model generated values for each of the quantities τ, μ_I, and σ_I^2 for three different human genes H2B ($L = 128$aa), β-actin ($L = 375$aa), and KDM5B ($L = 1549$aa), using reported parameters of $k_i = 0.03\ \text{s}^{-1}$ and $\bar{k}_e = 10\ \text{s}^{-1}$ [6]. For further comparison, Fig 2C compares estimates of τ (top), μ_I (middle), and $\sigma_I^2(bottom)$ for the β-actin gene for each of the three analyses, and as a function of different initiation and elongation rates. This comparison demonstrates that, at least for fast elongation rates, the full stochastic analysis and the moments-based computation are in excellent agreement to estimate the effective time as well as the mean and variance in the level of nascent proteins per RNA. However, when the initiation rate approaches $\bar{k}_e/nf$, ribosome collisions become more prevalent, which substantially lengthens the effective elongation time (Fig 2C top), and leads to a saturation of ribosomes (Fig 2C middle and bottom), and these nonlinear behaviors are not captured by the moment-based model. For longer genes, the simplified theoretical estimates from Eqs. 18-21 are also in good agreement with the complete model. For shorter genes, it becomes less realistic to approximate the tag region with a single point or a continuous ramp, and the error of this approximation leads to poorer estimates of the elongation time and the Poisson approximation over-estimates the true variance (see H2B in Table 1). However, even for short genes, the linear moments-based model, which includes the exact positions of all probes and the codon usage, provides a more accurate estimate of the true system behaviors.

Having demonstrated close agreement of the simplified theoretical models with the full stochastic simulations, we next explore how well different experiment designs should be expected to work to estimate translation parameters from single-RNA translation dynamics.

2.2 | Evaluation of Experimental Assays to Quantify Translation Kinetics

Using the models above, and if we knew the average time that ribosomes take to translate a single complete protein from a given gene, $\tau^{(g)}$, we could estimate $\bar{k}_e^{(g)}$ using Eq. 20.

With this in mind, we next consider three experimental approaches that have been used to estimate $\tau^{(g)}$ in recent experimental investigations (Fig 1C-E): fluorescence correlation spectroscopy (FCS), run-off assays (ROA), and fluorescence recovery after photobleaching (FRAP). Using our full stochastic models to generate synthetic data and the simplified theoretical model to interpret this data, we ask how accurately would each of these three assays work to identify $\bar{k}_e^{(g)}$ for a comprehensive list of 2,647 human genes from the PANTHER database [23] and under different imaging conditions corresponding to different frame rates or numbers of mRNA spots.

In the FCS approach, we compute the autocorrelation function, $G(\tau)$, of the simulated fluorescence intensities, and from $G(\tau)$ we estimate the time lag, τ_{FCS}, at which correlations disappear (see Fig 1C and Methods). In the ROA approach, we simulate the addition of a chemical compound, such as Harringtonine, which binds the 60S ribosome subunit and prevents ribosome assembly [24], and we record the average time, τ_{ROA}, at which protein fluorescence disappears from the RNA (see Fig 1D and Methods). To approximate variability in the specific time at which the drug reaches the mRNA and blocks ribosome initiation, we assume that the time of initiation blockage occurs at a normally distributed time of 60 ± 10 seconds [25]. In the FRAP analysis, we simulate an instantaneous fluorescence bleaching of all nascent proteins and then record the average time, τ_{FRAP}, at which fluorescence recovers to the average steady state level, Fig 1E [26]. To reduce the effect of noise in these calculations, we applied a linear fit to ROA and FRAP experiments and determined τ_{ROA} and τ_{FRAP} when these intensities intersect defined thresholds of zero intensity for ROA or the mean recovered intensity for FRAP. For FCS, we estimate τ_{05} as the time the autocorrelation function drops below 5% of the zero-lag covariance and calculate $\tau_{FSP} = \tau_{05}/0.95$.

The specific location of probes along the mRNA has different effects on the fluorescence kinetics for the three experimental analyses. The characteristic decorrelation time in FCS and recovery time in FRAP are both set by the time it takes a single ribosome to translate from the tag region to the end of the mRNA. To reflect this, we define the approximate probe location, np_{FCS} or np_{FRAP} in Eq. 20, as the beginning of the tag region. In this case, the beginning of the tag region is at the beginning of the gene, but in general, we note that moving the fluorescent tag regions downstream toward the 3' end would shorten the effective times measured using FCS or FRAP. In contrast, for the ROA, the characteristic time is defined by how long it takes from when translation initiation is blocked until all ribosomes complete translation. Because this time depends solely on the gene length, and not on the probe placement, we assume $np_{\text{ROA}} = 1$, independent of probe placement. In addition to these effects on average experiment timescale estimates, we note that placing probes as near as possible to the 5' end of the mRNA or using longer proteins increases the fluorescence signal-to-noise ratio for all three approaches and can reduce estimation uncertainties.

To generate simulated data, we assumed that all 2,647 genes in the library have a global average translation rate of $\bar{k}_e = 10\ \text{sec}^{-1}$ and an initiation rate of $k_i = 0.03\ \text{sec}^{-1}$. For each experiment type and each gene, we simulated time lapse microscopy data for 100 independent RNA and for 300 frames at 1/3 frames per second (FPS). We then estimated $\tau^{(g)}$ from these simulations using each of the three experimental methodologies, and we estimated the corresponding average elongation rate using the specific gene sequence and Eq. 4. Under these conditions, Figs 3A-C (top) show the resulting distributions of estimated $\tilde{k}_e$ for long genes (> 1000 codons, n = 658, purple), medium length genes ($500 - 1000$ codons, n = 1719, blue), and short genes (< 500 codons, n = 270, orange) using each of the three experimental approaches. When all genes were analyzed at the same imaging conditions (100 spots, 300 frames, 1/3 FPS), the FCS approach was the most accurate with root mean squared (RMSE) of 0.64, 1.27, and 1.40 for short, medium and long genes, respectively. For comparison, ROA had RMSE of 2.22, 2.52, and 1.78 and FRAP had RMSE of 5.22, 4.58,

and 2.68 for the same combinations of genes and imaging conditions.

We next extended our analysis to consider different numbers of spots and different frame rates at which to collect the data, but under the assumption that the total number of frames would remain fixed at 300. Fig 3A shows the corresponding resulting RMSE for different combinations of these experiment designs. As expected, we found the sampling rate and number of mRNA spots to directly affect the estimated $k_e^{(FCS)}$. FCS was the only technique capable to estimate the true elongation rate within a $RMSE_{\text{FCS}} \leq 2.0\,\text{sec}^{-1}$ for short, medium and long genes. For short genes, this could be accomplished with as few as 10 spots with a frame rate of 1/3 FPS. Medium length and long genes could also be accurately quantified with 10 spots at frame rates of 1/3 FPS or 1/10 FPS.

The ROA was also capable to estimate the elongation rate to an accuracy of $RMSE_{\text{ROA}} < 2.0\,\text{sec}^{-1}$ for medium and long genes, and for fast frame rates, the ROA approach could be more accurate than FCS. However, when applying the ROA method to short genes, we obtained $RMSE_{\text{ROA}} > 2.0\,\text{sec}^{-1}$ under all combinations of sampling rates and repetition numbers at 100 or fewer spots, Fig 3B. This effect can be explained in that the number of ribosomes actively translating each mRNA is small and highly susceptible to stochastic effects in the case of small genes. We also note that the error using ROA depends strongly on the precision of the estimate for the specific time at which translation is blocked after application of Harringtonine; if the average value of this time is unknown, or if variations exceed our assumed standard deviation of 10 seconds, then accuracy using ROA is severely diminished, especially for short genes.

We found that FRAP substantially overestimates the elongation rates for short size genes, which can be observed on Fig 3C, where it is shown that recovering a $RMSE_{\text{FRAP}} < 2.0\,\text{sec}^{-1}$ was not possible for any of the considered combinations of number of RNA spots and sampling rates. We argue that the estimate of elongation rates using FRAP is limited by the intrinsic formulation of the fluorescent probe design. FRAP requires an intensity generating mechanism to reestablish the fluorescence to a pre-perturbation steady state. For single-molecule translation studies, this mechanism relies on ribosomal initiation events that are rare and highly susceptible to variability [6, 7, 8, 9]. This variability is reflected in the estimated τ_{FRAP} and in the final estimated elongation rate. Even for the more favorable medium and long length genes, our results indicate that for FRAP, a large number of mRNA spots (>100 mRNA spots) would be needed to achieve accurate estimates (Fig 3C).

2.3 | Calibration of the Stochastic Translation Model using Quantitative Data from Single-RNA Translation Experiments

Having determined that the FCS approach provides the most consistent estimate of elongation rate for genes of different lengths, we next turn to published experimental FCS data that quantified the fluctuation dynamics for three human gene constructs of different lengths: KDM5B (1549 aa), β-actin (375 aa) and H2B (128 aa) [6]. Each construct encodes for an N-terminal 10X FLAG 'Spaghetti Monster' SM-tag (318 aa) followed by the specific protein of interest (POI), and the stop codon for each POI was followed by 24 repetitions of the MS2 tag in the 3' UTR region. For each construct, the MS2 signal was used to track the mRNA motion in three dimensions, and the co-localized fluorescence intensity of the FLAG SM-tag was quantified as a function of time. These movies were collected using frame rates of 1 sec for H2B (n=10), 3 sec for β-actin (n=17), and 10 sec for KDM5B (n=44), and each trajectory was tracked for 300 frames per mRNA. Figs 4A-C (left) show example time traces (in arbitrary units of fluorescence) for the nascent protein level per individual mRNA for each of the three genes.

To quantify the steady state variability of nascent proteins per mRNA, the units of protein fluorescence were converted to units of mature protein (ump) using a calibration construct that contains only a single epitope for FLAG ([6], see Methods). After calibration, the effective numbers of mature proteins per mRNA were estimated for 300 translating mRNA spots for each gene, the histograms of collected measurements were normalized to estimate the probability of

having i nascent protein per mRNA, $P_{\text{dat}}(i)$, for each gene, and the results are presented by the gray lines in Figs 4A-C, middle.

We explored if the full stochastic model could be fit to capture simultaneously the experimentally measured steady-state distribution of nascent proteins and the temporal dynamics of nascent protein fluctuations on single mRNA. For model calibration, we defined ε_{KS} as the Kolmogorov-Smirnoff distance between the simulated and measured distributions of nascent protein per mRNA:

$$\varepsilon_{\text{KS}} = \max_k \left| \sum_{i=0}^{k} P_{\text{mod}}(i|\theta) - \sum_{i=0}^{k} P_{\text{dat}}(i) \right|, \tag{24}$$

where θ represents the model parameters, and $P_{\text{mod}}(i|\theta)$ and $P_{\text{dat}}(i)$ were the simulated and measured probability for an mRNA to have a fluorescence intensity corresponding to i units of mature proteins (ump). As non-translating spots could not be separated from spots below a basal FLAG intensity in the experimental data measurements, comparison between simulations and measured distributions ignore all spots with an intensity value less than $1/2$ ump.

To compare temporal dynamics of the experiment and data, we match the normalized autocorrelation function for the fluorescence intensity predicted by the model to that of the measured time-lapse experiments (details in Methods). Specifically, we quantified the two-norm difference in the measured and simulated autocorrelations, ε_{t}, according to:

$$\varepsilon_{\text{AC}} = \sqrt{\frac{\sum_{t=1}^{T} (G_{mod}(t|\theta) - G_{data}(t))^2}{T}}, \tag{25}$$

where $G_{\text{mod}}(t|\theta)$ and $G_{\text{dat}}(t)$ were the simulated and measured autocorrelations, respectively.

To enforce that the model would fit simultaneously to both the steady state distributions and the temporal dynamics, we defined a multi-objective function to quantify the total difference between model and data as:

$$\varepsilon^{(g)} = \varepsilon_{\text{KS}}^{(g)} + \varepsilon_{\text{AC}}^{(g)}, \tag{26}$$

for any gene denoted by g.

Codon-dependent translation rates were assumed to be consistent among the three genes, as defined in Eq. 4, but the three genes were allowed to have different initiation rates, $\{k_i^{(g)}\}$. Under this assumption, the model has a total of four parameters. Upon fitting these parameters to minimize $\varepsilon = \varepsilon^{(\text{H2B})} + \varepsilon^{(\beta-\text{actin})} + \varepsilon^{(\text{KDM5B})}$, we found that the model could capture both the experimental distributions of nascent proteins per mRNA and the autocorrelation plots for all three genes, as shown in Fig 4A-C (middle and right). Optimized parameters and their uncertainties (see Methods) were found to be:

$$\begin{cases} \bar{k}_e = 9.8 \pm 1.24 sec^{-1}, \\ k_i^{(\text{H2B})} = 0.07 \pm 0.01 sec^{-1}, \\ k_i^{(\beta-\text{actin})} = 0.039 \pm 0.005 sec^{-1}, \\ k_i^{(\text{KDM5B})} = 0.02 \pm 0.002 sec^{-1}. \end{cases} \tag{27}$$

2.4 | Exploring Effects of Parameters on Ribosome Activities and Translation Dynamics

After determining that our model was sufficient to reproduce the experimentally measured fluctuation dynamics for H2B, β-actin, and KDM5B, we next extended our analyses to consider a broader range of translation parameters.

Specifically, we sought to explore the effects of variations to initiation and elongation rates as well as effects of synonymous codon substitutions or modulation of tRNA concentrations.

2.4.1 | Ribosome Collisions are Rare at Most Experimentally Observed Translation Initiation and Elongation Rates

Previous experimental reports [6, 7, 8, 9, 10] estimated a range of values from 0.01 to 0.08 sec^{-1} for the translation initiation rate, k_i, and range from 3 to 13 aa/sec for the average elongation rate, $\bar{k_e}$. Using β-actin gene as a reference, Fig 5-A depicts the variation in ribosome density as a function of the base parameters k_i and $\bar{k}_e$, and Fig 5B shows the number of times an average ribosome would collide with an upstream neighboring ribosome during a single round of translation. For most parameter combinations, ribosome loading was predicted to be very low (i.e., fewer than one ribosome per 100 codons) and collisions were rare (i.e., fewer than 10 collisions in an average round of translation). However, for slow elongation and fast initiation, such as those measured by Wang et al. [7]), a ribosome could collide with other ribosomes an average of ~ 20 times for a gene the length of β-actin. To further illustrate the effects that these initiation and elongation rates would have on ribosome dynamics on different genes, Fig. 5C shows simulated kymographs for SunTag24x-Kif18b [10], Flag-10x-KDM5B [6], and SunTag56x-Ki67 [9], each with their previously reported initiation and elongation rates. In addition, Appendix Figures S1 and S2 provide more detailed results of the translation elongation simulations for β-actin translation at multiple initiation rates and elongation rates, respectively. Each of these kymographs indicate that ribosome dynamics can vary from collision-free dynamics (SunTag24x-Kif18b and Flag-10x-KDM5B) to dynamics with multiple collisions (SunTag56x-Ki67) and that collisions can become more prevalent at high initiation rates or low elongation rates.

2.4.2 | Codon Usage Affects Translation Speed and Ribosome Loading

Simulations of genes H2B, β-actin, and KDM5B showed that each gene's codon order has an influence on the overall ribosome traffic dynamics, creating a non-uniform distribution of ribosomes along the mRNA (Appendix Figure S3). This observation of codon dependence led us to look more deeply into possible effects that optimization could have on observable translation dynamics. Appendix Figure S4 depicts simulated kymographs for the β-actin protein for three synonymous sequences containing: (i) natural codons, (ii) most frequent synonymous codon (optimized), and (iii) least frequent synonymous codon (de-optimized). For each case, Appendix Figure S4B illustrates the corresponding ribosome loading profiles; Appendix Figure S4C shows the simulated distribution of FLAG intensities in units of mature proteins, and Appendix Figure S4D presents the corresponding simulated fluorescence autocorrelation functions. Appendix Figure S5 and S6 show similar results for the H2B and KDM5B genes, respectively.

In all cases, optimized gene sequences speed-up ribosome dynamics, and de-optimized sequences cause a slower elongation rate that could be observed in the autocorrelation plots given in Appendix Figures S4D, S5D and S6D. Moreover, for constant initiation rates, faster elongation would lead to lower ribosome loading (Appendix Figures S4B, S5B, S6B) and therefore lower fluorescence intensity distributions (Appendix Figures S4C, S5C, S6C). All three genes under consideration had natural codon usage that was enriched for the most common codons (i.e., the natural and common codon usage dynamics are very similar), such that the translation rate, ribosome loading, and fluorescence intensity could be substantially altered only by substitution to rare codons. We note that the substitution of rare codons would lead to slower elongation and substantially higher numbers of ribosome collisions.

2.4.3 | Depletion of tRNA Levels can Induce Ribosome Traffic Jams

In addition to modulating translation speed through codon substitution, it is possible to perturb these dynamics through experimental modulation of tRNA concentrations. For example, Gorgoni et al., [17] used a mutated allele to the gene for $tRNA_{CUG}$ to reduce the concentration of the glutamine tRNA. To study how ribosome dynamics can be affected by the removal or addition of specific tRNA, we simulated the translation dynamics of H2B, β-actin, and KDM5B at several different concentrations for $tRNA_{CTC}$. Appendix Figure S7 shows the effect of decreasing $tRNA_{CTC}$ concentration on the ribosome association time (left) and elongation rate (right). The simulations show that ribosome dynamics are relatively unchanged provided that the $tRNA_{CTC}$ concentration remains above approximately 10% of the native level. In contrast, depleting $tRNA_{CTC}$ concentration below 10% of wild-type levels could lead to ribosome stalling, which was reflected in long ribosome association times and low effective elongation rates. Ribosome traffic-jams are observed under very low $tRNA_{CTC}$ concentration as shown in Appendix Figures S8 to S10. The prevalence of the CTC codon was found to be important in that the effect of $tRNA_{CTC}$ depletion occurs at higher $tRNA_{CTC}$ concentrations for the CTC codon rich KDM5B gene than for the other two constructs.

2.5 | RNA Sequence to NAscent Protein Simulation (RSNAPSIM)

To facilitate the simulation of single-molecule translation dynamics, all models and analyses described above have been incorporated into a user-friendly Python toolbox, which we have called RSNAPSIM. This toolbox combines a graphical user interface (GUI) divided into multiple tabs, graphical visualizations, and tables to present calculated biophysical parameters (see Fig 6). This simulator performs stochastic simulations considering the widely accepted mechanisms affecting ribosome elongation, such as codon usage and ribosome interference. The toolbox is available in Python 2.7/3.5+ and wrappers for optimized C++ code are provided with installation instructions.

RSNAPSIM takes as an input the gene sequence in Fasta format or an NCBI accession number. The user can decide on the type (FLAG, Sun-Tag, or Hemagglutinin), number, and placement of different epitopes upstream, downstream or within the protein of interest. The toolbox provides the user with a visualization of the gene sequence and the overall gene construct including the position of the POI and the positions of the Tag epitopes. From the concatenated tags and POI sequences, RSNAPSIM automatically generates a discrete single-RNA translation model with single amino acid resolution and codon-dependent translation rates. Once generated, these models can be simulated using stochastic dynamics, and the results can be quantified in terms of predicted translation spot intensity fluctuations (i.e., single-RNA translation time traces or kymographs), ribosomal density profiles, and fluorescence signal autocorrelation. The graphical user interface also provides for easy generation of simulated results for several different experimental assays, including FCS, FRAP, and ROA. From these simulation results, biophysical parameters such as the overall elongation rate or ribosome association rate are automatically calculated and returned to the user. The toolbox provides additional interfaces for the user to design and simulate gene sequences with substitution between natural, common, or rare codons for any combination of amino acids or to manually adjust the concentration of tRNA for specific codons. Simulations are saved automatically so that the user can compare translation dynamics for multiple different gene constructs.

The open source toolbox was tested in Mac, Windows, and Linux operating systems and is available at: `https://github.com/MunskyGroup/Translation_Simulator.git`. Simulating a gene with 1567 codons for 100 repetitions of 5000 seconds each takes less than 1 minute using a laptop computer with a core i7 and 32GB of RAM.

3 | DISCUSSION

Imaging translation in living cells at single-molecule resolution is still a nascent technology with a limited number of experimentally studied genes [6, 7, 8, 9, 10, 14, 15], but the number of such studies is expected to grow considerably in the near future [12]. Computational models can aid in this research by predicting the best utilization of experiments to quantify biophysical parameters from single-molecule data. Here, we introduce a theoretical framework that includes the most widely accepted mechanisms affecting ribosome elongation, including codon usage and ribosome interference [27]. To complement previous models that have sought to reproduce data from earlier bulk cellular assays [17], and ribosome profiling data [28, 29], our focus has been to integrate single-mRNA stochastic dynamics models with data from *in vivo* single-RNA translation dynamics experiments.

We developed a general codon-dependent model, where nascent protein distributions and autocorrelation functions were generated by detailed stochastic simulations that tracked the positions of ribosomes relative to their neighbors. However, in the absence of perturbations to change initiation and elongation rates, most ribosomes do not encounter others during elongation (Fig 2), at least not at currently accepted elongation and initiation rates from the literature [6, 7, 8, 9, 10]. This observation justifies an assumption of sparse ribosome loading and independent ribosome motion, which allow the linear reaction rate reformulation of the codon-dependent translation model into a simplified stochastic moment model (Section 2.1.1) and further reduction led to analytical expressions for the steady state mean and variance of fluorescence in units of mature protein levels per mRNA (Eq. 21) and for the decorrelation time (Eq. 18). For initiation rates at or below reported experimental values, the simplified analytical model and the full model are in strong agreement (Fig 2). However, increasing initiation rates relative to the base elongation rate, inserting more rare codons into the sequence, or depleting tRNA levels for some codons will increase the number of ribosome collisions and violate the simplifying assumptions (Figs 2C, 5). In such circumstances, the full stochastic model predicts slower effective elongation rates, longer ribosome association times, and accumulation of more ribosomes per mRNA.

With the full and reduced models in hand, it become possible to predict how well three modern methodologies would estimate elongation rates from single-molecule measurements: fluorescence correlation spectroscopy (FCS) [12], fluorescence recovery after photo-bleaching (FRAP) [12, 9, 8], and run-off assays (ROA) after perturbation with inhibitory drugs [10, 7]. Through simulations on 2,647 genes, we demonstrated that estimating elongation rates for long genes (>1000 codons) could be achieved with great accuracy using any of these methodologies, provided that a minimal number of mRNA spots are considered and with an appropriate temporal resolution as demonstrated in Fig 3. However, our results suggest that FCS would be the most likely method to provide an accurate elongation rate estimate (Fig 3A), especially for small and medium size genes. Although our simulation results suggest that FCS is the best single-molecule option to estimate elongation rates, it is important to remark that FCS analysis requires the tracking and measurement of intensity for single spots over long periods of time, and such measurements are susceptible to photo-bleaching and molecular motion. The former issue has been addressed through application of optical techniques such as highly inclined thin illumination microscopy [30] and the latter could be addressed through application of molecular tethers to reduce motion [10]. On the computational side, one could potentially address concerns of bleaching or motion relative to the imaging plane by including hyper-parameters to describe these dynamics and then fit these hyper-parameters concurrently with model parameters using Bayesian analyses.

Run-off assays using Harringtonine to prevent translation initiation can give accurate estimates when genes are long, but the accuracy of such an approach is highly diminished for shorter genes (Fig 3B) or if the precise time of drug action on the mRNA is not known. Our analyses suggest that run-off assays directly depend on the number of ribosomes actively translating the mRNA at the time of perturbation, and since this number is highly susceptible to stochasticity on small genes, the ROA would require analysis of a much larger number of spots to achieve accurate rest estimates. Our

analyses show that FRAP gives poor estimates for all genes of all sizes, and for all tested experimental designs, Fig 3C. The recovery of the intensity after photobleaching depends heavily on the initiation rate, which has been found to be an order of magnitude smaller than the elongation rate, making the recovery a highly stochastic process as well. We directly compared the error size for the studied methods, obtaining that the error in FRAP and ROA is two times larger than in FCS, Appendix Figure S11.

Using FCS data, we demonstrated that a codon-dependent translation model containing one universal average elongation rate and one gene-dependent initiation rate could capture quantitatively the distribution of nascent proteins per actively translating mRNA, as well as the temporal dynamics for three different genes expressed in human U2OS cells (Fig 4). Combining these estimates of initiation and elongation rates with reported values for the same rates identified using other methods and for other genes, we could predict ribosome dynamics and nascent protein intensities for reported gene sequences [6, 10, 7, 8, 9, 14, 15], (Fig 5). Those results allowed us to conclude that relatively fast elongation rates help maintain substantial space between ribosomes on a single mRNA. As a result, these ribosomes should not often collide, and the final ribosome-mRNA association times should remain unchanged for typical initiation rates, natural codon usage, and normal tRNA availability, as shown in Appendix Figure S3. Nevertheless, ribosome dynamics may be affected by genetic or environmental perturbations, such as increased initiation rates (Appendix Figure S1), reduction of elongation rates (Appendix Figure S2), enrichment for rare codons (Appendix Figure S4 to S6), or depletion of tRNA (Appendix Figure S7 to S10).

The present model and RSNAPSIM toolkit have intentionally been made as general and adaptable as possible to efficiently simulate and capture the most accepted mechanisms taking place during translation, i.e. codon-dependent elongation and ribosome interference. At present, the specific rates of codon-dependent elongation are only approximate and based on the prevalence of the the corresponding tRNA in the human genome. By modifying this assumption, it is possible to further improve fits for the elongation dynamics shown in Fig 4, and one could find codon dependent rates to explain the diversity of experimentally measured elongation rates depicted in Fig 5. For now, we argue that data from fewer than a dozen genes (and in different cell lines) is as yet insufficient to fully constrain codon dependent rates for all 64 codons. However, as new data is collected for more and more genes, we envision that it will become possible to tune these parameters with greater precision and to capture a greater complement of genes.

In addition to variation in initiation, elongation, codon usage, and tRNA concentrations, many other factors have been described to affect ribosome dynamics. These include, but are not limited to, ribosome stalling or drop-off, pauses due to secondary structures of the specific mRNA, and the electrostatic and hydrophobic interactions between the mRNA and the ribosome [27, 29]. We expect that increased prevalence of single-RNA translation experiments will add to the current understanding and reveal additional mechanisms taking place during translation. At the same time, such discoveries are bound to create new layers of model complexity. Although these mechanisms have not yet been implemented in our present model, they can be captured easily through modification of the set of elongation parameters, $k_e(i)$. For example, the RSNAPSIM toolbox allows for direct modification of elongation rates at an specific codon, which can be used to mimic pauses at certain locations. Furthermore, all of the computational analyses describe above are easily adapted to allow for simultaneous multi-frame and multi-color translation, as we have implemented in a parallel study to capture the dynamics of single-mRNA translational frame-shifting [14]. A main limitation in the experimental determination and quantification of translation mechanisms is the specific design of experiment to make that quantification. For example, in its current form, the introduction of tag regions in the open reading frame of the gene of interest can dramatically alter the overall translation dynamics. As depicted in Fig 1B, the tag region is around 300 codons in length, and this added length can substantially bias the measurement biophysical parameters, especially when quantified using FRAP or runoff assays (see Fig 3). On one hand, our model can help to explain these differences (Appendix Figure S11), but more importantly, the models themselves can be used to test in simulation

different experimental designs that are more likely to reveal important biophysical mechanisms or parameters. We envision that user-friendly simulations, such as those provided by RSNAPSIM, can be used to optimize combinations of probe placement, gene length, codon usage differences, video frame rates, drug-based perturbations, or specifications of movie length. Such simulation-based designs can be conducted prior to any new experimental analysis and then used again to fit the results of those experiments, to pinpoint discrepancies that may reveal new mechanisms, and to refine model parameters and mechanisms. Such an integration of experiment and computational model can help set the stage for more efficient experiments that specifically target and quantify the full complement of factors that modulate translation dynamics in living cells.

4 | MATERIALS AND METHODS

4.1 | Studied gene constructs

To constrain our analyses, we use published gene sequences used on single-molecule translation studies. An initial set of sequences were obtained from Morisaki et al., [6], these constructs encode an N-terminal region with 10 repeats of FLAG-SM-tag followed by one of three different genes of interest: KDM5B (1549 aa), β-actin (375 aa) and H2B (128 aa), the 3' UTR region contains 24 repetitions of the MS2 stem-loops. A second source of gene sequences comes from Yan, et al., [10], this gene construct encodes 24 repeats of SunTag followed by the gene of interest kif18b (1800 aa), and the 3' UTR contains 24 repeats of the PP7 bacteriophage coat protein. A sequence encoding 56 SunTag repeats, the gene of interest Ki67 (3177 aa), and the 3' UTR containing 132 repeats of MS2 stem-loops was obtained from Pichon et al., [9]. Finally, multiple gene constructs were build using 10 repeats of FLAG-SM-tag followed by a human gene. The studied human genes come from a comprehensive list of 2,647 gene sequences obtained from the PANTHER database [23].

4.2 | Correction to mean and variance of fluorescence intensity for the theoretical model

Neglecting ribosome exclusion, and under an assumption of memory-less initiation with exponential rate k_i, the number of ribosomes to initiate translation in a fixed time, τ, is described by a Poisson distribution with mean and variance equal to $k_i\tau$. For a single probe site, we can fix τ as the time it takes a ribosome to move from that site to the end of the mRNA, and the mean and variance of nascent protein fluorescence can be estimated in terms of units of mature protein fluorescence according to Eq. 21.

However, for probes that are spread out across a finite tag region, this distribution requires a slight correction to account for ribosomes within the probe region that only exhibit partial protein fluorescence. Let $\alpha(s)$ denote the intensity, scaled in units of mature protein, exhibited by a ribosome at the position, s, along the mRNA as follows:

$$\alpha(s) = \begin{cases} s/L_{\mathrm{t}} & \text{for } 0 \leq s < L_{\mathrm{t}} \\ 1 & \text{for } L_{\mathrm{t}} \leq s \leq L \end{cases} \tag{28}$$

Under an assumption of uniform codon usage, a given ribosome on the mRNA has equal probability to be at any site along the mRNA. If there are an average of mu mRNA total on the mRNA, then the number at each location is approximated by a Poisson distribution with mean and variance both equal to $mu/L \cdot ds$. Recall that the mean of the sum of two independent random variables is the sum of two means. Therefore, to find the total mean intensity contribution for all

ribosomes on an average mRNA (Eq. 22), we can integrate along the length of the mRNA to find:

$$\mu_I = \int_0^L \frac{\mu}{L}\alpha(s)ds, \tag{29}$$

$$= \left(1 - \frac{L_t}{2L}\right)\mu. \tag{30}$$

Similarly, we recall that the variance of a random variable with variance σ^2 and scaled by α is equal to $\alpha^2\sigma^2$ and the variance for the sum of two such variables is the sum of the corresponding variances. Therefor, by noting that $\mu = \sigma^2$, we can find the total variance of intensity on a single mRNA (Eq. 23) as:

$$\sigma_I^2 = \int_0^L \frac{\mu}{L}\alpha(s)^2 ds, \tag{31}$$

$$= \left(1 - \frac{2L_t}{3L}\right)\mu. \tag{32}$$

4.3 | Fluorescence Correlation Spectroscopy (FCS)

FCS is usually implemented by computing and comparing the autocorrelations of fluorescence intensities of one or more particles within small fixed volumes [31, 32], but similar correlation analyses have been used to quantify intensity fluctuations for tracked single particles [2]. For our analysis, we compute the temporal autocorrelation times of the FLAG fluorescence signal intensity for a moving volume that is centered around the moving RNA spot.

To estimate the rate of translation elongation, we took the following approach: first, each experimental and simulated intensity time courses were centered to have zero mean by subtracting the average intensity of the time series. Next, we computed the correlation function of the fluorescence intensity for each intensity spot according to the standard formula:

$$G(\tau) = \mathbb{E}\{(I_t - \mu_t)(I_{t+\tau} - \mu_{t+\tau})\}, \tag{33}$$

where τ denotes the time delay and $\mathbb{E}\{v\}$ denotes the expectation of some arbitrary value v.

To reduce the effects of high-frequency shot noise and tracking errors that are not considered in the model, the null-delay autocorrelation $G(0)$ was removed from the analysis [33]. The autocorrelations for $\tau > 0$ were then normalized with respect $G(\Delta t)$, where $\Delta t > 0$ corresponds to frame rate for that experiment (i.e., $G(\Delta t)$ is the co-variance of spot intensity between two successive frames). After normalization, we implemented the *multi-tau algorithm* described in detail elsewhere [34]. For statistical purposes, autocorrelations for multiple intensity time courses were calculated, and their value was averaged. Final results are reported as mean values and standard error of the mean (SEM). This signal analysis allowed us to measure the dwell time (τ_{FCS}) at which $G(\tau) = 0$, from which the average ribosome elongation rate can be calculated as:

$$k_e^{(FCS)} = L/\tau_{FCS}. \tag{34}$$

4.4 | Parameter Uncertainty

Parameter uncertainty analyses was calculated by building parameter distributions that reproduce results within a 10% error, calculated from 1,000 independent simulations using randomly selected parameter values. Simulations were

performed on the W. M. Keck High Performance Computing Cluster at Colorado State University.

4.5 | Numerical Methods

For solving the model under stochastic dynamics we used the direct method from Gillespie's algorithm [35] coded in Matlab 2018b and Python 2.7. ODE models were solved in Python 2.7.

4.6 | Acknowledgments

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM124747 and by the WM Keck Foundation. The work reported here was partially supported by a National Science Foundation grant (DGE-1450032). Any opinions, findings, conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the funders. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors thank Kenneth Lyon and members of the Munsky lab for their feedback on the presented analyses and for their testing of the RSNAPSIM software.

5 | CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

TABLE 1 **Comparing model dynamics.**

	Stochastic Model	Moments-Based Model	Theoretical Model
$KDM5B, L = 1867$aa			
mean (μ)	5.2 ± 0.02	5.07	5.01
var (σ^2)	4.5 ± 0.03	4.90	4.93
period (τ)	187.78 ± 0.94	180.0	185.23
$\beta - actin, L = 693$aa			
mean (μ)	1.31 ± 0.004	1.34	1.42
var (σ^2)	1.08 ± 0.003	1.17	1.28
period (τ)	62.17 ± 0.27	60.0	60.94
$H2B, L = 446$aa			
mean (μ)	0.75 ± 0.002	0.77	0.82
var (σ^2)	0.56 ± 0.001	0.59	0.68
period (τ)	42.96 ± 0.07	42.0	41.80

Mean and variance of intensity are given in units of mature proteins (ump). The period (τ) has units of seconds. Elongation and initiation rates are $\bar{k}_e = 10\text{s}^{-1}$ and $\bar{k}_i = 0.03\text{s}^{-1}$, respectively. Lengths include the tag region of 318aa. Stochastic simulations were performed for 500 simulated spots, with a frame rate of 1 sec, and for 2,000 frames. Error values represent the standard deviation of 3 repetitions of independent simulations.

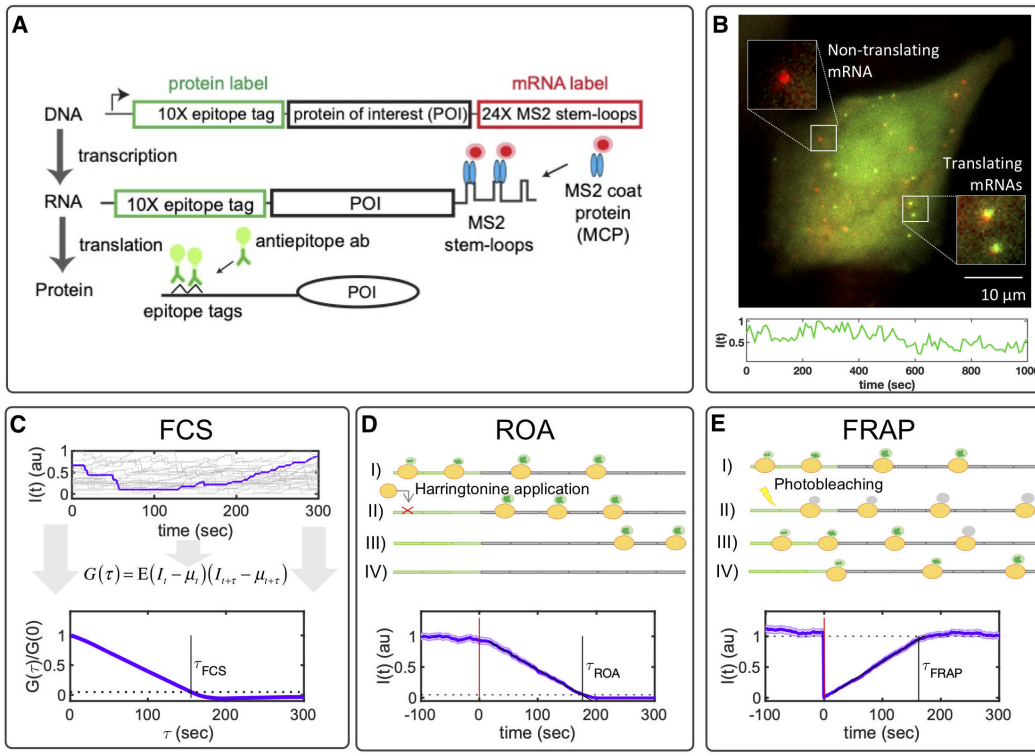

FIGURE 1 Translation studies with single-molecule resolution. A) Imaging single-molecule translation dynamics is achieved by the measurement of fluorescence spots that are produced when nascent proteins display epitopes that are recognized by antibody fragments bound to fluorescent probes. The gene construct encodes a 10X FLAG SM tag followed by a protein of interest (POI) and the 24X MS2 tag in the 3' UTR region. B) Microscopy image showing translation at single-molecule resolution; red spots represent single mRNA, and green spots represent nascent proteins. Below is a representative trace showing the intensity fluctuation dynamics of a single-transcript translating 10X-FLAG-KDM5B. C) Simulated time courses representing the characteristic single-molecule fluctuation dynamics. A representative trace is selected and highlighted. At the bottom of the figure is given the autocorrelation function (*G*) calculated from simulated time courses. The time at which the autocorrelation function hits zero represents the dwell-time (τ_{FCS}). D) Harringtonine inhibits the translation initiation step by binding to the ribosomal 60S subunit. Plot shows the average fluorescence after Harringtonine treatment. Without new initiation events, the fluctuations diminish causing the intensity to drop to zero at time τ_{ROA}. E) FRAP causes a rapid drop in the fluorescent intensity and a subsequent recovery that is proportional to the time needed by ribosomes to produce new nascent proteins with non-photobleached probes. The bottom plot shows the temporal dynamics of FRAP, where it can be observed by the abrupt decrease in intensity and a recovery time (τ_{FRAP}) correlated with the gene length. All simulations correspond to KDM5B for 100 spots and a frame rate of 1 FPS. Error bars represent the standard error of the mean.

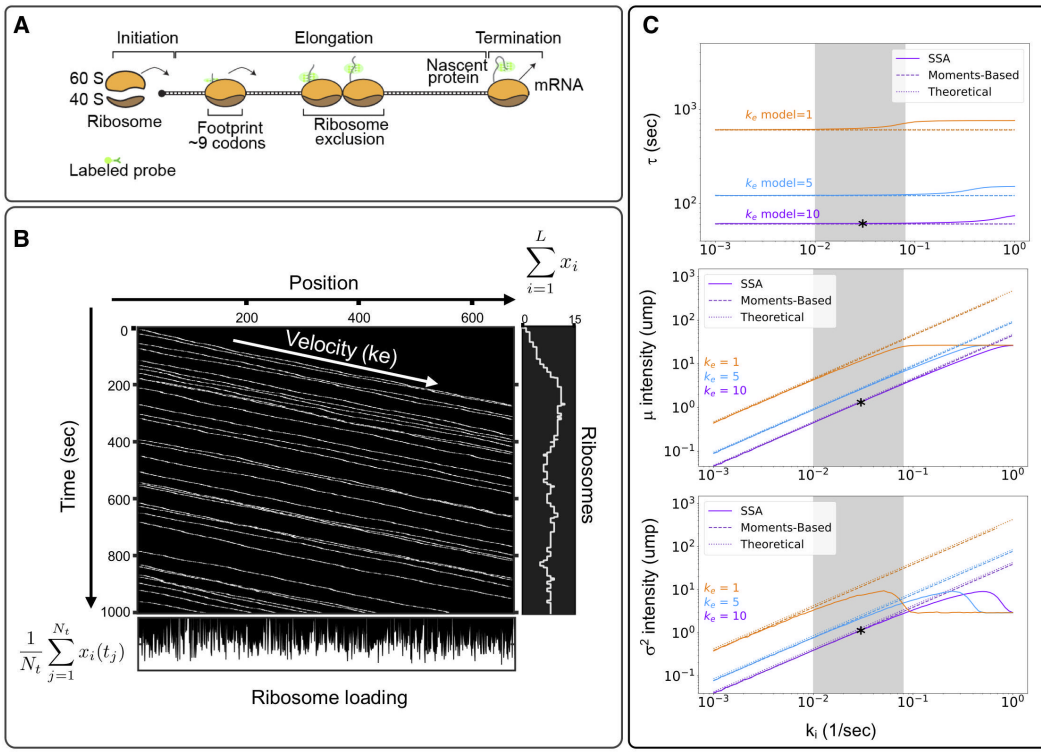

FIGURE 2 **Modeling single-molecule translation.** A) Translation is divided into three main processes: initiation, elongation, and termination. The ribosome footprint represents the physical space occluded by the ribosome, enforcing that that no two ribosomes can occupy the same space and time. B) Kymographs represent ribosome movement as a function of time. In the figure, the x-axis is the codon position and y-axis is time. Each line represents a single ribosome trajectory. The average slope is proportional to the effective ribosome elongation rate. The plot to the right shows the relationship between ribosome movement and fluorescence intensity, and the plot below shows the ribosome loading at each codon position, calculated as the average ribosome reads at each codon position for multiple time points N_t. C) Comparison of the average elongation time (top) and the mean (middle) or variance (bottom) of fluorescence intensity as calculated using the simplified model (Eqs. 18 to 21), a linear moments-based model (Eqs. 9 to 17), and a full stochastic model (Eqs. 1 to 5). Gray area represents previously reported parameter values for ribosome initiation. Panels B and C correspond to simulations for the β-actin. Asterisks represent the specific parameter combination used for Table 1.

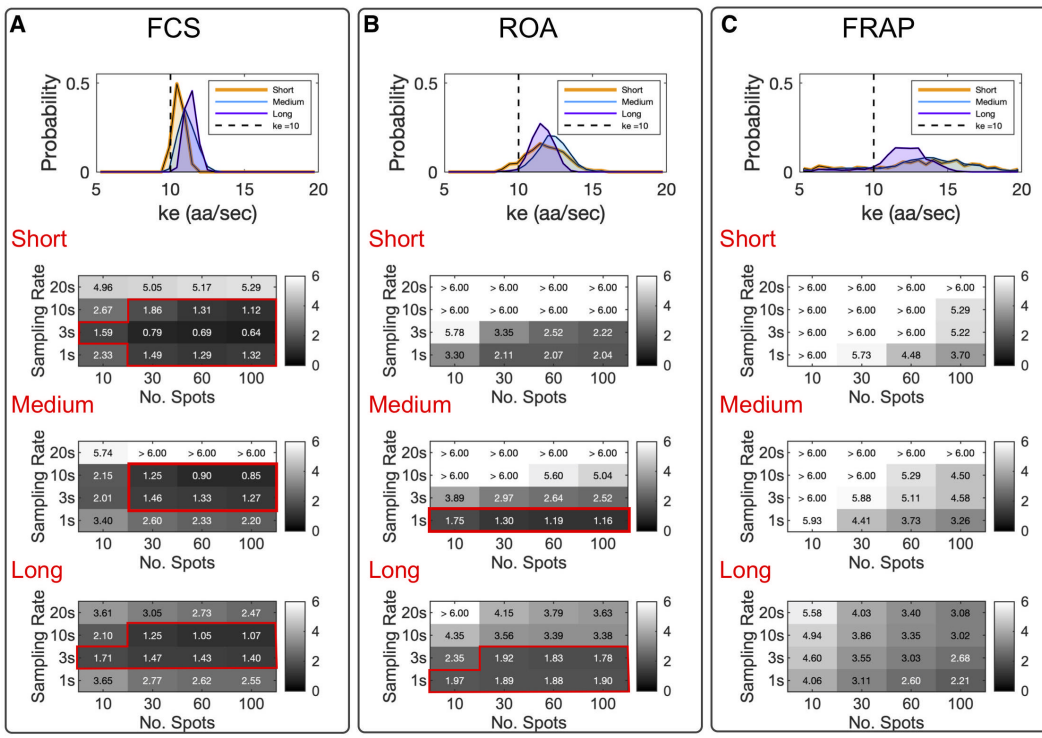

FIGURE 3 Comparing experimental methodologies to estimate ribosome elongation rates. Elongation rate estimate experiments were simulated for 2,647 human genes, using (A) fluorescence correlation spectroscopy (FCS), (B) run-off assays (ROA), and (C) fluorescence recovery after photobleaching (FRAP). Top panels show the distributions of estimated $\bar{k}_e$ for long genes (> 1000 codons, n = 658, purple), medium length genes (500 – 1000 codons, n = 1719, blue), and short genes (< 500 codons, n = 270, orange) using 100 mRNA spots for 300 frames at 1/3 FPS. The true elongation rate is denoted by a vertical dashed line. Bottom panels show the RMSE in elongation rate estimation as a function of the number of mRNA spots and the sampling rate. Red boxes highlight all experimental designs that yield a RMSE < 2.0. The 'true' elongation rate was set at $\bar{k}_e = 10$, and the initiation rate was fixed at $k_i = 0.03$ sec^{-1} for all simulations.

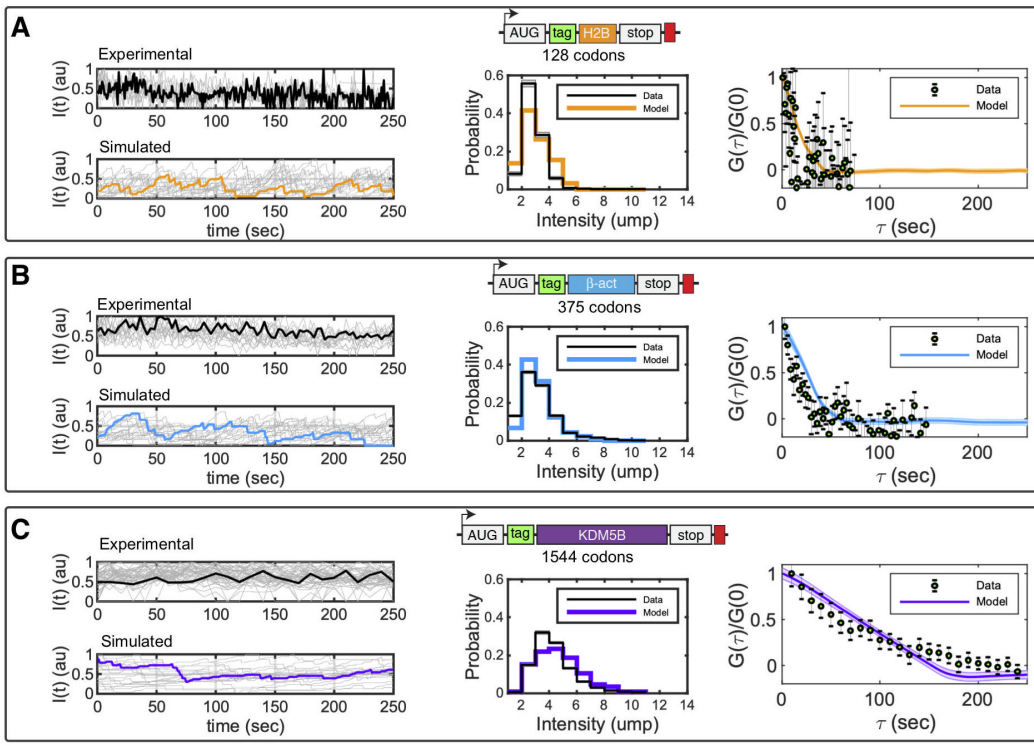

FIGURE 4 Fitting single-molecule data with the full stochastic model. Experimental data shows the fluctuation dynamics of a gene constructs encoding a N-terminal 10X FLAG 'Spaghetti Monster' SM-tag (green) followed by a protein of interest and finally a 24X MS2 tag (red) in the 3' UTR region. Three proteins were studied: A) KDM5B (magenta), B) β-actin (cyan) and C) H2B (orange). Middle figures show the simulated (colors) and measured (black) probability for an mRNA to have a fluorescence intensity corresponding to i units of mature proteins (ump). Right images show the autocorrelation function (G) calculated from experimentally measured (black error bars) and computationally simulated (colors) autocorrelation functions. Error bars in the experimental data and shadow bars in the simulated autocorrelation plots represent the standard error of the mean. Elongation and initiation rates were obtained by parameter optimization, using the Hooke and Jeeves Algorithm ([36]). Optimized parameters and their uncertainties (see Methods) are provided in Eq. 27.

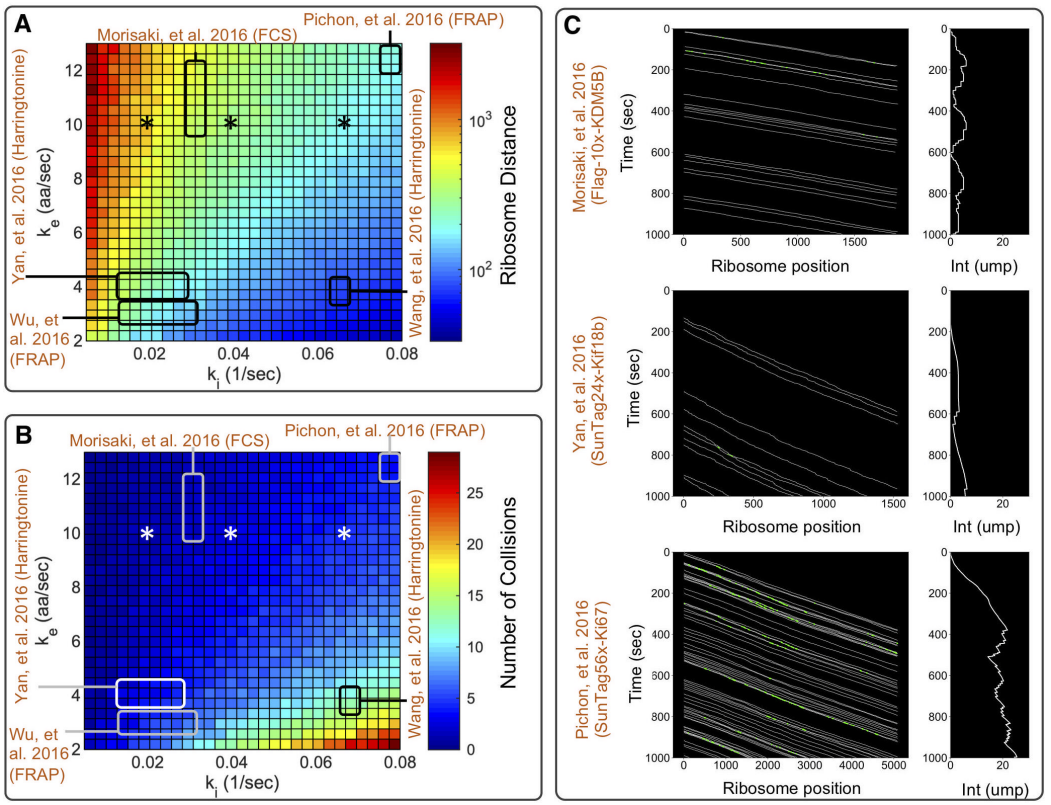

FIGURE 5 Ribosome dynamics under experimentally reported initiation and elongation rates. A) Simulated mean number of codons between ribosomes for the β-actin gene as function of initiation and elongation constants. In the plot, previous literature initiation and elongation values are highlighted by the squares [6, 10, 7, 8, 9], and values estimated in this study are denoted by asterisks. B) Simulated number of collisions per ribosome as a function of initiation and elongation constants. C) Top panel, kymograph showing the ribosomal dynamics for 24X-SunTag-Kif18b using experimentally determined parameters k_i = 1/100 sec^{-1} and $\bar{k}_e$ = 3.1 aa/sec [10]. Center panel, kymograph showing the ribosomal dynamics for 10X-FLAG-KDM5B using experimentally determined parameters k_i = 1/30 sec^{-1} and $\bar{k}_e$ = 10 aa/sec [6]. Bottom panel, kymograph showing the ribosomal dynamics for 56X-SunTag-ki67 using experimentally determined parameters k_i = 1/13 sec^{-1} and $\bar{k}_e$ = 13.2 aa/sec [9]. White lines in kymographs represent single ribosome positions, and green spots represent ribosome collisions.

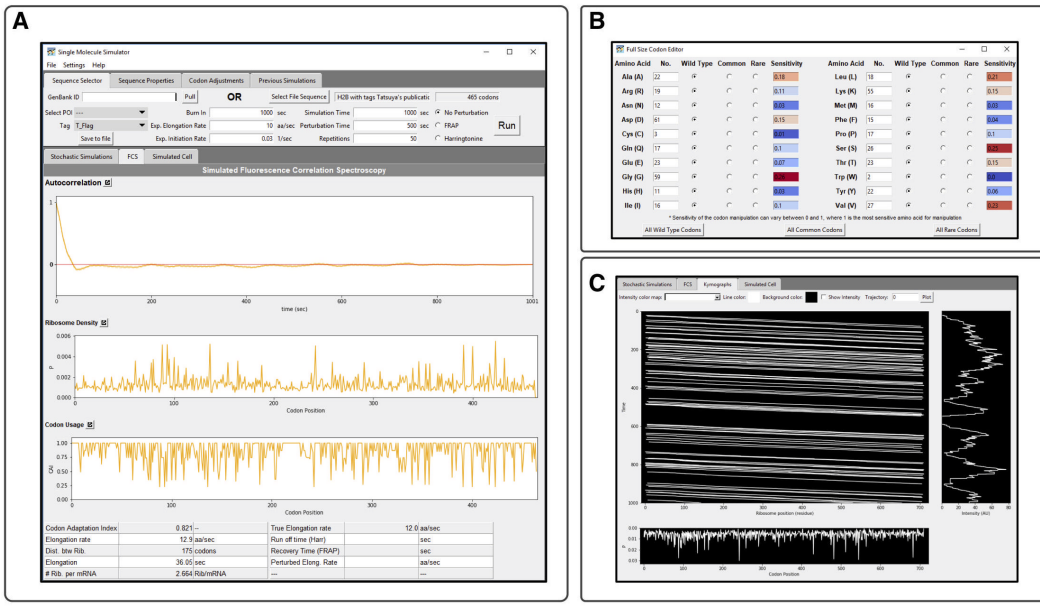

FIGURE 6 RNA Sequence to NAscent Protein Simulation (RSNAPSIM). A) RSNAPSIM is divided into four upper tabs and three lower tabs. Upper tabs allow the user to select and adjust sequences and then run simulations under varying conditions. Sequence selector allows the user to load a raw text file or GenBank file for their simulation needs. An option to poll GenBank via accession number is also available. All simulation parameters are also set on this tab. B) After a file is loaded, RSNAPSIM allows the user to change the tRNA copy numbers and codon types under the Codon Adjustment tab. Post simulation, the lower tabs display simulation information such as average intensity over time of N simulations. C) Screen-shot of a kymograph. The kymograph tab allows the user to create their own kymographs with varying display options. The other tabs are as follows: The Stochastic Simulation tab shows the time course data from the selected simulations. Fluorescence correlation spectroscopy tab displays simulate single-molecule translation dynamics, the autocorrelation function, and biophysical parameters, such as the elongation constant or ribosomal density. All functionality in the GUI is also available in a command line module for Python included with RSNAPSIM.

REFERENCES

[1] Bertrand E, Chartrand P, Schaefer M, Shenoy SM, Singer RH, Long RM. Localization of ASH1 mRNA particles in living yeast. Molecular cell 1998;2(4):437–445.

[2] Larson DR, Zenklusen D, Wu B, Chao JA, Singer RH. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. Science 2011;332(6028):475–478.

[3] Hocine S, Raymond P, Zenklusen D, Chao JA, Singer RH. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. Nature methods 2013;10(2):119.

[4] Mor A, Suliman S, Ben-Yishay R, Yunger S, Brody Y, Shav-Tal Y. Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells. Nature cell biology 2010;12(6):543.

[5] Buxbaum AR, Wu B, Singer RH. Single β-actin mRNA detection in neurons reveals a mechanism for regulating its translatability. Science 2014;343(6169):419–422.

[6] Morisaki T, Lyon K, DeLuca KF, DeLuca JG, English BP, Zhang Z, et al. Real-time quantification of single RNA translation dynamics in living cells. Science 2016;352(6292):1425–1429.

[7] Wang C, Han B, Zhou R, Zhuang X. Real-time imaging of translation on single mRNA transcripts in live cells. Cell 2016;165(4):990–1001.

[8] Wu B, Eliscovich C, Yoon YJ, Singer RH. Translation dynamics of single mRNAs in live cells and neurons. Science 2016;352(6292):1430–1435.

[9] Pichon X, Bastide A, Safieddine A, Chouaib R, Samacoits A, Basyuk E, et al. Visualization of single endogenous polysomes reveals the dynamics of translation in live human cells. J Cell Biol 2016;p. jcb–201605024.

[10] Yan X, Hoek TA, Vale RD, Tanenbaum ME. Dynamics of translation of single mRNA molecules in vivo. Cell 2016;165(4):976–989.

[11] Lyon K, Stasevich TJ. Imaging Translational and Post-Translational Gene Regulatory Dynamics in Living Cells with Antibody-Based Probes. Trends in Genetics 2017;.

[12] Morisaki T, Stasevich T. Quantifying Single mRNA Translation Kinetics in Living Cells. Cold Spring Harbor perspectives in biology 2018;10(11).

[13] Biswas J, Liu Y, Singer RH, Wu B. Fluorescence Imaging Methods to Investigate Translation in Single Cells. Cold Spring Harbor perspectives in biology 2018;p. a032722.

[14] Lyon KR, Aguilera LU, Morisaki T, Munsky B, Stasevich TJ. Live-cell single RNA imaging reveals bursts of translational frameshifting. bioRxiv 2018;`https://www.biorxiv.org/content/early/2018/11/24/478040`.

[15] Boersma S, Khuperkar D, Verhagen BM, Sonneveld S, Grimm JB, Lavis LD, et al. Multi-color single molecule imaging uncovers extensive heterogeneity in mRNA decoding. bioRxiv 2018;p. 477661.

[16] Lareau LF, Hite DH, Hogan GJ, Brown PO. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. Elife 2014;3:e01257.

[17] Gorgoni B, Ciandrini L, McFarland MR, Romano MC, Stansfield I. Identification of the mRNA targets of tRNA-specific regulation using genome-wide simulation of translation. Nucleic acids research 2016;44(19):9231–9244.

[18] Mahlab S, Tuller T, Linial M. Conservation of the relative tRNA composition in healthy and cancerous tissues. RNA 2012;18(4):640–652.

[19] Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic acids research 2000;28(1):292–292.

[20] Zhou HM, Brust-Mascher I, Scholey JM. Direct visualization of the movement of the monomeric axonal transport motor unc-104 along neuronal processes in livingcaenorhabditis elegans. Journal of Neuroscience 2001;21(11):3749–3755.

[21] Singh A, Hespanha JP. Approximate moment dynamics for chemically reacting systems. IEEE Transactions on Automatic Control 2011;56(2):414–418.

[22] W Gardiner C. Handbook of stochastic methods for physics, chemistry and the natural sciences, vol. 25; 1986.

[23] Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nature protocols 2013;8(8):1551.

[24] Akulich KA, Andreev DE, Terenin IM, Smirnova VV, Anisimova AS, Makeeva DS, et al. Four translation initiation pathways employed by the leaderless mRNA in eukaryotes. Scientific reports 2016;6:37905.

[25] Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. science 2009;324(5924):218–223.

[26] Axelrod D, Koppel D, Schlessinger J, Elson E, Webb WW. Mobility measurement by analysis of fluorescence photobleaching recovery kinetics. Biophysical journal 1976;16(9):1055–1069.

[27] Bonnin P, Kern N, Young NT, Stansfield I, Romano MC. Novel mRNA-specific effects of ribosome drop-off on translation rate and polysome profile. PLoS Computational Biology 2017;13(5):e1005555.

[28] Gritsenko AA, Hulsman M, Reinders MJ, de Ridder D. Unbiased quantitative models of protein translation derived from ribosome profiling data. PLoS computational biology 2015;11(8):e1004336.

[29] Duc KD, Song YS. The impact of ribosomal interference, codon usage, and exit tunnel interactions on translation elongation rate variation. PLoS genetics 2018;14(1):e1007166.

[30] Tokunaga M, Imamoto N, Sakata-Sogawa K. Highly inclined thin illumination enables clear single-molecule imaging in cells. Nature methods 2008;5(2):159.

[31] Krichevsky O, Bonnet G. Fluorescence correlation spectroscopy: the technique and its applications. Reports on Progress in Physics 2002;65(2):251.

[32] Jazani S, Sgouralis I, Pressé S. A method for single molecule tracking using a conventional single-focus confocal setup. The Journal of chemical physics 2019;150(11):114108.

[33] Coulon A, Ferguson ML, de Turris V, Palangat M, Chow CC, Larson DR. Kinetic competition during the transcription cycle results in stochastic RNA processing. Elife 2014;3.

[34] Coulon A, Larson DR. Fluctuation analysis: dissecting transcriptional kinetics with signal theory. In: Methods in enzymology, vol. 572 Elsevier; 2016.p. 159–191.

[35] Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics 1976 Dec;22(4):403–434.

[36] Hooke R, Jeeves TA. "Direct Search"Solution of Numerical and Statistical Problems. Journal of the ACM (JACM) 1961;8(2):212–229.