

RESEARCH ARTICLE

Unbiased Quantitative Models of Protein Translation Derived from Ribosome Profiling Data

Alexey A. Gritsenko^{1,2,3}, Marc Hulsman¹, Marcel J. T. Reinders^{1,3}, Dick de Ridder^{1,2,3,4*}

1 The Delft Bioinformatics Lab, Department of Intelligent Systems, Delft University of Technology, Delft, The Netherlands, **2** Platform Green Synthetic Biology, Delft, The Netherlands, **3** Kluyver Centre for Genomics of Industrial Fermentation, Delft, The Netherlands, **4** Bioinformatics Group, Wageningen University, Wageningen, The Netherlands

* dick.deridder@wur.nl

Abstract

Translation of RNA to protein is a core process for any living organism. While for some steps of this process the effect on protein production is understood, a holistic understanding of translation still remains elusive. *In silico* modelling is a promising approach for elucidating the process of protein synthesis. Although a number of computational models of the process have been proposed, their application is limited by the assumptions they make. Ribosome profiling (RP), a relatively new sequencing-based technique capable of recording snapshots of the locations of actively translating ribosomes, is a promising source of information for deriving unbiased data-driven translation models. However, quantitative analysis of RP data is challenging due to high measurement variance and the inability to discriminate between the number of ribosomes measured on a gene and their speed of translation. We propose a solution in the form of a novel multi-scale interpretation of RP data that allows for deriving models with translation dynamics extracted from the snapshots. We demonstrate the usefulness of this approach by simultaneously determining for the first time per-codon translation elongation and per-gene translation initiation rates of *Saccharomyces cerevisiae* from RP data for two versions of the Totally Asymmetric Exclusion Process (TASEP) model of translation. We do this in an unbiased fashion, by fitting the models using only RP data with a novel optimization scheme based on Monte Carlo simulation to keep the problem tractable. The fitted models match the data significantly better than existing models and their predictions show better agreement with several independent protein abundance datasets than existing models. Results additionally indicate that the tRNA pool adaptation hypothesis is incomplete, with evidence suggesting that tRNA post-transcriptional modifications and codon context may play a role in determining codon elongation rates.



OPEN ACCESS

Citation: Gritsenko AA, Hulsman M, Reinders MJT, de Ridder D (2015) Unbiased Quantitative Models of Protein Translation Derived from Ribosome Profiling Data. *PLoS Comput Biol* 11(8): e1004336. doi:10.1371/journal.pcbi.1004336

Editor: Costas D. Maranas, The Pennsylvania State University, UNITED STATES

Received: October 21, 2014

Accepted: May 10, 2015

Published: August 14, 2015

Copyright: © 2015 Gritsenko et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

Author Summary

Translation, the process of synthesizing proteins from mRNA templates, is an essential biological process in all living organisms. A better understanding of this process will have ramifications in various fields—from gene regulation, disease understanding and medicine to biotechnology and synthetic biology. Nonetheless, a holistic understanding of the processes remains elusive, making computational modelling a promising approach for studying it. However, accurate modelling of translation is challenging due to many assumptions made by such models and due to the sheer number of parameters that need to be specified. Here, we propose to fit models of translation onto ribosome profiling measurements, which record snapshots of the locations of actively translating ribosomes on mRNAs from millions of cells. We develop statistical and computational methods for fitting the Totally Asymmetric Exclusion Process (TASEP) models of translation on these measurements and verify them by deriving highly accurate translation models for the baker's yeast *Saccharomyces cerevisiae*, which outperform existing models on independent datasets. We find that fitted elongation rate parameters from the derived models deviate significantly from the widely accepted tRNA pool adaptation hypothesis.

Introduction

The process of protein synthesis is central to all living organisms. It has been actively researched for over five decades, and by now the individual steps of this process are known in great detail at the molecular and mechanistic levels [1]. Gene adaptation to the tRNA pool, mRNA secondary structure strength, codon order and local amino acid charge were independently implicated in shaping rates of protein production [2–4]. However, many disciplines would benefit from a holistic view of how these factors collectively influence translation. In particular, in biotechnology this knowledge would allow for tuning protein expression as desired with ramifications for cost-effective production of medicines and biofuels using microbes [5]. However, owing to the biological complexity of the process and the difficulty of measuring kinetic rates of the individual steps of protein synthesis, the development of computational models that would enable such applications lagged behind.

Only recently, the accumulated knowledge was integrated into several state-of-the-art models of increasing complexity. Zhang and Ignatova [6] proposed a “static” model for predicting the local speed of translation within a gene from codon-specific elongation rates derived from tRNA concentrations; their approach was extended by Reuveni *et al.* [7], who suggested using a “dynamic” model in which ribosomes initiate translation at the first codon and block each other while moving towards the end of the mRNA transcript. Siwiak and Zielenkiewicz [8] and Shah *et al.* [9] independently proposed static and dynamic full-cell models that additionally integrated the intracellular concentrations of ribosomes, mRNA and tRNA molecules, and their diffusion inside the cell in a single model. While predictions made by these models are usually in accordance with the current understanding of translation, most of their core assumptions (e.g. codon translation rates) have not been subjected to comparison against measured data.

Ribosome profiling (RP) [10, 11], a relatively new technique based on high-throughput sequencing of ribosome-protected RNA fragments (footprints), is nowadays often employed for studying translation [12–15]. It provides noisy snapshots of the locations of actively translating ribosomes attached to mRNA transcripts, thereby convolving the number of ribosomes and their speed of translation (a few stalled ribosomes can generate similar sets of footprints as

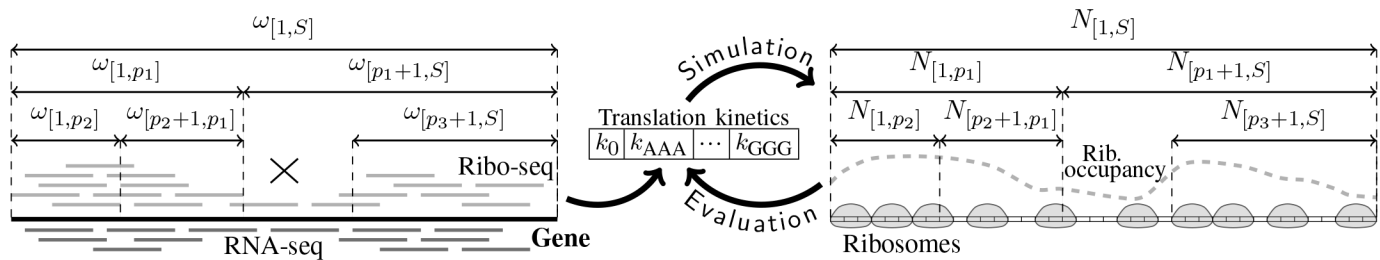


Fig 1. Schematic overview of the proposed approach for inferring translation kinetics from RP data. To obtain a segment tree representation of the RP data (left) mapped ribo-seq (light grey) and RNA-seq (dark grey) reads are assigned to nested segments of decreasing lengths (starting from segments $[1, S]$ equivalent to the full-length CDSes) while there is sufficient data. Ribosome densities ω for each segment are computed for the available replicates and are used to parameterize the log-normal distributions describing measurement error of these segments. To determine per-gene translation initiation rates k_0 and per-codon elongation rates k_{AAA}, \dots, k_{GGG} many candidate sets of translation rates are tested. For every candidate set the TASEP model of translation is simulated with the proposed rates for all genes in the *model simulation step* (right). Ribosome occupancy, i.e. the relative amount of time ribosomes spend at a particular location on the mRNA, obtained from the simulation (dashed grey) is then aggregated per segment to compute the average occupancies N , which are compared the log-normal distributions of the corresponding segments of the segment tree representation in the *model evaluation step*. Evaluation results are used by a genetic algorithm to propose new candidate sets of rates and repeat the simulation-evaluation cycle until the search for translation rates converges. To simplify notation, the gene index g is dropped for all gene-specific variables in the figure.

doi:10.1371/journal.pcbi.1004336.g001

many ribosomes involved in rapid translation). While in principle these data allow for simultaneously reasoning about ribosome counts and their local speed, such analysis is hampered by the limited understanding of the error model and biases of RP data [16]. To date RP measurements have been analyzed either at the level of full genes [8, 9] or at single codon resolution [4, 17]. While only the latter allows for analyzing the dynamics of translation, it is not clear whether codon-resolution measurements are sufficiently reliable for such quantitative analysis (see S1 Text). To overcome the measurement reliability issue several studies [18–20] performed “meta-codon” analysis by pooling observations from different occurrences of a particular codon together to produce an estimate of the codon elongation time. It is unclear, however, to what extent such estimates are affected by ribosomal interference.

We propose a set of methods for deriving full translation kinetics of an organism from RP data (see Fig 1). Our approach is conceptually similar to Ciandrini *et al.* [21], who inferred translation initiation rates of yeast genes from polysome profiling data, except that we use RP for deriving these rates and additionally determine the translation elongation rates. The method is based on a novel “segment tree” multi-scale interpretation of the RP data that captures ribosome translation dynamics along mRNAs without sacrificing reliability due to measurement noise. We use this interpretation to simultaneously extract, for the first time, per-gene translation initiation rates and per-codon translation elongation rates for the bakers yeast *Saccharomyces cerevisiae* by fitting two version of the TASEP (Totally Asymmetric Exclusion Process), a simple dynamic model of translation [22], on the segment tree estimates. To make fitting tractable, we devised a highly efficient initiation rate approximation scheme and combined it with a novel Monte Carlo simulation strategy inside an evolutionary optimization algorithm.

Fitted TASEP models match the RP data significantly better than the state-of-the-art models, and their predicted protein production rates are confirmed by several independent protein abundance (PA) datasets. In particular our models show significantly better agreement with PA than existing models when the measurements are corrected for mRNA levels, i.e. when only the effect of translation on protein levels is considered. Interestingly, the fitted codon elongation rates deviate significantly from the tRNA pool adaptation hypothesis.

Materials and Methods

Ribosome profiling data

RP data for yeast *Saccharomyces cerevisiae* strain S288C [23] containing ribosome footprint reads (ribo-seq) and fragmented mRNA reads (RNA-seq) measured in duplicate were obtained from the NCBI Short Read Archive (accession SRP028552). Reads were trimmed and mapped to the latest *S. cerevisiae* strain S288C reference genome taken from the Saccharomyces Genome Database (SGD, Cherry *et al.* [24]) in two stages, and assigned to gene coding sequences (CDSes) obtained from SGD. Aligned ribosome footprint and mRNA reads were assigned to single positions within the CDSes based on respectively their inferred A-sites or the centre position of the read (see S1 Text for details).

Measurement resolution

To obtain a high-resolution map of mRNA and ribosome density without sacrificing measurement accuracy, for each gene we construct a segment tree of density measurements from nested parts of the CDSes (Fig 1, left). By pooling reads from all segment positions, average densities per segment can be calculated more reliably than would be possible at single codon resolution (see also S1 Text), while recording these densities for nested segments of decreasing lengths allows for indirectly capturing the change in density along a transcript.

Starting from an initial segment $[l, r]$ equivalent to the complete CDS we count the number of ribo-seq reads $R_{[l,r]}$ and RNA-seq reads $M_{[l,r]}$ assigned to this segment. These counts are normalized by the total number of ribo- and RNA-seq reads aligned to all CDSes (N_R and N_M respectively) and the segment length $L_{[l,r]} = r - l + 1$ to obtain ribosome and mRNA densities $d_{[l,r]}^{\text{Ribo}} = \frac{R_{[l,r]}}{L_{[l,r]}N_R}$ and $d_{[l,r]}^{\text{mRNA}} = \frac{M_{[l,r]}}{L_{[l,r]}N_M}$ for the current segment. To obtain the sought *per transcript*

ribosome density (later referred to as density ratio) the ratio of the two measurements $\omega_{[l,r]} = \frac{d_{[l,r]}^{\text{Ribo}}}{d_{[l,r]}^{\text{mRNA}}}$ is calculated. The average segment ribosome density given by this ratio is normalized for transcript abundance and allows for directly comparing segments from different genes to each other. A cut point p is then chosen and the process is repeated recursively for segments $[l, p]$ and $[p + 1, r]$ (see Fig 1, left). The aim behind calculating $d_{[l,r]}^{\text{mRNA}}$ for each segment independently instead of estimating a single gene-specific value is to remove any local sequencing bias (presumed to be identical between RNA- and ribo-seq since very similar protocols are used for library preparation [23]) from the ratio estimates. Density measurements are computed for each replicate individually, but the same segment cut points are used in order to merge replicates later. Cut points are chosen such that the combined number of RNA- and ribo-seq reads across replicates is divided equally between the left and the right segments (see S1 Text for details).

The recursive tree construction continues while there are sufficient reads for making reliable density estimates (at least 128 reads in the two replicates summed together for RNA-seq and ribo-seq, separately; see S1 Text for details on choosing these thresholds) and segment length is large enough, $L_{[l,r]} \geq 20$ codons. The segment length cutoff aims at keeping the segments long enough to average out any measurement error due to incorrect read assignment or sequence bias. Prior to interpreting the measurements, we additionally remove a systematic density-dependent bias present in the density and ratio measurements using the available replicate information (see S1 Text).

This procedure was used to construct segment trees for 4,892 genes with a total of 60,466 nested density estimates left after removing genes classified as dubious or located on the mitochondrial chromosome.

Statistical treatment of the measurements

In order to accurately capture variance of RP data, we assume that the measured segment densities follow a log-normal distribution around the density values. A similar assumption is often made for transcriptome measurements and is justified by the observation that inter-replicate errors (i.r.e.), i.e. the ratios of replicated mRNA and ribosome density measurements, follow a log-normal distribution (S1 Fig and Ingolia *et al.* [10]). It then holds that density ratios $\omega_{[l,rj]}$ ($j \in \mathcal{J}^g$, where \mathcal{J}^g is the set of all segments of gene g) from different replicates also follow a log-normal distribution $\ln\mathcal{N}(\mu_j, \sigma_j)$ as ratios of log-normally distributed random variables—the mRNA and ribosome segment densities. Here μ_j and σ_j are used as shorthands for $\mu_{[l,rj]}$ and $\sigma_{[l,rj]}$ respectively.

To determine the parameters of this distribution we estimate μ_j for the j -th segment from the available replicated measurements as the log of their geometrical mean. Ideally, a separate shape parameter σ_j should also be estimated per segment, but, given the number of replicates, doing so would not yield reliable estimates. Instead it was chosen to group segments from all genes based on their length, and estimate shape parameters σ_k^{group} for group k from the i.r.e. of measurements from that group (see S1 Text and S2 Fig).

The proposed measurement distribution $\ln\mathcal{N}(\mu_j, \sigma_{k_j}^{\text{group}})$, where k_j denotes the length group of the j -th segment, is used throughout this paper as an error model for fitting TASEP models of translation on RP data and for comparing different models with the data.

Data interpretation and model evaluation

Computational models of translation typically provide the ability to extract steady-state codon occupancy probabilities obtained from model simulations, i.e. estimates of the chance that a particular position of an mRNA is occupied by an actively translating ribosome. Similar to the ribosome profiling measurements these occupancy profiles are determined by the local speed of translation and the number of ribosomes translating an mRNA. This allows for evaluating how well a given model matches the RP data by comparing the average segment occupancies and the segment tree ratio estimates (see Fig 1, right).

Quantitative measurements obtained via high-throughput sequencing such as the mRNA and ribosome densities (and hence their ratios) are measured in arbitrary units. Without explicit assumptions on the physiological characteristics of the analyzed organism, such as the full size of its transcriptome [8] or the number of ribosomes per cell [9], and on the efficiency of individual experimental steps, it is impossible to estimate sequencing depth of the RP measurements (i.e. the average number of reads per ribosome or the average number of reads per kilobase of transcript) and therefore impossible to express the measured values in physiologically meaningful units (e.g. number of ribosomes per transcript). Additionally, this unit mismatch complicates the comparison of modeled ribosome occupancies to the measured densities. To derive a model evaluation criterion, we first assume that an unknown scaling factor C that transforms model output into measurement data units is given, and propose a method for calculating it later.

Let n_i^g be the model-predicted ribosome occupancy at position i of gene g and $T^g = \{(\mu_j^g, \sigma_j^g) \mid j \in \mathcal{J}^g\}$ be the set of ratio distribution parameters for segments $[l_j^g, r_j^g]$. Here the upper index g denotes the gene, and for a more succinct notation we use the lower index j in place of $[l_j^g, r_j^g]$. For segment j on gene g the probability of the predicted occupancies given the segment ratio estimates can be expressed as

$$p(C, N_j^g \mid \mu_j^g, \sigma_j^g) \propto f_C(N_j^g; \mu_j^g, \sigma_j^g), \tag{1}$$

where $N_j^g \equiv \sum_{i=1}^{r_j^g} n_i^g / (r_j^g - l_j^g + 1)$ is the predicted average occupancy on segment j of gene g , and $f_C(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x + \ln C - \mu)^2}{2\sigma^2}}$ is the log-normal probability density function describing the density ratio measurement error scaled by factor $\frac{1}{C}$. This formulation is used for comparing the predicted occupancies to the estimated values in a probabilistic fashion. Assuming independence between ratio estimates of the same gene and between genes, the probability of observing all estimates, denoted by n , can be expressed as

$$p(C, n|T) \propto \prod_g \prod_{j \in J^g} f_C(N_j^g; \mu_j^g, \sigma_j^g), \tag{2}$$

In practice these calculations are more easily performed in log space and the constant factors are dropped:

$$\psi(C, n|T) = \sum_g \sum_{j \in J^g} \ln f_C(N_j^g; \mu_j^g, \sigma_j^g) \sim \sum_g \sum_{j \in J^g} \left[-\frac{1}{2(\sigma_j^g)^2} (\ln N_j^g - \mu_j^g + \ln C)^2 - \ln N_j^g \right] \tag{3}$$

We use $\psi(C, n|T)$ as the objective function for quantifying how well model-predicted ribosome occupancies match measured data.

To choose the scaling factor C , we note that it is the only free parameter of $\psi(C, n|T)$ if model output n and segment tree estimates T are given. In that case, the value of C maximizing ψ can be determined analytically:

$$\ln C = \left(\sum_{g,j \in J^g} \frac{1}{(\sigma_j^g)^2} (\mu_j^g - \ln N_j^g) \right) / \left(\sum_{g,j \in J^g} \frac{1}{(\sigma_j^g)^2} \right) \tag{4}$$

Throughout this paper, different models are evaluated at a scaling factor C maximizing their fit to the data (i.e. maximizing ψ). While the unknown true scaling factor is determined by the physiological properties of the cell, the efficiency of the experimental protocols and characteristics of the high-throughput sequencing measurements (see section “Initiation rate approximation” and [S1 Text](#)), evaluating models at the best possible scale allows for a more fair evaluation as it does not penalize models in cases when the model and the true scales mismatch.

The TASEP model of translation

TASEP (Totally Asymmetric Exclusion Process) models mRNAs g as one-dimensional lattices of length S^g and ribosomes as abstract “particles” occupying L sites corresponding to codons ([Fig 2](#)). These particles hop on (translation initiation) and off (translation termination) the lattice at the first and last sites with rates k_0^g and $k_{S^g}^g$ respectively. They only move towards the end of the lattice (hence the totally asymmetric) by hopping one site at a time with site-specific elongation rate k_i^g . Ribosomes interact with each other by “excluding” a volume of L sites that they cover on the lattice, meaning that a ribosome cannot continue to the next codon if it is already covered by another ribosome. The exact location of the active site among the L covered codons does not change the rules governing ribosome motion [[22](#)], but the choice of L may influence simulation dynamics in cases of high ribosome queueing. Typically, values $9 \leq L \leq 11$ are used [[8, 9, 16, 21](#)]; $L = 10$ was chosen for our simulations as it best matches the RP footprint size distribution [[10](#)].

TASEP captures the high-level physical interaction between ribosomes and transcripts by describing the ribosomes as travelling on the mRNAs. While in practice a number of varying translation scenarios are possible (e.g. RER-bound translation with ribosomes glued to the

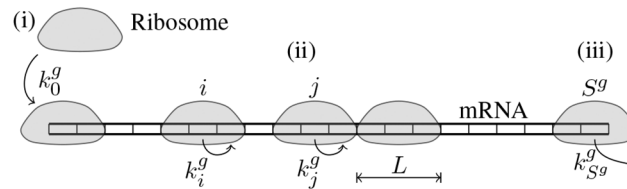


Fig 2. In TASEP mRNAs are modeled as one-dimensional lattices of S^g sites (codons) and ribosomes—as particles occupying L sites ($L = 3$ in the figure). During translation (i) ribosomes attach to the first codon with rate k_0^g if the beginning of the mRNA is not occupied by other ribosomes (initiation); (ii) ribosomes move from position i to $i + 1$ with a site-specific rate k_i^g if not blocked by another ribosome (elongation); and, finally, (iii) after reaching the last codon, they detach with rate $k_{S^g}^g$ (termination).

doi:10.1371/journal.pcbi.1004336.g002

endoplasmic reticulum and moving very slowly while the mRNA is instead pulled through the ribosomes [25]), the rich set of behaviors attainable by TASEP makes it a suitable framework for modelling translation. It requires specification of a large number of parameters, namely the gene- and site-specific elongation rates k_i^g (with the stop codon elongation rate functioning as the termination rate) and the gene-specific initiation rates k_0^g . To reduce the number of parameters we assume that the site-specific elongation rates are codon-specific and do not differ between genes. This commonly made assumption [7, 16, 21, 26] is necessary for determining model parameters from RP data as it makes the model fitting problem tractable. Depending on the experiment, either elongation rates consistent with the tRNA pool adaptation hypothesis were fixed to allow fitting the initiation rates only, or all model parameters were fit on the available data.

Monte Carlo simulations

Evaluation and fitting of the TASEP model requires an efficient way of obtaining steady-state ribosome occupancies. TASEP models allow limited analytical tractability and, to our knowledge, no analytical results for the steady-state codon occupancy probabilities are available for the general case. Additionally, existing TASEP mean-field approaches poorly approximate codon occupancies [27], a quantity of particular importance to this study, leaving stochastic simulations as the only suitable approach.

TASEP steady-state codon occupancies were obtained by simulating the model using a Monte Carlo algorithm, i.e. by randomly selecting an event (translation initiation, elongation or termination) in every simulation step and, if no other ribosomes interfere with the event, executing it with a probability proportional to its rate. To speed up simulation we developed a continuous time simulation method similar to the Gillespie algorithm [28], but based on the use of the Erlang distribution to only sample times between *state-changing events*, i.e. events that change the configuration of ribosomes attached to an mRNA.

Formally, the times between consecutive initiation or elongation events at position i are assumed to be exponentially distributed with rates k_0^g and k_i^g respectively (i.e. the corresponding model rate parameters, Fig 2). Let $o_i, i = 1, \dots, S^g$ be the current state of the simulated molecule:

$$o_i = \begin{cases} 1, & \text{codon } i \text{ is occupied by a ribosome (is at its A - site)} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

Then the time between *any* two consecutive events is also exponentially distributed with rate $k = k_0^g + \sum_{i=1}^{S^g} o_i k_i^g$ as the minimum of independent exponentially distributed random variables. Once an event occurred, the probability that it was event j is given by $p_j = o_j k_j^g / k$ (it is

assumed that ribosomes are always available to initiation translation, i.e. $o_0 = 1$). Some of the events cannot be executed due to ribosomes blocking each other and do not lead to a state change. If k_+ is the sum of rates of events leading to a state change, then the number of events between consecutive state changes, denoted as e , follows a geometric distribution with parameter $p_+ = k_+/k$ and the time Δt between state changing events follows the Erlang distribution with shape e and rate k as the sum of iid exponential random variables. The simulation proceeds by repeated random sampling of the number of events, the time between events and the event type s from the appropriate probability distributions; and updating ribosome locations in accordance to the sampled event:

$$s \sim \text{Categorical}(p_0, p_1, \dots, p_{ss}), \quad e \sim \text{Geometric}(p_+), \quad \Delta t \sim \text{Erlang}(e, k). \quad (6)$$

Simulating only state-changing events allows the simulation to progress faster, especially in cases of high ribosome queueing. The total time T_i^g spent by ribosomes at position i and the total simulation time T^g are recorded to estimate the per-transcript ribosome occupancy at this position as $n_i^g = T_i^g/T^g$, which is then used for comparing the model to RP data. Similarly the total number of translation terminations F^g is used to estimate the protein production rate $\mathcal{J}^g = F^g/T^g$.

To reach steady-state distribution of ribosomes on mRNA irrespective of the CDS length, each mRNA was simulated until 1000 translation termination events occurred. After that the model was further simulated for up to 10^7 additional steps or until the average ribosome occupancy in the segments of interest was estimated with high precision (absolute error $\epsilon < 10^{-3}$). The latter stopping criterion is based on the observation that average ribosome occupancy over a fixed segment of the mRNA can be reliably estimated before per-position occupancies can. Segment densities were first estimated after 5×10^5 simulation steps and then every 10^6 steps. Simulation was stopped if the absolute error between consecutive estimates was smaller than ϵ .

Initiation rate approximation

In addition to the elongation rates, large TASEP models require specification of hundreds gene translation initiation rates prior to simulation. Direct measurements of the initiation rates rates are unavailable and instead their values are often inferred from other sources such as ribosome profiling [8, 9] or polysome size measurements [21] data. Initiation rates estimated in such a way depend on the rates of translation elongation used in the analysis, and hence need to be optimized together with the elongation rates of the TASEP model. This leads to an explosion of the number of parameters that need to be determined, stressing the need for highly efficient initiation rate approximation strategies if the initiation and elongation rates are to be determined from the RP data simultaneously.

The problem of determining initiation rates was previously tackled by approximations neglecting ribosome queueing [8, 9], and by near-exhaustive computational search [21]. We propose a method that is a compromise between the two approaches—it allows approximating gene initiation rates for the TASEP model from RP data at a fraction of the computational cost of an exhaustive search. Briefly, we add an additional parameter \tilde{C} , the “proposed” scaling factor, to the list of model parameters that need to be estimated. This parameter is identical to the scaling factor C from Eq (4), but is used within the model to obtain biologically meaningful initiation rates. We calculated the value of \tilde{C} from the number actively translating ribosomes [29] and the number of mRNA molecules [30] per cell using a procedure proposed by Siwiak and Zielenkiewicz [8]. Given some estimate of the elongation rates and \tilde{C} we then find optimal initiation rates using a novel numerical approximation of ribosome density for TASEP models that is based on the observations of Cinandrini *et al.* [21]. This approach allows us to decouple

initiation rates from elongation rates and greatly reduces the number of model parameters that need to be fitted explicitly (next section). We used this method to efficiently (re-)approximate initiation rates of genes for each new set of elongation rates k_i^g . A full description of the approach is available in the [S1 Text](#).

Model fitting

When fitting the TASEP models, translation rates that maximize $\psi(C, n|T)$ are sought. Lacking a closed-form solution, we employed the Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES [31]) to find these rates.

We considered two different TASEP models: TASEP^{init} and TASEP^{elong}. In TASEP^{init} the elongation rates are fixed at values consistent with the tRNA pool adaptation hypothesis and initiation rates are approximated as described earlier. In the TASEP^{elong} model none of the parameters are fixed: also the codon-specific elongation rates are optimized with the CMA.

Since TASEP simulation output is invariant to scaling of translation rates, many equally good solutions exist. To constrain the search the elongation rate of codon GAA was fixed at its initial tRNA pool adaptation hypothesis value. The codon was chosen as it is present in many genes and segments ([S5 Fig](#)). Further details regarding the use of CMA can be found in the [S1 Text](#).

Despite the efficient Monte Carlo simulation and translation rate search strategies, model fitting remains a very CPU-intensive task. To speed up computations in practice, the models were fitted using hundreds of CPUs in parallel as individual genes can be simulated independently.

Because TASEP simulations of different genes are independent of each other, it may be unclear how to interpret the fitted elongation and initiation rates, as they must depend on such global biophysical quantities as the number of tRNAs or ribosomes within the cell. Nevertheless, the final simulation results are compared to whole-genome RP measurements. We can therefore expect that if our TASEP simulations agree well with RP data, the fitted translation rates used within the simulations account for the necessary biophysical parameters. Thus they should be regarded as the *effective* initiation and elongation rates that account for the relevant biophysical characteristics of the cell and growth conditions. We note that translation rates determined in such a way are condition-specific, and would likely change if fitted on a dataset obtained under different growth conditions.

Comparison to other models

To obtain a baseline for evaluating the performance of fitted TASEP models we also evaluated several existing state-of-the-art static and dynamic models of translation and compared them to each other based on their agreement with the RP data as given by [Eq \(1\)](#). SMOPT [9] and Zhang's model [6] were chosen for evaluation on the segment tree data as other state-of-the-art models, namely the Ribosome Flow Model [7] and the model from Siwiak and Zielenkiewicz [8], do not provide ribosome occupancy profiles compatible with the segment tree interpretation. The latter model was however compared to the fitted TASEP models based on several independent PA datasets.

When comparing models' predictions using independent protein abundance datasets, the "initiation frequency" P , "total amount of protein molecules produced from transcripts of particular type" B and the "total time of translation of one protein molecule from a given transcript" T from Siwiak and Zielenkiewicz [8] were respectively treated as translation initiation rate, the product of J and mRNA levels, and the inverse of J ; the average gene total elongation

time from SMOPT [9] was treated as the inverse of J ; \mathcal{P} from Ciandrini *et al.* [21] was treated as J .

Experimental setup

Since the sets of genes included in SMOPT and the segment trees differ, to facilitate comparison, all models were evaluated on a set of 3,617 genes (49,894 segments) that were in common between all models after removing very long genes to speed up TASEP simulations (31 genes longer than 2,000 codons). This set of genes was used to fit TASEP models inside a 5-fold stratified cross-validation (CV) loop over genes, in which the CV folds were chosen to balance the number of genes and segments between folds. In every step of the CV 1 fold was used for fitting (training set) and 4 folds were used for model evaluation (test set). Smaller training sets were used to reduce model fitting time. To evaluate predictions of the proposed TASEP models we also fitted them on all segment tree estimates. And to further reduce fitting time on this large dataset, codon elongation rates of the TASEP^{elong} model were set to the geometric mean of elongation rates from CV folds, and only the initiation rates were estimated from the data.

To simplify comparison of different models, we computed CV objectives for all evaluated models, including the models that did not require any parameter fitting (i.e. SMOPT and Zhang's model). While the static Zhang model does not explicitly model the translation initiation step, SMOPT and TASEP models require initiation rates to be defined for every gene in the test sets in order to calculate the CV objective. We used the original initiation rates inferred from the RP data [9, 10] for SMOPT, and approximated TASEP initiation rates using the test set segment tree measurements.

The tRNA pool adaptation hypothesis

Some of the experiments required the translation elongation rates to be defined. For those experiments we used translation elongation rates k_{AAA}, \dots, k_{GGG} consistent with the tRNA pool adaptation hypothesis, which could be seen as a statement that codons recognized by more abundant tRNAs are translated faster. The exact values for the elongation rates were defined based on the tRNA Adaptation Index (tAI [32]), which quantifies the decoding efficiency of a codon by simultaneously considering abundances of all tRNA species recognizing it and the strength of wobble base pairing between the codon and the anticodons of the isoacceptor tRNAs. The elongation rates k_{AAA}, \dots, k_{GGG} were calculated as the inverse of the codon translation times taken from the Ribosome Flow Model [33]; and translation termination rates (i.e. $k_{TAG}, k_{TAA}, k_{TGA}$) were set to 1.

Comparison to tAI and CAI

The tAI and CAI (Codon Adaptation Index [34]) are the most commonly used codon indices. They quantify respectively the extent to which a particular sequence consists of codons recognized by abundant tRNAs, and the extent to which a particular sequence consists of codons present in highly expressed (e.g. ribosomal and glycolytic) genes. These indices are often used as a proxy for translational efficiency of a gene and are employed to optimize its sequence for expression in a different host organism. Having determined elongation rates for the TASEP^{elong} model, we sought to understand whether these rates suggest a different optimization scheme than the one given by tAI or CAI.

For each codon the tAI (CAI) assigns a number—the absolute adaptiveness of that codon to the tRNA pool (codons used in highly expressed genes). To facilitate comparison between the different indices, following the definition of the CAI, we define the relative adaptiveness of a codon as its absolute adaptiveness normalized by the maximum adaptiveness among

synonymous codons. We then use the relative adaptiveness for CAI, tAI and an index based on the $TASEP^{elong}$ elongation rates (described below), when comparing optimization schemes.

We note that from the definitions of tAI [32] and elongation rates consistent with the tRNA pool hypothesis (previous section and [7]) it follows that the tAI absolute codon adaptiveness and the elongation rates are proportional to each other, and use this observation to define a codon index based on the fitted $TASEP^{elong}$ elongation rates. We define the relative adaptiveness of a codon according to $TASEP^{elong}$ as its elongation rate normalized as described above.

Other datasets

Protein abundance measurements were taken from Newman *et al.* [35] (YEPD and SD media) and Ghaemmaghami *et al.* [36]. 5'- and 3' UTR lengths were determined based on Nagalakshmi *et al.* [37] and Yassour *et al.* [38] as the mean length obtained from the two studies.

Results

Segment trees reliably capture density changes along transcripts

Segment density ratios are estimates of the average number of ribosomes engaged in translation of a given segment (measured in arbitrary units), and are expected to become more reliable if the segment length is increased. Fig 3 shows that estimates obtained for longer segments are indeed more reliable (smaller σ values) with the longest segments (rightmost group) being nearly as reliable as the full-CDS estimates from all genes (S2 Fig). We note that although group widths increase almost exponentially, potentially collecting segments with different i.r.e. in the top group, the constructed groups map very well to individual levels of the segment trees because lengths of segments with each new level are halved on average. This mapping thus provides important additional information to the segment trees about the increasing reliability of measurements that are located higher within the tree.

In this way, segment trees establish a tradeoff between measurement reliability and measurement resolution by combining the use of trustworthy estimates high in the tree (corresponding to longer segments, describing high-level gene behavior) with the use of many less reliable estimates located lower in the tree that describe the local density variation. As can be seen from the visualization of the raw data for gene YLR449W and its segment tree reconstruction in Fig 4, our multi-scale approach, that combines measurements from different scales (segment lengths), allows for implicitly capturing changes in ribosome density along transcripts, while at the same time keeping the average ribosome density across larger segments close to the observed levels. This representation also encodes uncertainty about the density ratio at a particular region of the gene, even if that region is not directly represented by a segment from the tree. For example, region (85, 104) (highlighted in the figure) is covered by 6 segments (i.e. has depth 5 within the tree) and has one of the tightest confidence intervals (CIs) in the reconstruction. At the same time region (105, 120) was not measured at the two lowest scales (has a depth 3) and its average density has to be derived from the density values of other segments and our uncertainty about them, leading to a wider CI. This example demonstrates how segment trees capture changes in ribosome density along the transcript, which are crucial for fitting translation rates and evaluating competing models.

Knowledge-based models do not fit RP data

Small standard deviations of the scaling factors and objective scores (determined using CV) of the evaluated models shown in Table 1 suggest that the (fitted) models perform consistently

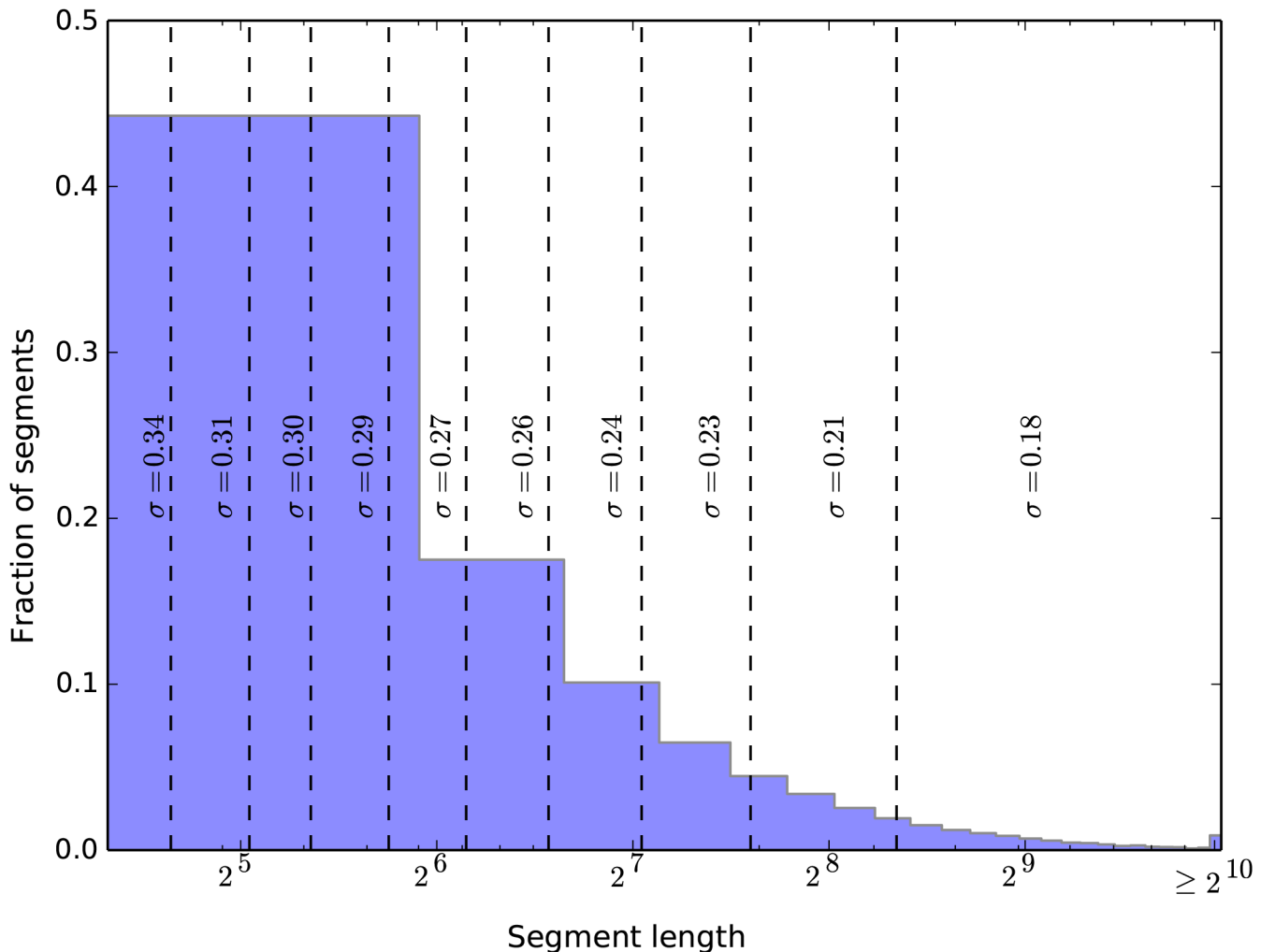


Fig 3. Segment length histogram overlaid with the shape parameters of the density ratio distributions for segment length groups (separated by dashed lines) shows that shorter segments tend to have more variable measurements. Segments were separated based on their length into 10 equal-content groups (group edges adjusted to allow for unique segment assignment), and the shape parameters σ were calculated from the inter-replicate errors of the measurements falling within each group (S1 Table).

doi:10.1371/journal.pcbi.1004336.g003

across different folds. The objective scores also show that knowledge-based models (i.e. the SMoPT and Zhang models) based on a manual choice of numerous translation-related parameters explain the ribosome density measurements significantly worse than the two models fitted on RP data. This can also be concluded from a visual inspection of the predictions made by these models for one of the genes in Fig 4C, which shows that their ribosome occupancies tend to “miss” the measured density ratios. For the Zhang model this could be explained by the absence of gene-specific initiation rates in the model, whereas SMoPT often overshoots the measured density ratios, presumably because it over-estimates initiation rates by neglecting ribosome queuing.

The TASEP^{init} model simulated with tAI-based elongation rates and fitted initiation rates achieves a significantly higher objective scores than the two state-of-the-art models. It is further improved by the TASEP^{elong} model, for which the elongation rates are additionally fit on the segment tree measurements. Fig 5 shows that superior objective function values of the fitted models translate to better predictions of the measured ribosome density (Pearson correlation

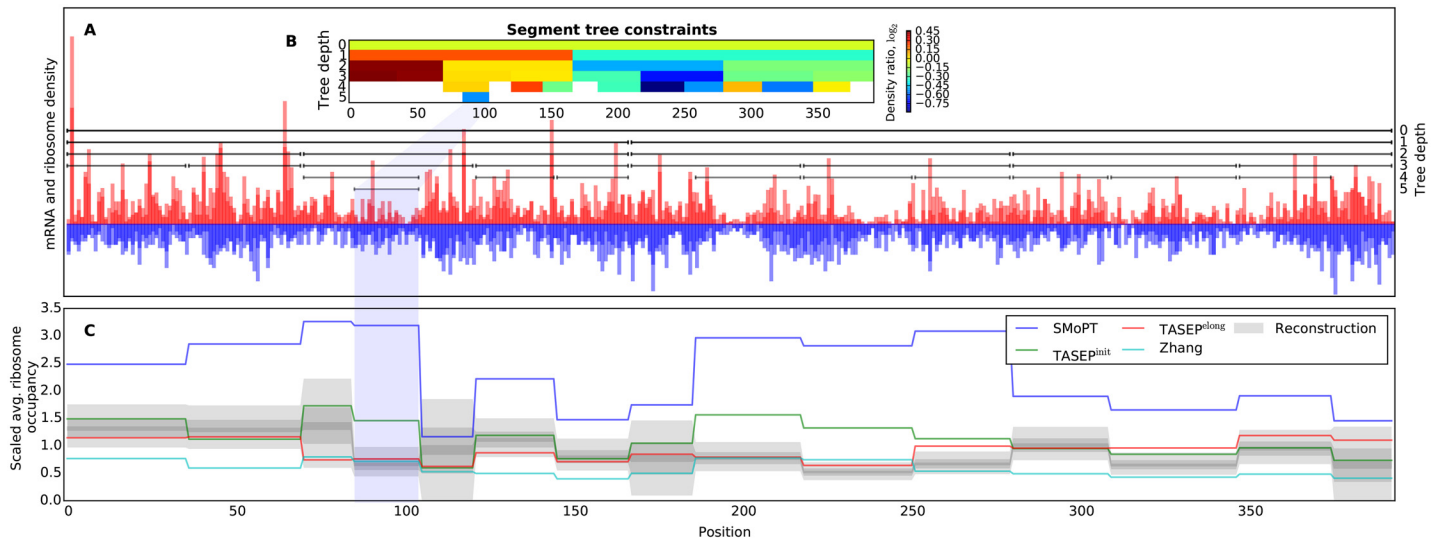


Fig 4. Ribosome profiling data, segment tree and simulated ribosome occupancy for gene YLR449W. (A) Ribo-seq (red) and mRNA-seq (blue) read density shown at single-codon resolution. Densities from the available replicates are overlaid with each other. Horizontal lines show the beginning and end of segments from the segment tree constructed for these densities. (B) Heatmap of the average density ratios recorded for each of the scales (i.e. depths) within the segment tree. (C) Reconstruction of the per-transcript ribosome density from the segment tree (gray) shown as 90%, 50% and 10% confidence intervals (shades of grey). The reconstruction was obtained by sampling from the joint probability distribution derived from the segment tree (see S1 Text). Simulated ribosome occupancy for several considered models (blue, green, red and cyan solid lines) was averaged within segments and scaled to match the data.

doi:10.1371/journal.pcbi.1004336.g004

coefficient $r = 0.77$ vs. 0.45 , $p < 10^{-293}$). Although the predictions are generally better for longer segments, improvements can be observed at all scales (see S3 Fig). While due to its relative simplicity only a weak positive correlation was expected for the Zhang model, for reasons unclear, a highly significant ($p < 10^{-293}$) negative correlation is observed (Fig 5, left). This demonstrates that current knowledge-based models are not supported by RP measurements and highlights the importance of a critical evaluation of existing translation models against independent measurements.

TASEP predictions are supported by independent datasets

Although TASEP^{init} and TASEP^{elong} outperformed existing models in the CV experiments, care has to be taken when interpreting these results as only the TASEP models were fitted directly on the segment tree measurements. We sought to obtain additional confirmation of the models' performance and to determine if they make biologically meaningful predictions. To this end we compared the protein production and translation initiation rates given by

Table 1. Objective ψ and scaling factor C for the evaluated models computed on the test folds inside a 5-fold CV loop.

Model	Fitted	ln C	Objective ψ
Zhang	No	-4.55 ± 0.00	$-600\ 286 \pm 4449$
SMoPT	No [†]	5.04 ± 0.01	$-244\ 834 \pm 2962$
TASEP ^{init}	Init.	5.40 ± 0.00	$99\ 144 \pm 2137$
TASEP ^{elong}	Yes	5.41 ± 0.02	$114\ 865 \pm 4335$

[†]—RP data Ingolia *et al.* [10] was used in the original publication to set initiation rates.

doi:10.1371/journal.pcbi.1004336.t001

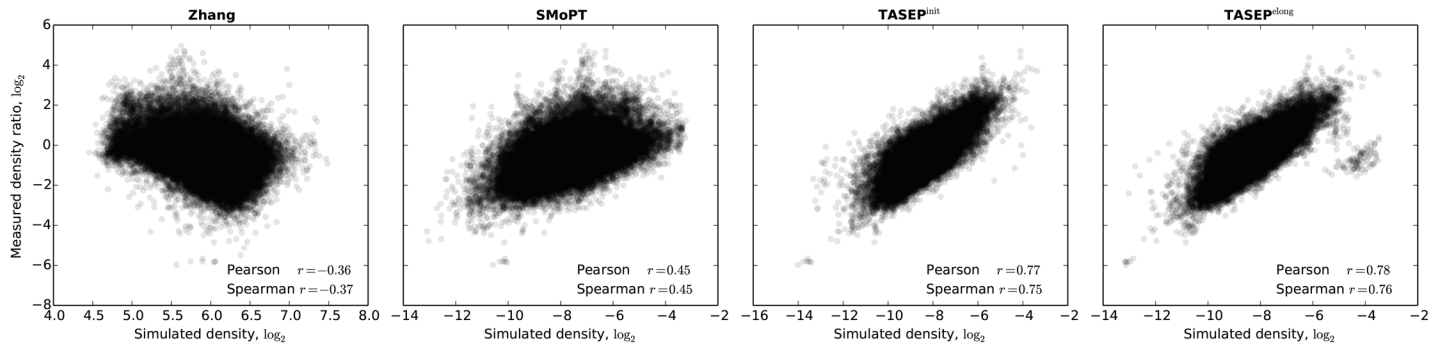


Fig 5. Measured segment density ratios $\mu_{[i,r]}$ plotted against the segment-averaged predicted ribosome occupancies for several existing and proposed models. Ribosome occupancy predictions made by the fitted models show significantly better agreement with the RP data. Reported correlations are highly significant ($p < 10^{-293}$).

doi:10.1371/journal.pcbi.1004336.g005

TASEP models fitted on all segment tree estimates to several independent large-scale PA datasets (see [Materials and Methods](#)).

Most importantly, we found that for both models the predicted protein production rates (PPRs) J positively correlate with the PA measurements ([Table 2](#)). As expected, because J describes PPR per transcript, these correlations improve when the product of J and mRNA levels ($J \times \text{mRNA}$; mRNA levels taken from the RP data) is considered. Even when both J and PAs are corrected for mRNA levels (thereby removing transcriptional regulatory influences in order to study translational regulation in isolation), the remaining (partial) correlation between J' and PA' is still significant, indicating that our TASEP models adequately capture the effects of protein translation on protein levels. These correlations are superior compared to correlations observed for state-of-the-art models ([Table 3](#)), especially when the partial correlations are considered. While strong positive partial correlations would be expected, we find these only for the fitted TASEP models. Unexpectedly low and negative partial correlations between PA' and J' for other models, together with strong correlations between PPR and mRNA levels ([Table 4](#)) suggest that existing models are overfit on transcript levels and may not accurately model

Table 2. Correlations of TASEP predictions with independent PA datasets. Spearman rank correlation coefficients r for are reported; J' is the partial correlation between J and PA given mRNA.

	TASEP ^{init}		
	Newman YEPD	Newman SD	Ghaemmaghani
Init. rate	$r = 0.56^{***}$	$r = 0.55^{***}$	$r = 0.49^{***}$
J	$r = 0.57^{***}$	$r = 0.56^{***}$	$r = 0.50^{***}$
$J \times \text{mRNA}$	$r = 0.72^{***}$	$r = 0.70^{***}$	$r = 0.63^{***}$
J'	$r = 0.52^{***}$	$r = 0.49^{***}$	$r = 0.39^{***}$
	TASEP ^{elong}		
	Newman YEPD	Newman SD	Ghaemmaghani
Init. rate	$r = 0.54^{***}$	$r = 0.53^{***}$	$r = 0.49^{***}$
J	$r = 0.56^{***}$	$r = 0.53^{***}$	$r = 0.49^{***}$
$J \times \text{mRNA}$	$r = 0.72^{***}$	$r = 0.70^{***}$	$r = 0.63^{***}$
J'	$r = 0.52^{***}$	$r = 0.48^{***}$	$r = 0.39^{***}$

*— p -value $< 10^{-5}$
 **— p -value $< 10^{-20}$
 ***— p -value $< 10^{-100}$

doi:10.1371/journal.pcbi.1004336.t002

Table 3. Correlations of predictions made by existing models with independent PA datasets. Spearman rank correlation coefficients r are reported.

Siviak and Zielenkiewicz			
	Newman YEPD	Newman SD	Ghaemmaghami
Init. rate	$r = 0.45^{**}$	$r = 0.48^{***}$	$r = 0.40^{***}$
J	$r = 0.33^{**}$	$r = 0.36^{**}$	$r = 0.37^{***}$
$J \times \text{mRNA}$	$r = 0.58^{***}$	$r = 0.54^{***}$	$r = 0.50^{***}$
J'	$r = -0.12^*$	$r = -0.07$	$r = -0.01$
SMoPT			
	Newman YEPD	Newman SD	Ghaemmaghami
Init. rate	$r = 0.45^{**}$	$r = 0.49^{***}$	$r = 0.44^{***}$
J	$r = 0.21^{**}$	$r = 0.23^{**}$	$r = 0.26^{**}$
$J \times \text{mRNA}$	$r = 0.45^{**}$	$r = 0.46^{**}$	$r = 0.46^{***}$
J'	$r = -0.26^{**}$	$r = -0.21^*$	$r = -0.13^*$
Ciandrini <i>et al.</i> [21]			
	Newman YEPD	Newman SD	Ghaemmaghami
Init. rate	$r = 0.44^{***}$	$r = 0.43^{***}$	$r = 0.43^{***}$
J	$r = 0.45^{***}$	$r = 0.44^{***}$	$r = 0.44^{***}$
$J \times \text{mRNA}$	$r = 0.57^{***}$	$r = 0.56^{***}$	$r = 0.55^{***}$
J'	$r = 0.10^*$	$r = 0.10^*$	$r = 0.14^*$

*— p -value $< 10^{-5}$
 **— p -value $< 10^{-20}$
 ***— p -value $< 10^{-100}$

doi:10.1371/journal.pcbi.1004336.t003

translation. These findings provide an independent confirmation that the TASEP models with fitted translation rates accurately capture the dynamics of the *S. cerevisiae* translation machinery.

Looking more in detail (Table 4), we find that for both models the fitted initiation rates agree well with the rates inferred by the existing full-cell models of Shah *et al.* (SMoPT), and of Siviak and Zielenkiewicz. However, we did not find the previously reported strong negative correlation between initiation rates and CDS length [9, 21]. We note that this correlation is also not supported by the model of Siviak and Zielenkiewicz. The initiation rates also exhibit a weak correlation with the 3' UTR lengths (similar correlations also found for several other models), supporting the hypothesis of more efficient translation re-initiation in genes with longer 3' UTRs.

Interestingly, we did not find the tendency for genes with short 5' UTRs to exhibit high initiation rates suggested by Shah *et al.* and supported by Ciandrini *et al.* [21] in our models or the model of Siviak and Zielenkiewicz. We also note that no relationship or a negative relationship can be observed between initiation rates and 5' UTR lengths corrected for CDS lengths can be found in most considered models. This suggests that the previously observed inverse relationship between 5' UTR lengths and initiation rates may not be indicative of a 5' UTR-mediated initiation rate regulation mechanism, but could be merely a consequence of a positive correlation between 5' UTR lengths and CDS lengths.

While correlations observed for the fitted models do not change between TASEP^{init} and TASEP^{elong} (Table 4), the latter model makes considerably better ribosome occupancy predictions. It can be seen from the example in Fig 4C that fitting the elongation rates allows the

Table 4. Comparison of TASEP predictions to existing models. Spearman rank correlation coefficients r are reported. When “corrected for” column is non-empty, partial correlations are reported.

Variable 1	Variable 2	Corrected for	Correlation coeff.	p-value
TASEP ^{init} init. rates	SMoPT init. rates		$r = 0.67$	$p < 10^{-298}$
	Siwiak and Zielenkiewicz init. rates		$r = 0.74$	$p < 10^{-298}$
	Ciandrini <i>et al.</i> init. rates		$r = 0.47$	$p < 10^{-197}$
TASEP ^{elong} init. rates	TASEP ^{init} init. rates		$r = 0.94$	$p < 10^{-298}$
	SMoPT init. rates		$r = 0.65$	$p < 10^{-298}$
	Siwiak and Zielenkiewicz init. rates		$r = 0.73$	$p < 10^{-298}$
	Ciandrini <i>et al.</i> init. rates		$r = 0.46$	$p < 10^{-182}$
	TASEP ^{init} init. rates		$r = -0.07$	$p < 10^{-4}$
CDS lengths	TASEP ^{elong} init. rates		$r = -0.05$	$p < 10^{-2}$
	SMoPT init. rates		$r = -0.52$	$p < 10^{-257}$
	Siwiak and Zielenkiewicz init. rates		$r = -0.02$	$p > 10^{-1}$
	Ciandrini <i>et al.</i> init. rates		$r = -0.65$	$p < 10^{-298}$
	TASEP ^{init} init. rates		$r = -0.01$	$p > 10^{-1}$
5' UTR lengths	TASEP ^{elong} init. rates		$r = -0.02$	$p > 10^{-1}$
	SMoPT init. rates		$r = -0.06$	$p < 10^{-3}$
	Siwiak and Zielenkiewicz init. rates		$r = 0.00$	$p > 10^{-1}$
	Ciandrini <i>et al.</i> init. rates		$r = -0.09$	$p < 10^{-10}$
	TASEP ^{init} init. rates	CDS lengths	$r = 0.00$	$p > 10^{-1}$
	TASEP ^{elong} init. rates	CDS lengths	$r = -0.01$	$p > 10^{-1}$
	SMoPT init. rates	CDS lengths	$r = 0.03$	$p > 10^{-1}$
	Siwiak and Zielenkiewicz init. rates	CDS lengths	$r = 0.03$	$p < 10^{-1}$
	Ciandrini <i>et al.</i> init. rates	CDS lengths	$r = -0.06$	$p < 10^{-3}$
	TASEP ^{init} init. rates		$r = 0.04$	$p < 10^{-2}$
3' UTR lengths	TASEP ^{elong} init. rates		$r = 0.04$	$p < 10^{-1}$
	SMoPT init. rates		$r = 0.06$	$p < 10^{-3}$
	Siwiak and Zielenkiewicz init. rates		$r = 0.07$	$p < 10^{-5}$
	Ciandrini <i>et al.</i> init. rates		$r = 0.03$	$p < 10^{-1}$
	TASEP ^{init} init. rates	CDS lengths	$r = 0.04$	$p < 10^{-1}$
	TASEP ^{elong} init. rates	CDS lengths	$r = 0.04$	$p < 10^{-1}$
	SMoPT init. rates	CDS lengths	$r = 0.07$	$p < 10^{-4}$
	Siwiak and Zielenkiewicz init. rates	CDS lengths	$r = 0.08$	$p < 10^{-6}$
	Ciandrini <i>et al.</i> init. rates	CDS lengths	$r = 0.02$	$p > 10^{-1}$
	TASEP ^{init} init. rates		$r = 0.36$	$p < 10^{-115}$
	TASEP ^{elong} init. rates		$r = 0.33$	$p < 10^{-93}$
	SMoPT init. rates		$r = 0.58$	$p < 10^{-298}$
	Siwiak and Zielenkiewicz init. rates		$r = 0.33$	$p < 10^{-117}$
	Ciandrini <i>et al.</i> init. rates		$r = 0.62$	$p < 10^{-298}$
	TASEP ^{init} J		$r = 0.34$	$p < 10^{-97}$
TASEP ^{elong} J		$r = 0.37$	$p < 10^{-115}$	
SMoPT J		$r = 0.65$	$p < 10^{-298}$	
Siwiak and Zielenkiewicz J		$r = 0.69$	$p < 10^{-298}$	
Ciandrini <i>et al.</i> J		$r = 0.63$	$p < 10^{-298}$	
mRNA levels	Newman YEPD PA		$r = 0.58$	$p < 10^{-209}$
	Newman SD PA		$r = 0.57$	$p < 10^{-194}$
	Ghaemmaghami PA		$r = 0.54$	$p < 10^{-273}$
CDS lengths	Newman YEPD PA		$r = -0.13$	$p < 10^{-10}$

(Continued)

Table 4. (Continued)

Variable 1	Variable 2	Corrected for	Correlation coeff.	p-value
	Newman SD PA		$r = -0.14$	$p < 10^{-12}$
	Ghaemmaghami PA		$r = -0.16$	$p < 10^{-22}$
	Newman YEPD PA	mRNA	$r = 0.32$	$p < 10^{-60}$
	Newman SD PA	mRNA	$r = 0.28$	$p < 10^{-42}$
	Ghaemmaghami PA	mRNA	$r = 0.21$	$p < 10^{-36}$
	mRNA		$r = -0.53$	$p < 10^{-298}$
	5' UTR lengths		$r = 0.14$	$p < 10^{-20}$
	3' UTR lengths		$r = -0.03$	$p < 10^{-1}$

doi:10.1371/journal.pcbi.1004336.t004

segment-averaged ribosome occupancy of TASEP^{elong} to follow the reconstructed density considerably better than any of other model.

Fitted elongation rates are not explained by adaptation to tRNA levels alone

Since the TASEP^{elong} model achieves a significantly better fit to the RP data compared to TASEP^{init} with tAI-based rates (Table 1), having fitted its elongation rates on different CV folds, we sought to interpret the obtained values and their variance. We first, however, confirmed that elongation rates determined from different RP datasets agree qualitatively with each other by fitting a new TASEP^{elong} model on the dataset of Ingolia *et al.* [10] and comparing its translation rates to the original model (see S1 Text).

It can be seen from Fig 6 that despite the generally large SDs, for many codons the elongation rates fitted in different folds of the CV are spread compactly around codon-specific values. This is clearly visible for codons with smaller SDs (green and blue), for which similar rates were found in different folds. Nonetheless the rate SDs differ considerably between codons. While the majority of the fitted elongation rates are consistently different from tAI-based rates, only for 13 codons this difference is statistically significant (single sample *t*-test for population mean difference, $p < 0.05$; Fig 6, S2 Table): GAC, TTG, CCA, CAA, GCC, GGT, GAT, TTT, CAG,

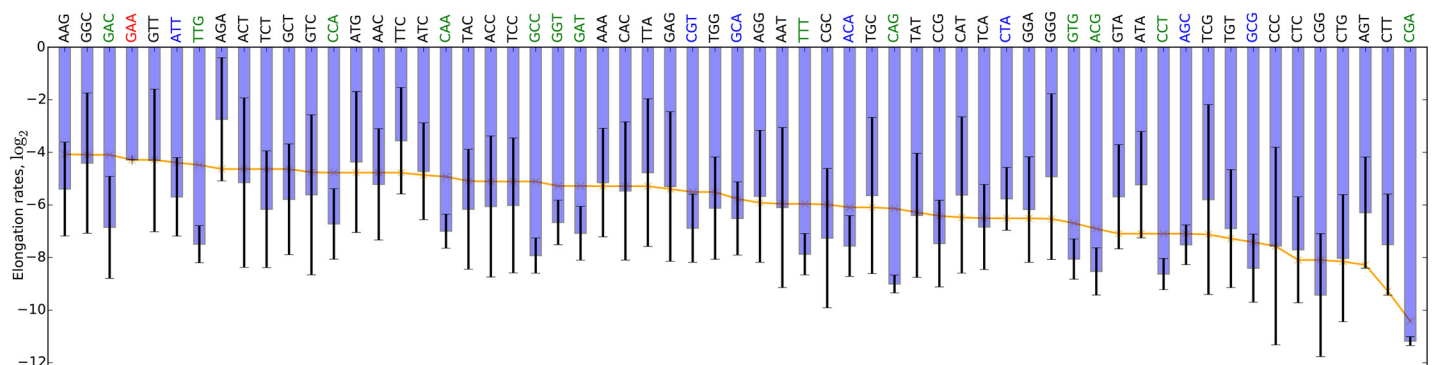


Fig 6. Mean and SD of the codon elongation rates fitted in different CV folds, compared to the tAI-based rates. For many codons elongation rates (depicted as mean and SD, blue bars) are reproducible across CV folds. This becomes evident for codons with smaller SDs (blue labels, $\sigma < 1.5$), and codons whose elongation rates are significantly different from the tAI-based rates (green labels; *t*-test, $p < 0.05$). tAI-based rates (orange line) are plotted as a reference. The rate of codon GAA (red label) was not optimized. Stop codons were excluded from the figure as their fitted termination rates remained very close to the original values of 1.

doi:10.1371/journal.pcbi.1004336.g006

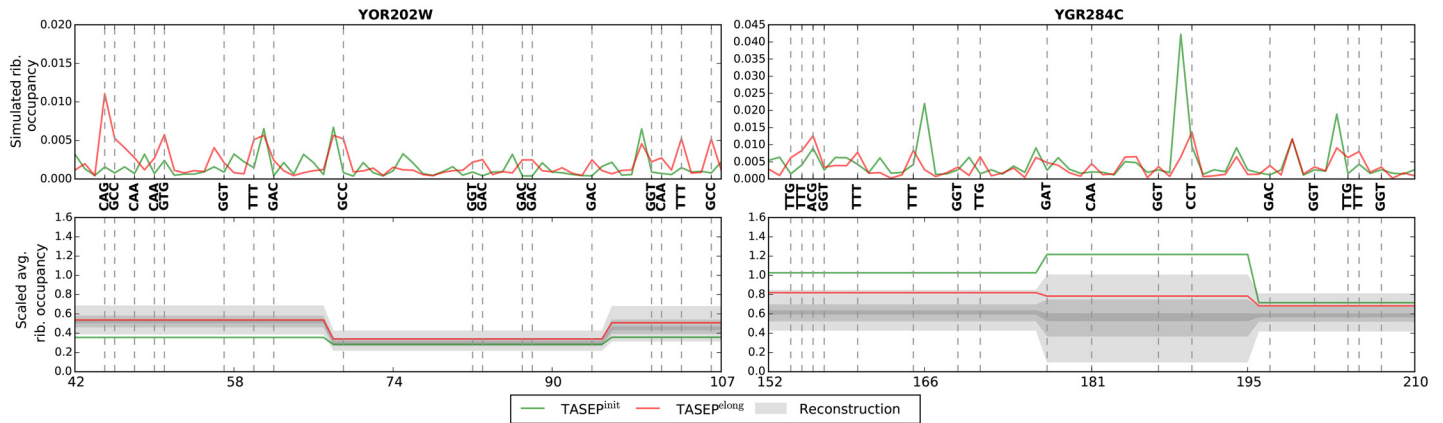


Fig 7. Ribosome density reconstruction (gray, bottom panels) and simulated ribosome occupancy (top) for selected regions of genes YOR202W (left panels) and YGR284C (right panels) plotted for the TASEP^{init} (green) and TASEP^{elong} (red) models. Presence of codons with significantly different elongation rates (vertical dashed lines) increases simulated ribosome occupancy. Higher increase can be observed for segments containing more such codons. This is clearly seen for gene YOR202W (left) with similar initiation rates in the TASEP^{init} and TASEP^{elong} models (0.24×10^{-4} and 0.22×10^{-4} respectively), for which the predicted occupancy only increases when fitted elongation rates are used. For most genes, such as YGR284C (right) this increase in density is compensated by reducing the initiation rate (from 0.72×10^{-4} to 0.36×10^{-4}), which leads to an overall better agreement between simulated ribosome occupancy and the segment tree measurements (bottom right). To keep the visualization comprehensible, only selected regions of genes YOR202W and YGR284C were used. However, the described trends also hold for the remainder of these genes and for other genes.

doi:10.1371/journal.pcbi.1004336.g007

GTG, ACG, CCT and CGA. Although these differences between the tAI-based and fitted elongation rates are challenging to explain, their presence suggests that additional unknown factors are shaping these rates.

Having identified differences in elongation rates between the TASEP^{init} and TASEP^{elong} models, we sought to understand their effect on models' predictions. As could be expected from the similar correlations in Table 4 and Fig 5, the two models make very similar PPR and ribosome density predictions (S4 Fig). However, ribosome density predicted by the TASEP^{elong} model with fitted elongation rates agrees better with RP measurements. To understand the cause of this improvement we looked for genes whose fit to the RP data improved when fitted elongation rates were used. These genes can be classified into two groups: (i) genes that have a very similar initiation rate in both models (Fig 7, left); and (ii) genes that have a considerably lower initiation rate in the TASEP^{elong} model (Fig 7, right). Because all 13 codons with significantly different elongation rates were predicted to be slower, their presence in CDSes generally leads to higher predicted ribosome occupancy, especially if the genes initiation rate remains unchanged. For genes from the first group, such as YOR202W shown on the left panel of Fig 7, this already results in a more accurate ribosome occupancy prediction. For most other genes, the second group, this increase in codon elongation times yields ribosome occupancy that is too high under the current initiation rate. For these genes (e.g. YGR284C on the right panel of Fig 7) a smaller fitted initiation rate is required to reduce ribosome occupancy that would otherwise be too high due to the effects of slow codons and high ribosomal flux (due to high initiation rate). Together these effects allow the model to better match the ribosome density changes along the transcript.

Significance of the fitted elongation rates for codon optimization

Codon optimization, the process of substituting codons with synonymous alternatives that are elongated faster, thus contributing to the overall protein production rate, is routinely used to improve protein expression [39, 40]. Nonetheless, it remains a controversial tool because the

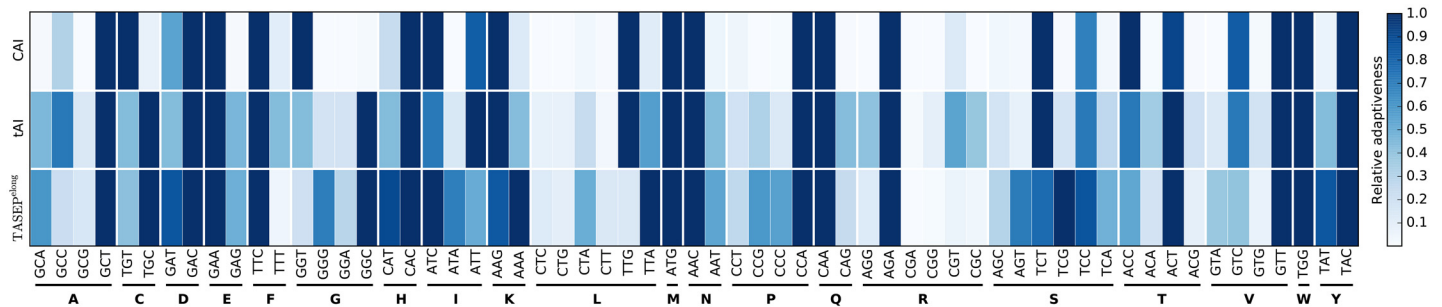


Fig 8. tAI and CAI compared to a measure of codon adaptation derived from the fitted TASEP^{elong} elongation rates. Relative adaptiveness of codons grouped by their corresponding amino acids (columns) plotted for the three measures of codon adaptation (rows) shows that the considered measures often agree on the optimal codon. In particular, the tAI and TASEP^{elong} measure agree on the optimal codons for all but 4 amino acids (I, K, L and S).

doi:10.1371/journal.pcbi.1004336.g008

same optimization techniques can lead to contradicting results when applied to different proteins [41]. Here we compare our fitted elongation rates to codon optimality estimated by the commonly used tAI [32] and CAI [34] indices.

We considered the relative adaptiveness of a codon (see [Materials and Methods](#)) given by the CAI, the tAI and the fitted elongation rates of the TASEP^{elong} model. Fig 8 shows that the three measures of codon adaptation often agree on the optimal codon for a particular amino acid (relative adaptiveness of 1.0, dark blue), which further demonstrates that our findings are in line with the earlier work. In particular, despite significant differences between the fitted elongation rates and elongation rates given by the tRNA adaptation hypothesis, the two sets agree on optimal codons for all but four amino acids. Only for isoleucine, leucine, lysine and serine the TASEP^{elong} model suggests codons ATC, AAA, TTA and TCG instead of ATT, AAG, TTG and TCT respectively. An interesting observation is that the bottom row in Fig 8 is much more blue than the top ones, suggesting codon optimization is less black-and-white than suggested by tAI and in particular CAI, meaning that many more codons are “reasonably good”, i.e. there may be less to gain by codon optimization than thought before. This observation is also corroborated by Leavitt and Alper [42], who noted that the level of control achievable in yeast through codon optimization is considerably smaller than what can be achieved through transcriptional regulation.

Translation initiation limits protein production

It is still unclear whether translation of endogenous yeast genes is limited by initiation or elongation [43, 44]. To test whether translation is limited by the initiation rates or by the elongation rates we artificially increased the initiation rate of each gene from the TASEP^{elong} model by 10%. To obtain robust results the experiment was repeated 5 times with different random initializations and the average increase in PPRs was calculated for every gene.

Fig 9 shows the relative differences in PPRs for all genes. In almost all cases (except 7 genes) the PPR increased substantially (relative difference > 0.02) when increasing the initiation rate, supporting the hypothesis that under exponential growth in the rich medium translation in *S. cerevisiae* operates in an initiation-limited regime. This also explains why fitting the codon elongation rates in TASEP^{elong} did not improve the PA correlations compared to the TASEP^{init} model. Elongation-limited production for these genes can be explained by the very high initiation rates predicted for them, which shift the rate-limiting step from translation initiation to translation elongation. Interestingly, groups of genes that had a low, medium and high PPR increase are enriched for several biological functions (FDR < 0.05, Fig 9). Notably, genes in the high increase group are involved in negative regulation of various biosynthetic and metabolic

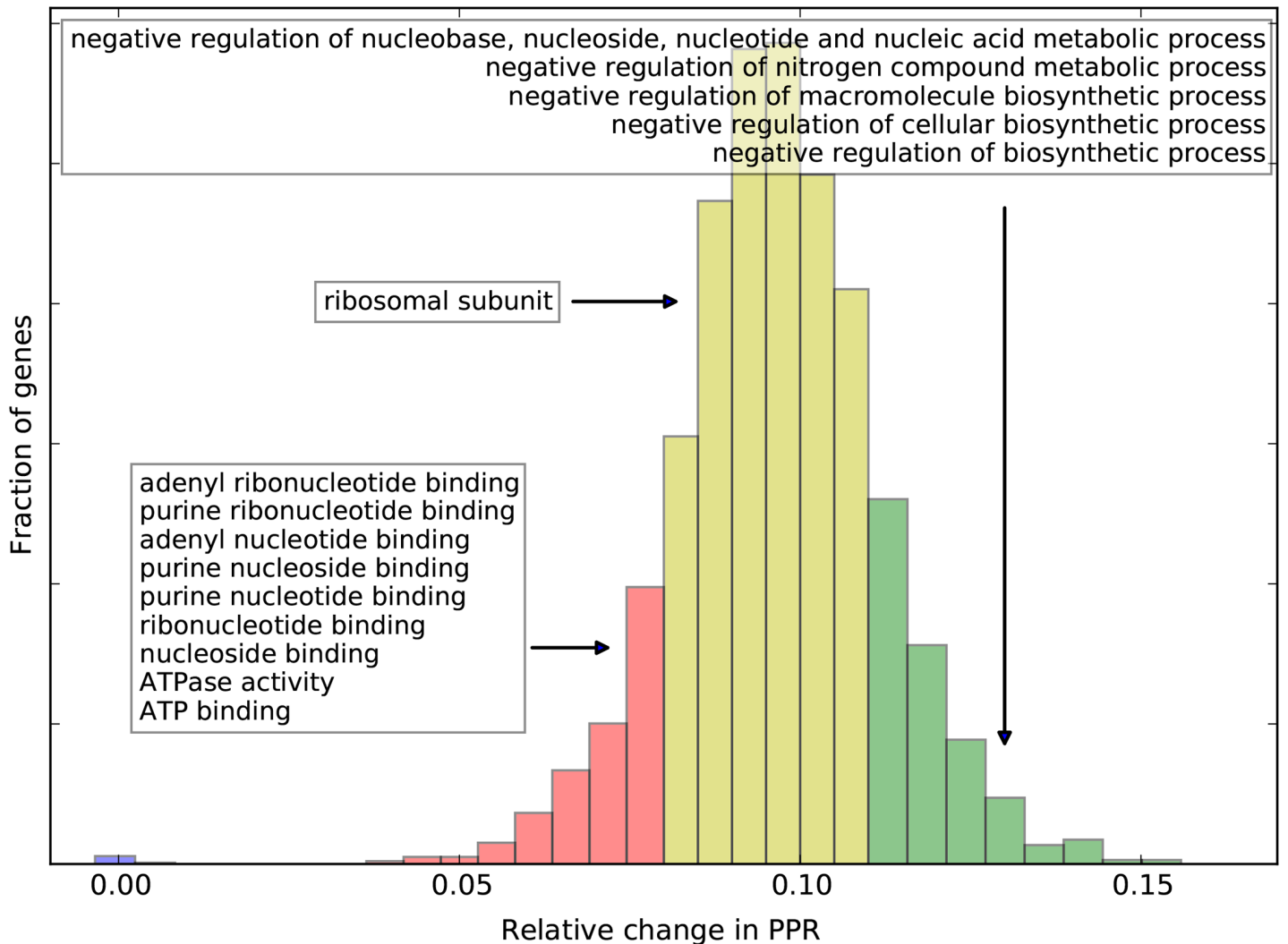


Fig 9. Distribution of the relative changes in PPR after a 10% increase in initiation rates shows that translation initiation is the rate limiting step for the protein production for most *S. cerevisiae* genes from the considered RP dataset. Groups of genes with low (≥ 0.02 and < 0.08 , red), medium (≥ 0.08 and < 0.11 , yellow) and high (≥ 0.11 , green) increase in PPRs are enriched for several biological functions (white boxes in the figure, FDR < 0.05).

doi:10.1371/journal.pcbi.1004336.g009

processes. This suggests that yeast cells may have evolved to rapidly “switch on” negative regulation by keeping a buffer of the required mRNA transcripts that are efficiently translated only once there is demand.

Discussion

For the first time, we described an approach that derives complete translation kinetics of an organism from ribosome profiling data and used it to simultaneously infer the translation elongation, translation initiation and protein production rates all together without neglecting the effects of ribosomal interference. We applied our methodology to the ribosome and RNA sequencing data of the baker’s yeast *Saccharomyces cerevisiae*. The fitted yeast translation models agree considerably better with independent protein abundance datasets than existing

models. In particular, our TASEP models are the only ones that maintain strong correlations with protein abundance after removing the effect of transcriptional regulation.

While translation initiation rates provided by the models are similar to rates from other studies, we did not find the previously reported negative correlation between initiation rates and CDS lengths. The observed negative correlations between PA and CDS length, which one would expect to see as a result of this correlation, can alternatively be explained by transcriptional regulation, i.e. the strong negative correlation between mRNA levels and CDS lengths (Table 4). An alternative explanation can be offered by a mechanism driven by amino acid chain elongation rather than translation initiation. For example, abortive translation or the degradation of misfolded proteins [45], since the chance of producing a misfolded protein is expected to increase with protein length.

We also found that translation elongation rates deviate considerably from the widely accepted tRNA pool adaptation hypothesis, for 13 codons significantly so. Differences in elongation rates of these codons between the tRNA pool adaptation hypothesis and TASEP^{elong} may be partially explained by nucleotide modifications of their respective tRNAs, which are known to modify the specificity and efficiency of messenger decoding [46]. As such, some of these 13 codons were shown to be affected by post-transcriptional nucleotide modifications of tRNAs in different organisms [47]. We speculate that for these codons the concentration of (un)modified tRNAs, rather than the total tRNA concentration, plays a non-negligible role in determining their elongation rates [18]. An additional factor that possibly contributes to the observed deviation from the tRNA pool adaptation hypothesis is its implicit assumption that different tRNA genes from the same family contribute equally to determining the rate of translation. This assumption should be revisited in light of the recent finding of Bloom-Ackermann *et al.* [48] that the contributions of different gene copies from the same tRNA family to the tRNA pool and cellular fitness are far from equal.

In our experiments we found that SDs of elongation rates from different CV folds differ markedly between codons. In order for the elongation rates to be specified with high precision by the RP data, small changes in the rates must lead to detectable differences in ribosome density. However, in light of our finding that yeast translation is initiation-limited and the observation of Bloom-Ackermann *et al.* [48] that *S. cerevisiae* is robust to deletions of tRNA genes, especially in rich medium used to produce the ribosome profiling measurements analyzed here, it is unlikely that in the considered physiological conditions the elongation rates exert a strong enough effect on ribosome density to allow the RP data to specify elongation rates with high precision. We speculate that found SDs reflect the robustness of the yeast translation system w.r.t. the codon translation rates, with the system being more sensitive to changes in rates of those codons that have smaller SDs. In this case, yeast translation appears to be robust to fold changes in codon translation rates and, consequently, to the aminoacyl-tRNA availability that these rates are thought to be determined by [44].

Alternatively, the SDs may reflect the extent to which codon translation rates change between CV folds due to codon context, i.e. the local sequence around a codon which may alter its elongation rate (see S1 Text, translation rate reproducibility analysis). It is unlikely that the TASEP model captures the full complexity of the translation process by assuming that codon elongation rates are determined solely by the codon identity, and not also by the sequence surrounding the codons as was previously suggested [2, 3]. Such a constraint limits the models ability to capture the underlying translation dynamics and may bias it towards fitting different rates on different sets of genes (e.g. CV folds) with varying codon contexts, thereby inflating the SDs. The observed variation in fitted elongation rates puts forward codon context as a factor that may significantly modulate the baseline elongation rates.

Using our models we found that under exponential growth in rich medium translation initiation appears to be the main limiting factor of protein production of endogenous genes in *Saccharomyces cerevisiae*, with protein production being limited by initiation rates for all but 7 genes with very high initiation rates. These findings suggest that rational design of 5' UTRs involved in translation initiation [49, 50] may be a more promising avenue for achieving protein overexpression than the routinely used codon optimization techniques. It is likely, however, that further overexpression could be achieved using codon optimization. Because once the gene is put under the translational control of an efficient 5'-UTR, which is usually the case in heterologous gene expression, translation elongation is expected to become a rate-limiting factor. In such cases we recommend performing codon optimization using the fitted TASEP^{long} elongation rates, which, while mostly agreeing with existing techniques, also demonstrate several differences.

Although we found that translation initiation appears to be the main factor limiting protein production in yeast under exponential growth in rich medium, it is possible that different mechanisms are dominant in other organisms. For example, Li *et al.* [51] and Guimaraes *et al.* [52] discuss greater contribution of protein elongation respectively by anti-Shine-Dalgarno sequences and codon usage in *E. coli*. Our method could be applied to ribosome profiling data of other organisms to delineate the relative contribution of initiation and elongation.

All translation models proposed to date, including TASEP^{init} and TASEP^{elong}, assume that translation elongation rates are not influenced by *codon context*, i.e. the sequence around a particular codon, although various factors affecting the speed of elongation have been suggested [2–4]. Variation in fitted elongation rates and the highly varying codon translation times recently observed by Dana and Tuller [53] suggest that codon context may play a more compelling role in determining translation rates than previously thought. Fortunately investigations of codon context are becoming feasible thanks to the growing adoption of ribosome profiling as a standard technique for studying translation. With the increasing amount of ribosome profiling measurements, data-driven approaches, such as the one described here, will become instrumental for delineating the effects of multiple competing translation mechanisms, for generating new hypothesis, and for constructing predictive models for use in other fields. These goals can be achieved by incorporating alternative translation mechanisms as sequence- and position-specific effects altering the codon elongation rates.

Supporting Information

S1 Text. Contains extended methods and supplementary results.

(PDF)

S1 Fig. Histograms of the log₂ inter-replicate errors (ratios of replicated measurements) of reliable ribosome and mRNA density measurements show that the full-CDS and segment tree density estimates follow comparable log-normal distributions. Distributions fitted into data (solid lines) are centered around zero, but their SDs differ.

(TIF)

S2 Fig. Histograms of the log₂ inter-replicate errors of reliable density ratio measurements show similar error profiles in full-CDS and segment tree estimates. The group shape parameters of the i.r.e. and the density ratio distributions are related as $\sigma_k^{\text{group}} = \frac{1}{\sqrt{2}} \sigma_k^{\text{i.r.e.}}$.

(TIF)

S3 Fig. Measured segment density ratios $\mu_{[l_p, r_j]}$ plotted against the segment-averaged predicted ribosome occupancies for segments of varying size and for several existing and proposed models. TASEP^{init} and TASEP^{elong} significantly improve over existing models for all

segment length groups.
(TIF)

S4 Fig. Agreement between the PPR (left) and gene-level average ribosome occupancy (right) predictions made by TASEP^{init} and TASEP^{long} models.
(TIF)

S5 Fig. Presence of codons in gene and segment sequences from the segment tree. Translation rate of codon GAA (red) was fixed in elongation rate fitting experiments as it is present in many genes and segments.
(TIF)

S6 Fig. Histogram of the running times (average over 3 replicates) of the TASEP model simulations for genes in the evaluation set. tAI-based elongation rates and initiation rates of 1.0 were used in the simulations.
(TIF)

S1 Table. Shape parameters of the density ratio distributions for segments grouped by length. Left (inclusive) and right (exclusive) edges give the range of segment lengths of a given group.
(PDF)

S2 Table. Mean and SD of the codon elongation rates fitted on different CV folds of the evaluation set. *p*-values of the single sample *t*-test are calculated to check whether the observed rates are significantly different from the tAI-based rates. All rates are given in log₂ space. Codons are colored as in the main text.
(XLS)

S3 Table. Detailed results of the GO term functional enrichment analysis.
(XLS)

S1 Dataset. Translation initiation and protein production rates for the derived models.
(CSV)

Acknowledgments

This work was supported by the research programme of the Kluyver Centre for Genomics of Industrial Fermentation, a subsidiary of the Netherlands Genomics Initiative (NGI) and was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

Author Contributions

Conceived and designed the experiments: AAG DdR. Analyzed the data: AAG. Wrote the paper: AAG MH MJTR DdR. Designed simulation algorithm: AAG MH.

References

- Schmeing TM, Ramakrishnan V (2009) What recent ribosome structures have revealed about the mechanism of translation. *Nature* 461: 1234–1242. PMID: [19838167](#)
- Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, et al. (2010) A role for codon order in translation dynamics. *Cell* 141: 355–367. doi: [10.1016/j.cell.2010.02.036](#) PMID: [20403329](#)
- Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, et al. (2011) Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol* 12: R110. doi: [10.1186/gb-2011-12-11-r110](#) PMID: [22050731](#)
- Charneski CA, Hurst LD (2013) Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. *PLoS biology* 11: e1001508. doi: [10.1371/journal.pbio.1001508](#) PMID: [23554576](#)

5. Keasling JD (2010) Manufacturing molecules through metabolic engineering. *Science* 330: 1355–1358. doi: [10.1126/science.1193990](https://doi.org/10.1126/science.1193990) PMID: [21127247](https://pubmed.ncbi.nlm.nih.gov/21127247/)
6. Zhang G, Ignatova Z (2009) Generic algorithm to predict the speed of translational elongation: implications for protein biogenesis. *PLoS One* 4: e5036. doi: [10.1371/journal.pone.0005036](https://doi.org/10.1371/journal.pone.0005036) PMID: [19343177](https://pubmed.ncbi.nlm.nih.gov/19343177/)
7. Reuveni S, Meilijson I, Kupiec M, Ruppin E, Tuller T (2011) Genome-scale analysis of translation elongation with a ribosome flow model. *PLoS computational biology* 7: e1002127. doi: [10.1371/journal.pcbi.1002127](https://doi.org/10.1371/journal.pcbi.1002127) PMID: [21909250](https://pubmed.ncbi.nlm.nih.gov/21909250/)
8. Siwiak M, Zielonkiewicz P (2010) A comprehensive, quantitative, and genome-wide model of translation. *PLoS computational biology* 6: e1000865. doi: [10.1371/journal.pcbi.1000865](https://doi.org/10.1371/journal.pcbi.1000865) PMID: [20686685](https://pubmed.ncbi.nlm.nih.gov/20686685/)
9. Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB (2013) Rate-Limiting Steps in Yeast Protein Translation. *Cell* 153: 1589–1601. doi: [10.1016/j.cell.2013.05.049](https://doi.org/10.1016/j.cell.2013.05.049) PMID: [23791185](https://pubmed.ncbi.nlm.nih.gov/23791185/)
10. Ingolia NT, Ghaemmaghami SA, Newman JRS, Weissman JS (2009) Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* 324: 218–223. doi: [10.1126/science.1168978](https://doi.org/10.1126/science.1168978) PMID: [19213877](https://pubmed.ncbi.nlm.nih.gov/19213877/)
11. Ingolia NT (2014) Ribosome profiling: new views of translation, from single codons to genome scale. *Nature Reviews Genetics* 15: 205–213. doi: [10.1038/nrg3645](https://doi.org/10.1038/nrg3645) PMID: [24468696](https://pubmed.ncbi.nlm.nih.gov/24468696/)
12. Ingolia NT, Lareau LF, Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147: 789–802. doi: [10.1016/j.cell.2011.10.002](https://doi.org/10.1016/j.cell.2011.10.002) PMID: [22056041](https://pubmed.ncbi.nlm.nih.gov/22056041/)
13. Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, et al. (2012) High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* 335: 552–557. doi: [10.1126/science.1215110](https://doi.org/10.1126/science.1215110) PMID: [22194413](https://pubmed.ncbi.nlm.nih.gov/22194413/)
14. Gerashchenko MV, Lobanov AV, Gladyshev VN (2012) Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proceedings of the National Academy of Sciences* 109: 17394–17399. doi: [10.1073/pnas.1120799109](https://doi.org/10.1073/pnas.1120799109)
15. Artieri CG, Fraser HB (2014) Evolution at two levels of gene expression in yeast. *Genome Research* 24: 411–421. doi: [10.1101/gr.165522.113](https://doi.org/10.1101/gr.165522.113) PMID: [24318729](https://pubmed.ncbi.nlm.nih.gov/24318729/)
16. Dana A, Tuller T (2012) Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. *PLoS computational biology* 8: e1002755. doi: [10.1371/journal.pcbi.1002755](https://doi.org/10.1371/journal.pcbi.1002755) PMID: [23133360](https://pubmed.ncbi.nlm.nih.gov/23133360/)
17. Qian W, Yang JR, Pearson NM, Maclean C, Zhang J (2012) Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS genetics* 8: e1002603. doi: [10.1371/journal.pgen.1002603](https://doi.org/10.1371/journal.pgen.1002603) PMID: [22479199](https://pubmed.ncbi.nlm.nih.gov/22479199/)
18. Zinshteyn B, Gilbert WV (2013) Loss of a conserved tRNA anticodon modification perturbs cellular signaling. *PLoS genetics* 9: e1003675. doi: [10.1371/journal.pgen.1003675](https://doi.org/10.1371/journal.pgen.1003675) PMID: [23935536](https://pubmed.ncbi.nlm.nih.gov/23935536/)
19. Gardin J, Yeasmin R, Yurovsky A, Cai Y, Skiena S, et al. (2014) Measurement of average decoding rates of the 61 sense codons *in vivo*. *eLife* 3: e03735. doi: [10.7554/eLife.03735](https://doi.org/10.7554/eLife.03735)
20. Lareau LF, Hite DH, Hogan GJ, Brown PO (2014) Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *eLife* 3: 301257. doi: [10.7554/eLife.01257](https://doi.org/10.7554/eLife.01257)
21. Ciandrini L, Stansfield I, Romano MC (2013) Ribosome traffic on mRNAs maps to gene ontology: genome-wide quantification of translation initiation rates and polysome size regulation. *PLoS computational biology* 9: e1002866. doi: [10.1371/journal.pcbi.1002866](https://doi.org/10.1371/journal.pcbi.1002866) PMID: [23382661](https://pubmed.ncbi.nlm.nih.gov/23382661/)
22. MacDonald CT, Gibbs JH, Pipkin AC (1968) Kinetics of biopolymerization on nucleic acid templates. *Biopolymers* 6: 1–25. doi: [10.1002/bip.1968.360060102](https://doi.org/10.1002/bip.1968.360060102) PMID: [5641411](https://pubmed.ncbi.nlm.nih.gov/5641411/)
23. McManus CJ, May GE, Spealman P, Shteyman A (2014) Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome research* 24: 422–430. doi: [10.1101/gr.164996.113](https://doi.org/10.1101/gr.164996.113) PMID: [24318730](https://pubmed.ncbi.nlm.nih.gov/24318730/)
24. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, et al. (2011) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic acids research: gkr1029*.
25. Shibata Y, Voeltz GK, Rapoport TA (2006) Rough sheets and smooth tubules. *Cell* 126: 435–439. doi: [10.1016/j.cell.2006.07.019](https://doi.org/10.1016/j.cell.2006.07.019) PMID: [16901774](https://pubmed.ncbi.nlm.nih.gov/16901774/)
26. Racle J, Picard F, Girbal L, Coccagn-Bousquet M, Hatzimanikatis V (2013) A genome-scale integration and analysis of *Lactococcus lactis* translation data. *PLoS computational biology* 9: e1003240. doi: [10.1371/journal.pcbi.1003240](https://doi.org/10.1371/journal.pcbi.1003240) PMID: [24130467](https://pubmed.ncbi.nlm.nih.gov/24130467/)
27. Shaw LB, Sethna JP, Lee KH (2004) Mean-field approaches to the totally asymmetric exclusion process with quenched disorder and large particles. *Physical Review E* 70: 021901. doi: [10.1103/PhysRevE.70.021901](https://doi.org/10.1103/PhysRevE.70.021901)

28. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry* 81: 2340–2361. doi: [10.1021/j100540a008](https://doi.org/10.1021/j100540a008)
29. Zenklusen D, Larson DR, Singer RH (2008) Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature structural & molecular biology* 15: 1263–1271. doi: [10.1038/nsmb.1514](https://doi.org/10.1038/nsmb.1514)
30. Miura F, Kawaguchi N, Yoshida M, Uematsu C, Kito K, et al. (2008) Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. *BMC genomics* 9: 574. doi: [10.1186/1471-2164-9-574](https://doi.org/10.1186/1471-2164-9-574) PMID: [19040753](https://pubmed.ncbi.nlm.nih.gov/19040753/)
31. Hansen N, Müller SD, Koumoutsakos P (2003) Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Computation* 11: 1–18. doi: [10.1162/106365603321828970](https://doi.org/10.1162/106365603321828970) PMID: [12804094](https://pubmed.ncbi.nlm.nih.gov/12804094/)
32. dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic acids research* 32: 5036–5044. doi: [10.1093/nar/gkh834](https://doi.org/10.1093/nar/gkh834) PMID: [15448185](https://pubmed.ncbi.nlm.nih.gov/15448185/)
33. Zur H, Tuller T (2012) RFMapp: ribosome flow model application. *Bioinformatics* 28: 1663–1664. doi: [10.1093/bioinformatics/bts185](https://doi.org/10.1093/bioinformatics/bts185) PMID: [22495755](https://pubmed.ncbi.nlm.nih.gov/22495755/)
34. Sharp PM, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research* 15: 1281–1295. doi: [10.1093/nar/15.3.1281](https://doi.org/10.1093/nar/15.3.1281) PMID: [3547335](https://pubmed.ncbi.nlm.nih.gov/3547335/)
35. Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, et al. (2006) Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* 441: 840–846. doi: [10.1038/nature04785](https://doi.org/10.1038/nature04785) PMID: [16699522](https://pubmed.ncbi.nlm.nih.gov/16699522/)
36. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, et al. (2003) Global analysis of protein expression in yeast. *Nature* 425: 737–741. doi: [10.1038/nature02046](https://doi.org/10.1038/nature02046) PMID: [14562106](https://pubmed.ncbi.nlm.nih.gov/14562106/)
37. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320: 1344–1349. doi: [10.1126/science.1158441](https://doi.org/10.1126/science.1158441) PMID: [18451266](https://pubmed.ncbi.nlm.nih.gov/18451266/)
38. Yassour M, Kaplan T, Fraser HB, Levin JZ, Pfiffner J, et al. (2009) *Ab initio* construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proceedings of the National Academy of Sciences* 106: 3264–3269. doi: [10.1073/pnas.0812841106](https://doi.org/10.1073/pnas.0812841106)
39. Gustafsson C, Govindarajan S, Minshull J, Komar A (2004) Codon bias and heterologous protein expression. *TRENDS in Biotechnology* 22. doi: [10.1016/j.tibtech.2004.04.006](https://doi.org/10.1016/j.tibtech.2004.04.006) PMID: [15245907](https://pubmed.ncbi.nlm.nih.gov/15245907/)
40. Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, et al. (2009) Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS one* 4: e7002. doi: [10.1371/journal.pone.0007002](https://doi.org/10.1371/journal.pone.0007002) PMID: [19759823](https://pubmed.ncbi.nlm.nih.gov/19759823/)
41. Lanza AM, Curran KA, Rey LG, Alper HS (2014) A condition-specific codon optimization approach for improved heterologous gene expression in *Saccharomyces cerevisiae*. *BMC systems biology* 8: 33. doi: [10.1186/1752-0509-8-33](https://doi.org/10.1186/1752-0509-8-33) PMID: [24636000](https://pubmed.ncbi.nlm.nih.gov/24636000/)
42. Leavitt JM, Alper HS (2015) Advances and current limitations in transcript-level control of gene expression. *Current opinion in biotechnology* 34: 98–104. doi: [10.1016/j.copbio.2014.12.015](https://doi.org/10.1016/j.copbio.2014.12.015)
43. Plotkin JB, Kudla G (2010) Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics* 12: 32–42. doi: [10.1038/nrg2899](https://doi.org/10.1038/nrg2899) PMID: [21102527](https://pubmed.ncbi.nlm.nih.gov/21102527/)
44. Gingold H, Pilpel Y (2011) Determinants of translation efficiency and accuracy. *Molecular systems biology* 7. doi: [10.1038/msb.2011.14](https://doi.org/10.1038/msb.2011.14) PMID: [21487400](https://pubmed.ncbi.nlm.nih.gov/21487400/)
45. Kraft C, Peter M, Hofmann K (2010) Selective autophagy: ubiquitin-mediated recognition and beyond. *Nature cell biology* 12: 836–841. doi: [10.1038/ncb0910-836](https://doi.org/10.1038/ncb0910-836) PMID: [20811356](https://pubmed.ncbi.nlm.nih.gov/20811356/)
46. Gustilo EM, Vendeix FA, Agris PF (2008) tRNA's modifications bring order to gene expression. *Current opinion in microbiology* 11: 134–140. doi: [10.1016/j.mib.2008.02.003](https://doi.org/10.1016/j.mib.2008.02.003) PMID: [18378185](https://pubmed.ncbi.nlm.nih.gov/18378185/)
47. Agris PF (2004) Decoding the genome: a modified view. *Nucleic acids research* 32: 223–238. doi: [10.1093/nar/gkh185](https://doi.org/10.1093/nar/gkh185) PMID: [14715921](https://pubmed.ncbi.nlm.nih.gov/14715921/)
48. Bloom-Ackermann Z, Navon S, Gingold H, Towers R, Pilpel Y, et al. (2014) A comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool. *PLoS genetics* 10: e1004084. doi: [10.1371/journal.pgen.1004084](https://doi.org/10.1371/journal.pgen.1004084) PMID: [24453985](https://pubmed.ncbi.nlm.nih.gov/24453985/)
49. Salis HM, Mirsky EA, Voigt CA (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nature biotechnology* 27: 946–950. doi: [10.1038/nbt.1568](https://doi.org/10.1038/nbt.1568) PMID: [19801975](https://pubmed.ncbi.nlm.nih.gov/19801975/)
50. Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, et al. (2013) Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proceedings of the National Academy of Sciences* 110: E2792–E2801. doi: [10.1073/pnas.1222534110](https://doi.org/10.1073/pnas.1222534110)
51. Li GW, Oh E, Weissman JS (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484: 538–541. doi: [10.1038/nature10965](https://doi.org/10.1038/nature10965) PMID: [22456704](https://pubmed.ncbi.nlm.nih.gov/22456704/)

52. Guimaraes JC, Rocha M, Arkin AP (2014) Transcript level and sequence determinants of protein abundance and noise in *Escherichia coli*. *Nucleic acids research* 42: 4791–4799. doi: [10.1093/nar/gku126](https://doi.org/10.1093/nar/gku126) PMID: [24510099](https://pubmed.ncbi.nlm.nih.gov/24510099/)
53. Dana A, Tuller T (2014) The effect of tRNA levels on decoding times of mRNA codons. *Nucleic acids research* 42: 9171–9181. doi: [10.1093/nar/gku646](https://doi.org/10.1093/nar/gku646) PMID: [25056313](https://pubmed.ncbi.nlm.nih.gov/25056313/)