Heterogeneous recruitment abilities to RNA polymerases generate nonlinear scaling of gene expression level with cell volume

Jie Lin
Center for Quantitative Biology, Peking University, Beijing, China and Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China

Qirun Wang
Center for Quantitative Biology, Peking University, Beijing, China
(Dated: April 8, 2021)

While most genes' expression levels are proportional to cell volumes, some genes exhibit nonlinear scaling between their expression levels and cell volume. Therefore, their mRNA and protein concentrations change as the cell volume increases, which often have crucial biological functions such as cell-cycle regulation. However, the biophysical mechanism underlying the nonlinear scaling between gene expression and cell volume is still unclear. In this work, we show that the nonlinear scaling is, in fact, a direct consequence of heterogeneous recruitment abilities of promoters to RNA polymerases based on a gene expression model at the whole-cell level. Those genes with weaker (stronger) recruitment abilities compared with the average ability spontaneously exhibit superlinear (sublinear) scaling with cell volume. Analysis of the promoter sequences and the nonlinear scaling of *Saccharomyces cerevisiae*'s mRNA levels shows that motifs associated with transcription regulation are indeed enriched in genes exhibiting nonlinear scaling, in concert with our model.

Homeostasis of gene expression level in a dynamically growing cell volume is widely observed across the domains of biology, namely, the concentrations of mRNAs and proteins of most genes are approximately constant [1–10]. Since the cell volume usually grows exponentially [9, 11, 12], this implies exponential growth of mRNA and protein copy numbers as well, and a linear scaling between the gene expression level and cell volume exists. Theoretical models have shown that the linear scaling between gene expression level and exponentially growing cell volume is a consequence of the limiting nature of gene expression machinery: RNA polymerases (RNAPs) and ribosomes [13], in agreement with experimental observations [10, 14–16]. However, along with the constant concentrations of mRNAs and proteins of most genes, there is a subset of genes exhibiting nonconstant concentrations as the cell volume increases. These genes are often crucial for cell-cycle and size regulation. They allow cells to sense their sizes based on the concentrations of proteins that have different scaling behaviors with cell volume, such as Whi5 and Cln3 in *Saccharomyces cerevisiae* [17–21]. Other examples include DNA-binding proteins such as histone proteins, whose expression levels scale with the total DNA amount instead of cell volume [22]. A fundamental question then arises: if a linear scaling between gene expression level and cell volume are by default for most genes, how can cells achieve nonlinear scaling for a subset of genes with cell volume in the meantime? In this paper, we show that the superlinear and sublinear scaling of gene expression level is, in fact, a direct consequence of the heterogeneous recruitment abilities of promoters to RNA polymerases. Given a unimodal distribution of recruitment abilities, those genes with their promoters' recruitment abilities below (above) the average spontaneously exhibit superlinear (sublinear) scaling with cell volume, while the majority of genes with recruitment abilities near the average exhibit approximately linear scaling.

In the following, we first introduce a coarse-grained gene expression model at the whole-cell level in which the promoters of all genes have the same recruitment abilities to RNAPs, which is analytically solvable. Then we consider a scenario in which all genes except a small subset of genes have the same recruitment abilities and show that the expression levels of those special genes can exhibit superlinear or sublinear scaling with cell volume depending on their relative magnitudes of recruitment abilities compared with the majority of genes. Then we extend the simplified model to allow a continuous distribution of recruitment abilities and show that our simplified model can quantitatively capture this more realistic scenario. Genes with recruitment abilities below the average value naturally exhibit superlinear (sublinear) scaling with cell volume, and active regulation is not necessary for nonlinear scaling. Finally, to verify our theoretical predictions, we analyze the nonlinear scaling of mRNA numbers *vs.* cell volume of *S cerevisiae* using the data from Ref. [21]. Our model predicts a positive correlation between the mRNA production rates and nonlinear scaling degrees among genes, and experimental data confirm this. We further analyze the promoter sequences of all genes with measured nonlinear scaling degrees and find that special motifs for transcription factor binding are indeed enriched in the promoters of genes exhibiting nonlinear scaling, in concert with our theoretical predictions.

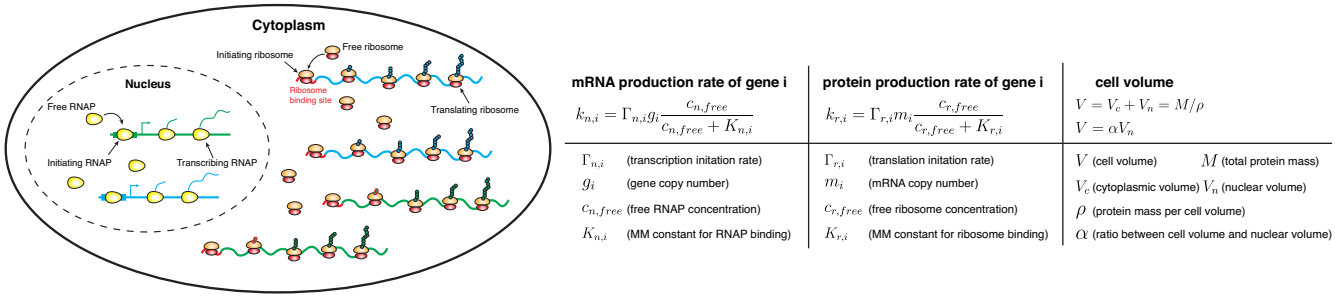

FIG. 1. A coarse-grained model of gene expression at the whole-cell level. Inside the nucleus, genes compete for free RNAPs to bind to their promoters and for simplicity we only show two genes in this schematic. The binding probability of free RNAPs to the promoters depends on the free RNAP concentration and the recruitment abilities of the promoters. In the cytoplasm, mRNAs compete for free ribosomes to bind to their ribosome binding sites and the binding probability depends on the free ribosome concentrations and the recruitment abilities of the mRNAs to ribosomes. The cell volume V, which includes the cytoplasmic volume (V_c) and nuclear volume (V_n), is proportional to the total protein mass M. The ratio between the nuclear volume and cell volume is constant.

RESULTS

Model of gene expression at the whole-cell level

We consider a coarse-grained model of gene expression. The mRNA production rate $k_{n,i}$ of one particular gene labeled by index i is proportional to its gene copy number (g_i) and the probability for its promoters to be bound by RNAPs ($P_{b,i}$),

$$k_{n,i} = \Gamma_{n,i} g_i P_{b,i}, \tag{1}$$

where

$$P_{b,i} = \frac{c_{n,free}}{c_{n,free} + K_{n,i}} \tag{2}$$

Here $\Gamma_{n,i}$ is the initiation rate of transcription of gene i: the rate that a promoter-bound RNAP starts transcribing the gene and producing mRNA. We assume the probability for one promoter to be bound by an RNAP follows the Michaelis-Menten equation where $c_{n,free}$ is the concentration of free RNAPs in the nucleus [23] (see the schematic in Figure 1). $K_{n,i}$ is the Michaelis-Menten (MM) constant which is inversely proportional to the binding rate of RNAPs on the promoters (see detailed derivations in Supplementary Information SI A). In the following, we use $1/K_{n,i}$ as a metric of the recruitment abilities of genes to RNAPs: a larger $K_{n,i}$ represents a smaller recruitment ability. Given the mRNA production rate, the mRNA copy number m_i changes as

$$\frac{dm_i}{dt} = k_{n,i} - \frac{m_i}{\tau_{m,i}} \tag{3}$$

where $\tau_{m,i}$ is the mRNA lifetime. Note that we mainly discuss eukaryotic cells in this work in which the transcription and translation processes are spatially separate, and nonspecific binding of RNAPs to DNA are mostly irrelevant [24, 25]. As we also mainly discuss the transcription of mRNA, RNAP here refers to RNAP II. Real transcription processes in eukaryotic cells are complex and may involve transcription factors, enhancers, and TATA-binding proteins [26, 27]. Their effects are coarse-grained into the Michaelis-Menten constant within our model, e.g., a transcription factor that increases the expression level of one gene is equivalent to reducing the Michaelis-Menten constant in the RNAP binding probability P_b.

Because the typical time scale for an RNAP to finish transcribing a gene is around one minute [28], much shorter than the typical doubling times, we assume that the dynamics of RNAPs along each gene is in the steady state. Therefore, the outgoing flux of RNAPs that finish transcribing a gene ($v_n n_i / L_i$) must be equal to the initiation rate of transcription ($P_{b,i}\Gamma_{n,i}$) where n_i is the number of transcribing RNAPs on gene i. Here L_i is the length of the gene i in the number of codons and v_n is the elongation speed of RNAP, which we approximate as a constant. This leads to the expression of n_i as $n_i = P_{b,i}\Lambda_{n,i}$ where $\Lambda_{n,i} = \Gamma_{n,i} L_i / v_n$ is the maximum possible number of transcribing RNAPs on gene i.

A similar model as transcription also applies to the translational process (Figure 1). The protein production rate of one particular gene is proportional to its corresponding mRNA number (m_i) and the translation initiation rate ($\Gamma_{r,i}$):

$$k_{r,i} = \Gamma_{r,i} m_i \frac{c_{r,free}}{c_{r,free} + K_{r,i}}. \tag{4}$$

Here $c_{r,free}$ is the concentration of free ribosomes in the cytoplasm and $K_{r,i}$ is the Michaelis-Menten constant of ribosome binding on the mRNAs (see an alternative formulation of translation model in Ref. [29]). In principle, the recruitment abilities of mRNAs to ribosomes and the initiation rates can also be heterogeneous, but our main results of this work are insensitive to these assumptions, as we show in the following. In the following analysis, we assume that most proteins are non-degradable, and our

results are equally valid for degradable proteins.

The total protein mass of a cell is $M = \sum_i p_i L_i$, in the number of amino acids where p_i is the protein copy number of gene i. The total protein mass is known to be proportional to the cell volume [30, 31]. Therefore, we assume a constant ratio between the total protein mass and cell volume, $V = M/\rho$, where ρ is the density in the unit of amino acid number/μm^3. We further decompose the cell volume into the cytoplasmic volume (V_c) and nuclear volume (V_n) so that $V = V_c + V_n$. We assume that the nuclear volume is proportional to the total cell volume such that $V = \alpha V_n$, supported by experimental observations [32].

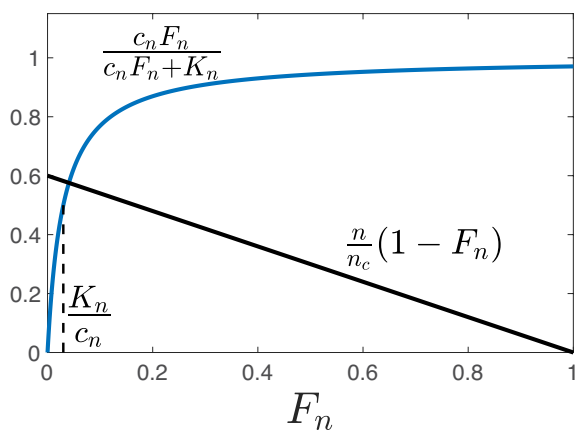

FIG. 2. Based on the conservation of total number of RNAPs, a self-consistent equation, Eq. (5), is derived for the case of homogeneous recruitment abilities to RNAPs. Assuming $K_n/c_n \ll 1$, the fraction of free RNAPs F_n that solves Eq. (5) must be much smaller than 1 if $n < n_c$.

A simplified model in which all genes share the same recruitment ability

In the following, we consider a simplified scenario assuming the promoters of all genes have the same recruitment ability to RNA polymerases so that $K_{n,i} = K_n$ for all i, which is analytically solvable (we will relax this assumption later). Using the conservation of the total number of RNAPs, including free RNAPs, initiating RNAPs, and transcribing RNAPs, we obtained the self-consistent equation to determine the fraction of free RNAPs in all RNAPs:

$$\sum_i g_i(1 + \Lambda_{n,i}) \frac{c_n F_n}{c_n F_n + K_n} = n - nF_n. \tag{5}$$

Here n is the total number of RNAPs, and F_n is the fraction of free RNAPs. c_n is the concentration of total RNAPs in the nucleus. Let us now consider a thought experiment in which all RNAPs suddenly become free. One would expect that these free RNAPs will have a strong tendency to rebind to the promoters; otherwise, a large fraction of RNAPs will be idle, which is clearly inefficient in normal cellular physiological states. Therefore, we assume that $K_n \ll c_n$ and it is straightforward to find that in this case, the F_n solving Eq. (5) must be much smaller than 1 if $n < n_c$ where $n_c = \sum_i g_i(1+\Lambda_{n,i})$ (Figure 2), which is the maximum of number of RNAPs the entire genome can hold. In this scenario, almost all RNAPs are either bound to the promoter or transcribing ($F_n \ll 1$). Taking $F_n = 0$ in Eq. (5), we obtain the binding probability as $P_b = n/n_c$ and using Eq. (1), the mRNA production rate becomes proportional to the number of RNAPs,

$$k_{n,i} = \Gamma_{n,i} \frac{g_i}{\sum_j g_j(1 + \Lambda_{n,j})} n. \tag{6}$$

This linear scaling between the mRNA production rate and the number of RNAPs [13], is the root of the linear scaling between mRNA level and cell volume, as we explain in the following.

Considering the translational process, we find that when the gene copy numbers are fixed, the mass fractions of proteins in the entire proteome are also approximately constant, which is valid regardless of whether the translation initiation rates and the recruitment abilities of mRNAs to ribosomes are heterogeneous or not among genes (SI B). Given constant mass fractions of proteins, the total number of RNAPs is proportional to the total protein mass and is also proportional to the cell volume. Therefore, the mRNA production rate is proportional to the cell volume as well, and this scenario of gene expression has been called Phase 1 of gene expression [13], which is the typical state of cells in normal conditions [6, 14, 33] and the main focus of this work. Since the mRNA lifetimes are typically much shorter than the doubling time [28], the mRNA numbers can be approximated as the product of mRNA production rates and mRNA lifetimes, $m_i \approx k_{n,i}\tau_{m,i}$, therefore proportional to the cell volume as well.

When the number of RNAPs exceeds the threshold value n_c, the linear scaling between mRNA numbers and cell volume breakdowns for all genes and the growth mode of cell volume also deviates from exponential growth (Phase 2). If the ribosome number exceeds some threshold value, the cell volume eventually grows linearly, which has been observed in budding yeast [33] (Phase 3, see Ref. [13] and detailed derivations in the SI B). We also discuss the effects of nonspecific binding of RNAPs on the transition between Phase 1 and Phase 2 in SI C and show that the condition of Phase 1 becomes more stringent in the presence of nonspecific binding.

A more realistic model in which genes can have different recruitment abilities

We now consider a more realistic scenario in which the recruitment abilities to RNAPs of different genes can be different. We start from a simple scenario in which all genes have the same recruitment ability $1/K_n$ except one special gene i has a recruitment ability $1/K_{n,i}$. Since the contribution of the particular gene to the global allocation of RNAPs is negligible, the probability for the promoters of most genes to be bound is still $P_b = c_n F_n/(c_n F_n + K_n) \approx n/n_c$ from which we find that the free RNAP concentration $c_n F_n$ as a function of n/n_c. The mRNA production rate for the particular gene i therefore becomes

$$k_{n,i} = \Gamma_{n,i} g_i \frac{K_n n}{K_{n,i} n_c - (K_{n,i} - K_n) n}. \tag{7}$$

As n gets closer to n_c (with $F_n \ll 1$ still satisfied), we find that if $K_{n,i} > K_n$ ($K_{n,i} < K_n$) the mRNA production rate of gene i exhibits a superlinear (sublinear) dependence on the RNAP number. Because the RNAP number is proportional to cell volume, this nonlinear relation is also valid between the mRNA production rates and cell volume, which leads to the nonlinear scaling between the mRNA copy numbers and cell volume. Furthermore, if the corresponding proteins have short lifetimes, their copy numbers will be proportional to the mRNA production rates as well. More importantly, we find that the nonlinear scaling in mRNA copy numbers also propagates to non-degradable proteins, suggesting that the nonlinear scalings in protein copy numbers are insensitive to their lifetimes (Methods).

In the following, we discuss a more realistic scenario that is a continuous distribution of $K_{n,i}$ to reflect the promoter structures' heterogeneity. In this case, we propose that the nonlinear scaling, Eq. (11), is still approximately valid for any gene if K_n is replaced by $\langle K_{n,i} \rangle$, the average value of $K_{n,i}$ among all genes with some appropriate weights (later we show that the appropriate average is weighted by the protein mass fractions). Therefore, those genes with $K_{n,i}$ considerably larger (smaller) than $\langle K_{n,i} \rangle$ exhibit superlinear (sublinear) scaling, while genes with $K_{n,i}$ near the average exhibit approximately linear scaling. To quantify the nonlinear degree between the expression level and cell volume of each gene, we introduce two parameters β and α respectively for mRNAs and non-degradable proteins (see detailed derivations and discussions of degradable proteins in Methods) so that their scalings with cell volume become

$$\widetilde{m}_i(t) = \widetilde{V}(t) \frac{1 + \beta_i}{1 + \beta_i \widetilde{V}(t)}, \tag{8}$$

$$\Delta \widetilde{p}_i(t) = C_i \ln(1 + \alpha_i \Delta \widetilde{V}(t)). \tag{9}$$

Here the mRNA and protein copy numbers are measured relative to their values at time $t = 0$, e.g., $\widetilde{m}_i(t) = m_i(t)/m_i(0)$. $\Delta \widetilde{V}(t) = \widetilde{V}(t) - 1$, and $\Delta \widetilde{p}_i(t) = \widetilde{p}_i(t) - 1$. C_i is a constant depending on the particular gene. Positive (negative) β_i and α_i represent sublinear (superlinear) scaling behaviors. They are related to the recruitment abilities of their corresponding genes to RNAPs (see the full expressions in Methods) and when $K_{n,i}$ is close to the mean Michaelis-Menten constant $\langle K_{n,i} \rangle$, we find that $\beta_i \propto \langle K_{n,i} \rangle - K_{n,i}$, $\alpha_i \propto \langle K_{n,i} \rangle - K_{n,i}$.

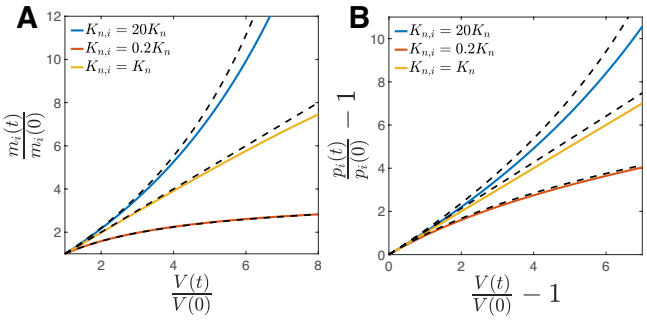

FIG. 3. (A) We consider two special genes with the Michaelis-Menten constants $K_{n,i} = 20K_n$, $K_{n,i} = 0.2K_n$. The majority of genes have $K_{n,i} = K_n$. The mRNA levels of the two special genes show superlinear and sublinear scalings with cell volume, in agreement with the theoretical predictions (dashed lines). The mRNA number and cell volume are normalized by their initial values. (B) Same analysis as (A) for non-degradable proteins. Deviations are expected since the actual mRNA lifetimes are finite.

Numerical simulations

We numerically tested our theoretical predictions by simulating a cell with 2000 genes, including RNAP and ribosome, which we coarse-grained as single proteins. Simulation details are summarized in Methods, and parameter values are shown in Table S1. To avoid the effects of time-dependent gene copy numbers, which is certainly important but also complicates our analysis, we considered the scenario in which the cell volume grows without DNA replication, e.g., cells arrested in the G1 phase, which is a common experimental protocol to study the effects of cell volume on gene expression [21, 33].

We first simulated the simplified model with homogeneous recruitment abilities to RNAPs. We confirmed the exponential growth of cell volume and the linear scaling between mRNA copy numbers, protein copy numbers, and cell volume (Figure S2). We then simulated the case when all genes share the same recruitment ability K_n, but two genes have different abilities that are respectively smaller and larger than K_n. Our theoretical predictions for the expression levels of mRNAs, non-degradable proteins (Figure 3A, B), and degradable proteins (Figure S3), assuming short mRNA lifetimes match the simulation results reasonably well.

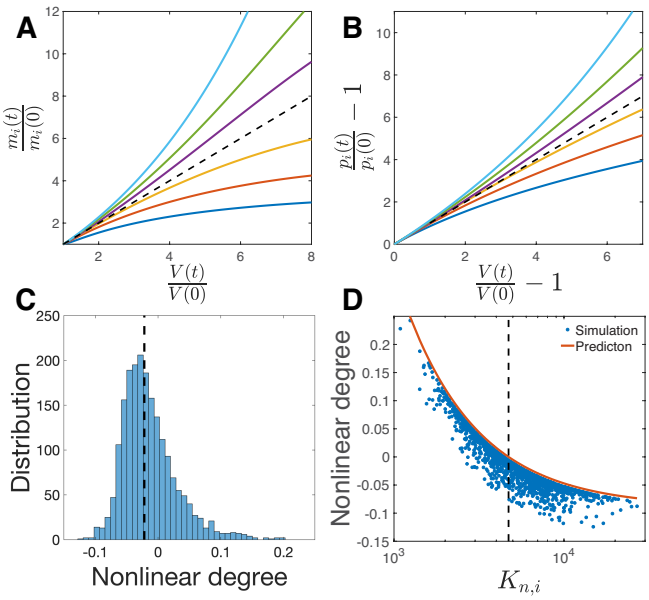

FIG. 4. (A) Examples of mRNA numbers *vs.* cell volume including superlinear and sublinear scaling. The dashed line is the $y = x$ line. (B) Examples of protein number increment *vs.* cell volume increment. The dashed line is the $y = x$ line. Here we show the results for non-degradable proteins and our results are equally valid for degradable proteins (Figure S4). (C) Distribution of the measured nonlinear degrees β of mRNA numbers from numerical simulations. The dashed line marks the location of the median value of the nonlinear degrees. (D) We compare the theoretically predicted nonlinear degrees of mRNA numbers and the measured one from numerical simulations. We note that deviations are expected since the theoretical predictions are based on the assumption that the lifetimes of mRNAs are infinitely short. The dashed line marks the location of the average $\langle K_{n,i}\rangle$ weighted by protein mass fractions at which the nonlinear degree is predicted to be zero.

Finally, we simulated the more realistic scenario in which $K_{n,i}$ is continuously distributed, which we modeled as a lognormal distribution. Our conclusions are independent of the specific choice of the form of distribution. Examples of volume dependence of several mRNAs and proteins are shown in Figure 4A, B. Nonlinear degrees of mRNAs and proteins are measured based on Eqs. (8, 9). The resulting distribution of mRNAs is shown in Figure 4C, and those of proteins are shown in Figure S4. To test our theoretical predictions on the relations between the nonlinear degrees β, α, and $K_{n,i}$ (Eqs. (20, 22)), we also need the expression of the appropriate average $\langle K_{n,i}\rangle$. We found that the appropriate average Michaelis-Menten constant is the one weighted by the initial protein mass fractions, $\langle K_{n,i}\rangle = \sum_i \phi_i K_{n,i}$, where ϕ_i is the mass fraction of proteins i in the entire proteome at time $t = 0$. The above expression of $\langle K_{n,i}\rangle$ leads to a good agreement between numerical simulations and theoretical predictions (Figure 4D). We set the Michaelis-Menten constants of RNAP and ribosome as the average value to ensure that their concentrations are approximately constant as the cell volume increases.

We note that the recruitment ability not only determines the nonlinear degree of volume scaling but also affects the mRNA production rate since a higher recruitment ability enhances the binding probability of RNAP to the promoter. Therefore, our model predicts a positive correlation between the mRNA production rates and the nonlinear degree β (Figure S4C). Our results suggest that those genes with sublinear scaling and smaller $K_{n,i}$ contribute more in the weighted average of $K_{n,i}$, therefore $\langle K_{n,i}\rangle < \overline{K_{n,i}}$ where $\overline{K_{n,i}}$ is the average over all genes with equal weights. Therefore, genes with $K_{n,i} \approx \overline{K_{n,i}}$ are expected to exhibit superlinear scaling (Figure 4D). We note that this does not imply most genes show superlinear scaling. In fact, we found that the median value of nonlinear degrees is close to zero therefore most genes still show approximately linear scaling (Figure 4C).

Analysis of experimental data and searching for motifs in the promoter sequences

We analyzed the genome-wide dataset from Ref. [21] where the volume dependences of mRNA levels are measured for *S. cerevisiae*. We calculated the nonlinear degrees of mRNA scaling with cell volume using Eq. (8) and obtained the resulting distribution (Figure 5A). The calculated nonlinear degree β is indeed highly correlated with the nonlinear scale calculated in Ref. [21] (Figure S5). The median value of nonlinear degrees is close to zero, suggesting that the majority of genes show approximately linear scaling, similar to our numerical simulations. We used GSEA (Gene Set Enrichment Analysis) [37, 38] to find annotated functional gene sets in KEGG [39–41] that are enriched in the superlinear and sublinear scaling regime with a threshold p value = 5.00e-2 (Figure S6). Interestingly, we found that the ribosomal genes and other translation-related genes, which correspond to the coarse-grained ribosomal genes in our model, are enriched in the superlinear regime with the average β over ribosomal genes about -0.2, which may be induced by the drug that blocks cell-cycle progress [21]. For RNAP related genes, we found that they indeed show linear scaling with cell volume as we assume in our coarse-grained model, which is crucial for the linear scaling between the mRNA copy numbers and cell volume (Figure S7). To confirm the validity of our conclusions in the presence of weakly superlinear scaling of ribosomes, we numerically simulated our gene expression model with the recruitment ability of ribosomal gene weaker than the average value and found that our results are still valid (Figure S8).

In Ref. [21], the expression of pre-selected activators for cell cycle were shown to be superlinear, while in-

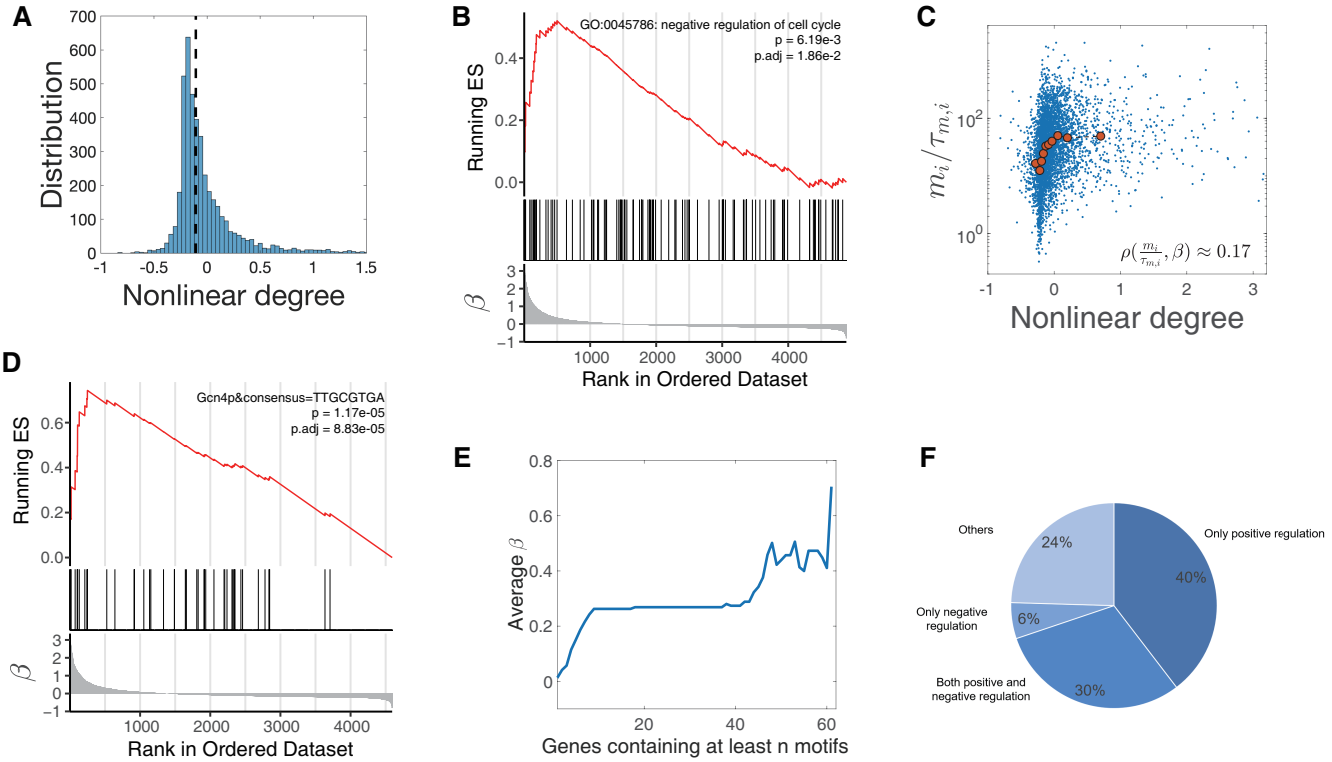

FIG. 5. (A) Distribution of the nonlinear degrees of cell volume scaling of mRNAs for *S. cerevisiae* among genes. The dashed line marks the location of the median value of the nonlinear degrees. In our analysis of the the dataset from Ref. [21], we consider genes with $-1 < \beta < 3.2$ including 95% of all the measured genes. (B) Genes annotated as negative regulation of cell cycle (GO:0045786) are enriched in the sublinear regime. In the bottom panel, genes are ordered by the nonlinear degree β from positive (sublinear) to negative (superlinear). In the middle panel, the vertical lines represent the locations of the inhibitors. The upper panel shows the running enrichment score (ES) for the gene set, where the score at the peak is the ES for this gene set. Plot was made using R package clusterProfiler [34]. (C) A positive and significant Pearson correlation (Pearson correlation test, p value < 2.20e-16) between the nonlinear degrees of mRNAs and their initial amounts divided by the lifetimes [35], which are proxies for the mRNA production rates. The red data points are median values after binning. (D) An example of motif that blue is enriched in the sublinear regime. Here, the vertical lines in the middle panel represent the locations of genes containing the particular motif in the promoter sequences. Note that the motif can also appear in the weakly superlinear regime, but clear diminishes in the strongly superlinear regime. The bottom and upper panels are the same as are described in (B). (E) We pick out all 80 motifs that are enriched in the promoters of sublinear genes and calculate the average β over genes with at least n motifs. Promoters are defined as 500 bp upstream of transcription start sites. (F) The functions of transcription factors associated with the 80 motifs enriched in the sublinear regime. Functional enrichment analysis was done using Metascape [36]. Positive regulation of transcription by RNA polymerase II (GO:0045944) is enriched with p value = 2.41e-32.

hibitors were shown to be sublinear. So we next checked the nonlinear degree of cell cycle regulators annotated in the Gene Ontology database [42–44] using GSEA. We found inhibitors enriched in the sublinear regime (Figure 5B), but activators were not enriched in the superlinear regime. We note that our results do not contradict with Ref. [21] and support the view that the ratio of activators to inhibitors could trigger cell cycle progress.

To detect the correlation between the mRNA production rate and β, we used the mRNA amounts (the product of RPKM and cell volume) at the smallest cell volume divided by the mRNA lifetimes from Ref. [35] as a proxy for the mRNA production rate according to Eq. (3). A weak but significant correlation is indeed observed, as shown in Figure 5C. We note that both of the heterogeneous initiation rates and gene copy numbers can reduce the correlation between the mRNA production rate and the Michaelis-Menten constant $K_{n,i}$.

To further support our theoretical predictions, we investigated the promoter sequences from the Yeastract database [45] of all genes included in our analysis. We expect that those genes with nonlinear scaling should have some special patterns in their promoter sequences which render them stronger or weaker recruitment abilities to RNAPs than the others. If this is the case, specific motifs should be enriched in the superlinear or sublinear regime. We used the CentriMo [46] tool of MEME suite [47] to detect the transcription factors binding motifs an-

notated in the Yeastract database [45] in the promoter sequences and then used GSEA to identify those motifs that are enriched in the nonlinear regime with a threshold p value = 5.00e-2. We found 80 motifs enriched in the sublinear regime (see one example in Figure 5D and Table S2). To further validate our results, we computed the average β for genes containing at least n motifs that are enriched in the sublinear regime and found that the average β indeed increases as a function of n (Figure 5E). As expected, 70% of the corresponding transcription factors exhibit positive regulation and therefore enhance the recruitment abilities to RNAPs of their target genes (Figure 5F). However, we did not find motifs enriched in the superlinear regime. Considering the cumulative effect of motifs on β as shown in Figure 5D, we propose that antagonistic effect may also exist among motifs, which suggests that motifs reducing the recruitment ability to RNAPs may reside in most genes. But in genes without superlinear scaling, their effects are counteracted by other motifs that enhance the recruitment ability.

DISCUSSION

For abnormally large cells, the number of RNA polymerases or ribosomes may exceed some threshold values so that the bottlenecks of gene expression become the templates of gene expression: gene copy numbers and mRNA copy numbers [14, 33]. However, in typical cellular physiological states, cells are far below the thresholds, and the limiting bottlenecks of gene expression are the copy numbers of RNAPs and ribosomes [10, 15, 16]. In this case, if the promoters of all genes share the same recruitment ability, the expression levels of all genes should exhibit linear scaling with cell volume both at the mRNA and protein level [13]. We extended this simple scenario to a more realistic case in which the recruitment abilities among genes are continuously distributed. We derived the dependence of the mRNA production rate on cell volume. We show that genes with recruitment abilities below (above) the average exhibit superlinear (sublinear) scaling with cell volume, a natural consequence of the heterogeneous distribution of recruitment abilities. In other words, active regulation is not necessary to achieve the nonlinear scaling for a subset of genes. We further show that the nonlinear scaling between the mRNA production rates and cell volume propagates to the mRNA copy numbers and proteins copy numbers. All of our theoretical predictions were confirmed by numerical simulations.

The recruitment abilities not only determine the nonlinear degrees of gene expression but also determines the mRNA production rates. Therefore, our model predicts that genes with higher (lower) mRNA production rates are more likely to exhibit sublinear (superlinear) scaling with cell volumes, in concert with experimental data in *S. cerevisiae* [21]. Furthermore, according to our theoretical models, motifs that enhance or reduce the promoters' recruitment abilities should exist in the superlinear or sublinear scaling genes. Indeed, we found a group of motifs enriched in the sublinear regime, and these motifs are associated with transcription factor binding sites that have positive regulation on the target genes. We note that in Ref. [21], the numbers of binding sites for specific cell-cycle transcription factors in the target genes were found to be positively correlated with the superlinear degrees of their mRNA levels, which may be due to other mechanisms not related to our model.

Our results can have far-reaching implications: the nonlinear scaling of gene expression level allows cells to sense their sizes based on the ratio of concentrations of different proteins, enabling cells to decide the timing of multiple cell-cycle events such as cell division. Our results suggest that sensing the concentration differences among a group of proteins as a measure of cell volume can be the most accessible option cells can take to achieve cell size regulation. Finally, the gene expression model proposed in this work is by definition at the whole-cell level. Therefore, it can be a valuable platform for the mathematical modeling of gene expression, especially for problems in which the competition among genes for the limiting resources of RNAPs and ribosomes are crucial.

METHODS

Derivation of the nonlinear scaling

We consider a simple model assuming all genes have the same MM constant K_n except one special gene i has a MM constant $K_{n,i}$. Since the contribution of the particular gene to the global allocation of RNAP is negligible, we find that

$$\frac{c_n F_n}{c_n F_n + K_n} \approx \frac{n - nF_n}{\sum_j g_j(1+\Lambda_{n,j})}. \tag{10}$$

We now focus Phase 1 so that $F_n \ll 1$ and the mRNA production rate for the particular gene becomes

$$k_{n,i} = \Gamma_{n,i} g_i \frac{K_n n}{K_{n,i} n_c - (K_{n,i} - K_n) n}. \tag{11}$$

If $K_{n,i} > K_n$, we find that the mRNA production rate of gene i exhibits a superlinear dependence on the RNAP number, therefore also the cell volume. If $K_{n,i} < K$, the mRNA production rate of gene i exhibits a sublinear behavior. The general solution for Eq. (3) becomes

$$m_i(t) = m_i(0) e^{-t/\tau_{m,i}} + \int_0^t e^{-\Delta t/\tau_{m,i}} k_{n,i}(t-\Delta t) d\Delta t. \tag{12}$$

In the limit that the lifetime of mNRA goes to zero, the number of mRNA becomes strictly proportional to the

mRNA production rate

$$m_i(t) = \tau_{m,i} k_{n,i}(t). \tag{13}$$

We now consider the dynamics of protein number which is

$$\frac{dp_i}{dt} = \Gamma_{r,i} \frac{c_r F_r}{c_r F_r + K_{r,i}} m_i - \frac{p_i}{\tau_{p,i}}. \tag{14}$$

The general solution for Eq. (14) is

$$p_i(t) = p_i(0) e^{-t/\tau_{p,i}} + \int_0^t e^{-\Delta t/\tau_{p,i}} \Gamma_{r,i} \frac{c_r F_r}{c_r F_r + K_{r,i}} m_i(t - \Delta t) d\Delta t. \tag{15}$$

In the following, we assume $c_r F_r$ to be constant, as a good approximation in Phase 1 (SI B) and consider two limiting cases, $\tau_{p,i} \to 0$ and $\tau_{p,i} \to \infty$.

When $\tau_{p,i} \to 0$, the protein number becomes strictly proportional to the protein production rate, which is proportional to the mRNA number,

$$\begin{aligned} p_i(t) &= \Gamma_{r,i} \frac{c_r F_r}{c_r F_r + K_{r,i}} m_i(t) \tau_{p,i} \\ &= \Gamma_{r,i} \frac{c_r F_r}{c_r F_r + K_{r,i}} \tau_{m,i} \tau_{p,i} k_{n,i}(t). \end{aligned} \tag{16}$$

Note that the second identity is valid when $\tau_{m,i} \to 0$ which is a good approximation when the lifetime of mRNA is much shorter than the doubling time.

When $\tau_{p,i} \to \infty$ and $\tau_{m,i} \to 0$ the dynamics of protein number becomes

$$\begin{aligned} p_i(t) - p_i(0) &= \int_0^t \Gamma_{r,i} \frac{c_r F_r}{c_r F_r + K_{r,i}} m_i(t') dt' \\ &= \Gamma_{r,i} \frac{c_r F_r}{c_r F_r + K_{r,i}} \Gamma_{n,i} g_i \tau_{m,i} \int_0^t \frac{K_n n}{K_{n,i} n_c - (K_{n,i} - K_n) n} dt'. \end{aligned} \tag{17}$$

In Phase 1, the number of RNA polymerase increases exponentially $n(t) = n(0) e^{\mu t}$, therefore the integral in Eq. (17) can be analytically calculated

$$\begin{aligned} &p_i(t) - p_i(0) \\ &= \Gamma_{r,i} \Gamma_{n,i} g_i \tau_{m,i} \frac{c_r F_r}{c_r F_r + K_{r,i}} \frac{K_n}{(K_n - K_{n,i}) \mu} \times \\ &\ln \Big(\frac{K_{n,i} n_c - (K_{n,i} - K_n) n(t)}{K_{n,i} n_c - (K_{n,i} - K_n) n(0)} \Big). \end{aligned} \tag{18}$$

The nonlinear scaling of mRNA copy number also propagates to non-degradable proteins.

To quantify the nonlinear degrees of mRNA copy number, we investigate the volume dependence of mRNA copy number and normalize the volume and mRNA number by their initial values. Using Eqs. (11, 13), we obtain

$$\widetilde{m}_i(t) = \widetilde{V}(t) \frac{1 - \frac{\Delta K_{n,i} n(0)}{K_{n,i} n_c}}{1 - \frac{\Delta K_{n,i} n(0)}{K_{n,i} n_c} \widetilde{V}(t)}. \tag{19}$$

Here $\widetilde{m}_i(t) = m_i(t)/m_i(0)$, $\widetilde{V}(t) = V(t)/V(0)$, $\Delta K_{n,i} = K_{n,i} - K_n$. We have used the fact that the RNAP concentration is constant therefore $n(t)/n(0) = V(t)/V(0)$. Note that when $K_{n,i}$ is continuously distributed, we should replace K_n by $\langle K_{n,i} \rangle$. Comparing with Eq. (8) in the main text, we find that

$$\beta_i = -\frac{\Delta K_{n,i} n(0)}{K_{n,i} n_c}. \tag{20}$$

When $\Delta K_{n,i} = 0$, $\beta_i = 0$ as expected. For proteins with short lifetimes, the above analysis is equally valid.

We also study the nonlinear degree of non-degradable proteins, and using Eq. (18), we find that

$$\begin{aligned} \Delta \widetilde{p}_i(t) &= C_i \ln \Big(1 - \frac{\Delta K_{n,i} (n(t) - n(0))}{K_{n,i} n_c - \Delta K_{n,i} n(0)} \Big) \\ &= C_i \ln \Big(1 - \frac{\Delta K_{n,i} n(0)}{K_{n,i} n_c - \Delta K_{n,i} n(0)} \Delta \widetilde{V}(t) \Big). \end{aligned} \tag{21}$$

Here we combine all the constants before the logarithmic term in Eq. (18) to C_i. Comparing with Eq. (9) in the main text, we find that

$$\alpha_i = \frac{\Delta K_{n,i} n(0)}{K_{n,i} n_c - \Delta K_{n,i} n(0)}. \tag{22}$$

In the limit $\Delta K_{n,i} \to 0$, we find that

$$\Delta p_i = \Gamma_{r,i} \Gamma_{n,i} g_i \tau_{m,i} \frac{c_r F_r}{c_r F_r + K_{r,i}} \frac{1}{\mu} \frac{n(t) - n(0)}{n_c}, \tag{23}$$

which is consistent with Eq. (17) in the case of $K_{n,i} = K_n$.

Details of numerical simulations

We summarize some of the parameters we used in the numerical simulations in Table S1. The gene copy numbers are time-independent which we set as 1 for all genes except the ribosome gene which we set as 5. Given an attempted growth rate μ_0, we set the attempted mass fraction of ribosomes in the entire proteome as

$$\phi_r = \mu_0 L_r / v_r. \tag{24}$$

Here L_r is the length of ribosome gene in the unit of codons. Note that the actual mass fraction is time-dependent and slightly deviates from the attempted value. Here to get Eq. (24), we assume that all ribosomes are actively translating and neglect the correction due to free ribosomes and initiating ribosomes. Given the mass fraction of ribosome, we assume that the copy number of RNAPs is about 10% of that of ribosomes and set the attempted mass fraction of RNAP as

$$\phi_n = 0.1 \phi_r L_n / L_r. \tag{25}$$

Here L_n is the length of RNAP gene. In Figure 3, we set $K_n = 6 \times 10^3/\mu m^3$, and the two special genes with $K_{n,i} = 20K_n$ and $K_{n,i} = 0.2K_n$. We also set the lifetimes of all mRNAs as 10 mins. In Figure 4, we set $K_{n,i}$ following a lognormal distribution so that its average $\overline{K_{n,i}} = 6\times10^3/\mu m^3$ and the coefficient of variation is equal to 0.5. Note the average weighted by protein mass fractions $\langle K_{n,i}\rangle$ is in general smaller than $\overline{K_{n,i}}$. We also set the lifetimes of mRNA following a lognormal distribution with a mean equal to 10 mins and a coefficient of variation equal to 1. In all simulations, we set the initial attempted total protein mass as $M_b = 10^9$ in the unit of amino acid number and the attempted critical RNAP number as $n_c = \sum_i g_i(1 + \Lambda_{n,i}) = 10^4$.

To find the appropriate value of $\Gamma_{n,i}$ that leads to the attempted n_c, we assume $c_nF_n \ll K_{n,i}$ so that

$$\frac{\phi_n}{\phi_r} \approx \frac{\Gamma_{n,n}\frac{g_n}{K_{n,n}}\tau_n L_n}{\Gamma_{n,r}\frac{g_r}{K_{n,r}}\tau_r L_r}, \tag{26}$$

from which we can find that $\Gamma_{n,n} = y\Gamma_{n,r}$ where y can be found using the above equation. We also set $\Gamma_{n,i} = x\Gamma_{n,r}$ for all i except the genes for RNAP ($i = 1$) and ribosome ($i = 2$), so that

$$\phi_r \approx \frac{\Gamma_{n,r}\frac{g_r}{K_{n,r}}\tau_r L_r}{\Gamma_{n,r}\frac{g_r}{K_{n,r}}\tau_r L_r + y\Gamma_{n,r}\frac{g_n}{K_{n,n}}\tau_n L_n + x\Gamma_{n,r}\sum_{i>2}\frac{g_i}{K_{n,i}}\tau_i L_i} \tag{27}$$

from which we can find the expression of x. Finally, using $\sum_i g_i(1 + \Gamma_{n,i}\frac{L_i}{v_n}) = n_c$, we find that

$$\Gamma_{n,r} = \frac{(n_c - \sum_i g_i)v_n}{g_rL_r + yg_nL_n + x\sum_{i>2} g_iL_i} \tag{28}$$

from which we find $\Gamma_{n,n}$ and $\Gamma_{n,i}$ for $i > 2$. Note that the approximation $c_nF_n \ll K_{n,i}$ for all i is merely to find the values of $\Gamma_{n,i}$ which do not affect our conclusions.

Before we run the simulations, we use the attempted ϕ_n, initial cell mass, and $\Gamma_{n,i}$ with g_i and $K_{n,i}$ to calculate the mRNA production rates for all genes from which we use $m_i = k_{n,i}\tau_{m,i}$ as the initial condition. Using the initial m_i, we compute the initial protein mass fractions as $\phi_i = m_iL_i/\sum_i m_iL_i$ and also update the mass fractions of RNAP and ribosome. The actual initial mass fraction of RNAP can slightly deviate from the attempted value. To make sure the RNAP copy number is continuous from the beginning of simulations, we also slightly shift the total protein mass from the attempted value so that $\phi_n(\text{actual})M_b(\text{actual}) = \phi_n(\text{attemped})M_b(\text{attempted})$.

To remove the transient effects, we take the simulation results at $t = 20$ as the initial values when we compare our results with theoretical predictions. To simulate degradable proteins, we choose 200 of the genes as degradable proteins with lifetimes $\tau_p = 10$ mins and for degradable proteins, we take the simulation results at $t = 100$ as the initial values.

[1] Harry A Crissman and John A Steinkamp, "Rapid, simultaneous measurement of dna, protein, and cell volume in single cells from large mammalian cell populations," The Journal of Cell Biology **59**, 766 (1973).

[2] SG Elliott and CS McLaughlin, "Rate of macromolecular synthesis through the cell cycle of the yeast saccharomyces cerevisiae," Proceedings of the National Academy of Sciences **75**, 4384–4388 (1978).

[3] Hanna Salman, Naama Brenner, Chih-kuan Tung, Noa Elyahu, Elad Stolovicki, Lindsay Moore, Albert Libchaber, and Erez Braun, "Universal protein fluctuations in populations of microorganisms," Phys. Rev. Lett. **108**, 238105 (2012).

[4] Samuel Marguerat and Jrg Bhler, "Coordinating genome expression with cell size," Trends in Genetics **28**, 560 – 565 (2012).

[5] Hermannus Kempe, Anne Schwabe, Frédéric Crémazy, Pernette J Verschure, and Frank J Bruggeman, "The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise," Molecular Biology of the Cell **26**, 797–804 (2015).

[6] Olivia Padovan-Merhar, Gautham P Nair, Andrew G Biaesch, Andreas Mayer, Steven Scarfone, Shawn W Foley, Angela R Wu, L Stirling Churchman, Abhyudai Singh, and Arjun Raj, "Single mammalian cells compensate for differences in cellular volume and dna copy number through independent global transcriptional mechanisms," Molecular Cell **58**, 339–352 (2015).

[7] Robert Ietswaart, Stefanie Rosa, Zhe Wu, Caroline Dean, and Martin Howard, "Cell-size-dependent transcription of flc and its antisense long non-coding rna coolair explain cell-to-cell expression variation," Cell Systems **4**, 622 – 635.e9 (2017).

[8] Xiao-yu Zheng and Erin K O'Shea, "Cyanobacteria maintain constant protein concentration despite genome copy-number variation," Cell Reports **19**, 497–504 (2017).

[9] Benjamin D Knapp, Pascal Odermatt, Enrique R Rojas, Wenpeng Cheng, Xiangwei He, Kerwyn Casey Huang, and Fred Chang, "Decoupling of rates of protein synthesis from cell expansion leads to supergrowth," Cell Systems **9**, 434–445 (2019).

[10] Xi-Ming Sun, Anthony Bowman, Miles Priestman, Francois Bertaux, Amalia Martinez-Segura, Wenhao Tang, Chad Whilding, Dirk Dormann, Vahid Shahrezaei, and Samuel Marguerat, "Size-dependent increase in rna polymerase ii initiation rates mediates gene expression scaling with cell size," Current Biology **30**, 1217 – 1230.e7 (2020).

[11] Manuel Campos, Ivan V Surovtsev, Setsu Kato, Ahmad Paintdakhi, Bruno Beltran, Sarah E Ebmeier, and Christine Jacobs-Wagner, "A constant size extension drives bacterial cell size homeostasis," Cell **159**, 1433–1446 (2014).

[12] Nathan Cermak, Selim Olcum, Francisco Feijo Delgado, Steven C Wasserman, Kristofor R Payer, Mark

A Murakami, Scott M Knudsen, Robert J Kimmerling, Mark M Stevens, Yuki Kikuchi, Arzu Sandikci, Masaaki Ogawa, Vincent Agache, Francois Baleras, David M Weinstock, and Scott R Manalis, "High-throughput measurement of single-cell growth rates using serial microfluidic mass sensor arrays," Nature Biotechnology **34**, 1052–1059 (2016).

[13] Jie Lin and Ariel Amir, "Homeostasis of protein and mrna concentrations in growing cells," Nature Communications **9**, 4496 (2018).

[14] Jacob Zhurinsky, Klaus Leonhard, Stephen Watt, Samuel Marguerat, Jürg Bähler, and Paul Nurse, "A coordinated global control over cellular transcription," Current Biology **20** (2010).

[15] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa, "Interdependence of Cell Growth and Gene Expression: Origins and Consequences," Science **330**, 1099 (2010).

[16] Eyal Metzl-Raz, Moshe Kafri, Gilad Yaakov, Ilya Soifer, Yonat Gurvich, and Naama Barkai, "Principles of cellular resource allocation revealed by condition-dependent proteome profiling," Elife **6**, e28034 (2017).

[17] Kurt M Schmoller, JJ Turner, M Kõivomägi, and Jan M Skotheim, "Dilution of the cell cycle inhibitor whi5 controls budding-yeast cell size," Nature **526**, 268–272 (2015).

[18] Frank S Heldt, Reece Lunstone, John J Tyson, and Béla Novák, "Dilution and titration of cell-cycle regulators may control cell size in budding yeast," PLoS computational biology **14**, e1006548 (2018).

[19] Athanasios Litsios, Daphne H. E. W. Huberts, Hanna M. Terpstra, Paolo Guerra, and Matthias Heinemann, "Differential scaling between g1 protein production and cell size dynamics promotes commitment to the cell division cycle in budding yeast," Nature Cell Biology **21** (2019).

[20] Yimiao Qu, Jun Jiang, Xiang Liu, Ping Wei, Xiaojing Yang, and Chao Tang, "Cell cycle inhibitor whi5 records environmental information to coordinate growth and division in yeast," Cell reports **29**, 987–994 (2019).

[21] Yuping Chen, Gang Zhao, Jakub Zahumensky, Sangeet Honey, and Bruce Futcher, "Differential scaling of gene expression with cell size may explain size control in budding yeast," Molecular Cell **78**, 359 – 370.e6 (2020).

[22] Kora-Lee Claude, Daniela Bureik, Petia Adarska, Abhyudai Singh, and Kurt M Schmoller, "Transcription coordinates histone amounts and genome content," bioRxiv (2020).

[23] Stefan Klumpp and Terence Hwa, "Growth-rate-dependent partitioning of rna polymerases in bacteria," Proceedings of the National Academy of Sciences **105**, 20245–20250 (2008).

[24] MARIE T Killeen and JACK F Greenblatt, "The general transcription factor rap30 binds to rna polymerase ii and prevents it from binding nonspecifically to dna." Molecular and Cellular Biology **12**, 30–37 (1992).

[25] Robert G. Roeder, "The role of general initiation factors in transcription by rna polymerase ii," Trends in Biochemical Sciences **21**, 327 – 335 (1996).

[26] Chin-Tong Ong and Victor G Corces, "Enhancer function: new insights into the regulation of tissue-specific gene expression," Nature Reviews Genetics **12**, 283–293 (2011).

[27] Andrew D Basehoar, Sara J Zanton, and B Franklin Pugh, "Identification and distinct regulation of yeast tata box-containing genes," Cell **116**, 699–709 (2004).

[28] Ron Milo and Rob Phillips, *Cell biology by the numbers* (Garland Science, 2015).

[29] Sophia Hsin-Jung Li, Zhiyuan Li, Junyoung O Park, Christopher G King, Joshua D Rabinowitz, Ned S Wingreen, and Zemer Gitai, "Escherichia coli translation strategies differ across carbon, nitrogen and phosphorus limitation conditions," Nature microbiology **3**, 939 (2018).

[30] Andrea K. Bryan, Vivian C. Hecht, Wenjiang Shen, Kristofor Payer, William H. Grover, and Scott R. Manalis, "Measuring single cell mass, volume, and density with dual suspended microchannel resonators," Lab Chip **14**, 569–576 (2014).

[31] Markus Basan, Manlu Zhu, Xiongfeng Dai, Mya Warren, Daniel Sévin, Yi-Ping Wang, and Terence Hwa, "Inflating bacterial cells by increased protein synthesis," Molecular Systems Biology **11** (2015), 10.15252/msb.20156178.

[32] Frank R Neumann and Paul Nurse, "Nuclear size control in fission yeast," The Journal of Cell Biology **179**, 593–600 (2007).

[33] Gabriel E Neurohr, Rachel L Terry, Jette Lengefeld, Megan Bonney, Gregory P Brittingham, Fabien Moretto, Teemu P Miettinen, Laura Pontano Vaites, Luis M Soares, Joao A Paulo, *et al.*, "Excessive cell growth causes cytoplasm dilution and contributes to senescence," Cell **176**, 1083–1097 (2019).

[34] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He, "clusterProfiler: an r package for comparing biological themes among gene clusters," OMICS: A Journal of Integrative Biology **16**, 284–287 (2012).

[35] Benjamin Neymotin, Rodoniki Athanasiadou, and David Gresham, "Determination of in vivo RNA kinetics using rate-seq," RNA **20**, 1645–1652 (2014).

[36] Yingyao Zhou, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K. Chanda, "Metascape provides a biologist-oriented resource for the analysis of systems-level datasets," Nature Communications **10** (2019), 10.1038/s41467-019-09234-6.

[37] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," Proceedings of the National Academy of Sciences **102**, 15545–15550 (2005).

[38] Vamsi K. Mootha, Cecilia M. Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstrle, Esa Laurila, and et al., "Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," Nature Genetics **34**, 267273 (2003).

[39] M. Kanehisa, "KEGG: Kyoto encyclopedia of genes and genomes," Nucleic Acids Research **28**, 27–30 (2000).

[40] Minoru Kanehisa, "Toward understanding the origin and evolution of cellular organisms," Protein Science **28**, 1947–1951 (2019).

[41] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Mari Ishiguro-Watanabe, and Mao Tanabe, "KEGG: integrating viruses and cellular organisms," Nucleic Acids Research **49**, D545–D551 (2020).

[42] Michael Ashburner, Catherine A. Ball, Judith A. Blake,

David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock, "Gene ontology: tool for the unification of biology," Nature Genetics **25**, 25–29 (2000).

[43] Seth Carbon, Eric Douglass, Benjamin M Good, Deepak R Unni, Nomi L Harris, Christopher J Mungall, Siddartha Basu, Rex L Chisholm, Robert J Dodson, Eric Hartline, Petra Fey, Paul D Thomas, Laurent-Philippe Albou, Dustin Ebert, Michael J Kesling, Huaiyu Mi, Anushya Muruganujan, Xiaosong Huang, Tremayne Mushayahama, Sandra A LaBonte, Deborah A Siegele, Giulia Antonazzo, Helen Attrill, Nick H Brown, Phani Garapati, Steven J Marygold, Vitor Trovisco, Gil dos Santos, Kathleen Falls, Christopher Tabone, Pinglei Zhou, Joshua L Goodman, Victor B Strelets, Jim Thurmond, Penelope Garmiri, Rizwan Ishtiaq, Milagros Rodríguez-López, Marcio L Acencio, Martin Kuiper, Astrid Lægreid, Colin Logie, Ruth C Lovering, Barbara Kramarz, Shirin C C Saverimuttu, Sandra M Pinheiro, Heather Gunn, Renzhi Su, Katherine E Thurlow, Marcus Chibucos, Michelle Giglio, Suvarna Nadendla, James Munro, Rebecca Jackson, Margaret J Duesbury, Noemi Del-Toro, Birgit H M Meldal, Kalpana Paneerselvam, Livia Perfetto, Pablo Porras, Sandra Orchard, Anjali Shrivastava, Hsin-Yu Chang, Robert Daniel Finn, Alexander Lawson Mitchell, Neil David Rawlings, Lorna Richardson, Amaia Sangrador-Vegas, Judith A Blake, Karen R Christie, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry M Sitnikov, Midori A Harris, Stephen G Oliver, Kim Rutherford, Valerie Wood, Jaqueline Hayles, Jürg Bähler, Elizabeth R Bolton, Jeffery L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Cody Plasterer, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lisa Matthews, James P Balhoff, Suzi A Aleksander, Michael J Alexander, J Michael Cherry, Stacia R Engel, Felix Gondwe, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Matt Simison, Marek S Skrzypek, Shuai Weng, Edith D Wong, Marc Feuermann, Pascale Gaudet, Anne Morgat, Erica Bakker, Tanya Z Berardini, Leonore Reiser, Shabari Subramaniam, Eva Huala, Cecilia N Arighi, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Alex Bateman, Marie-Claude Blatter, Emmanuel Boutet, Emily Bowler, Lionel Breuza, Alan Bridge, Ramona Britto, Hema Bye-A-Jee, Cristina Casals Casas, Elisabeth Coudert, Paul Denny, Anne Estreicher, Maria Livia Famiglietti, George Georghiou, Arnaud Gos, Nadine Gruaz-Gumowski, Emma Hatton-Ellis, Chantal Hulo, Alexandr Ignatchenko, Florence Jungo, Kati Laiho, Philippe Le Mercier, Damien Lieberherr, Antonia Lock, Yvonne Lussi, Alistair MacDougall, Michele Magrane, Maria J Martin, Patrick Masson, Darren A Natale, Nevila Hyka-Nouspikel, Sandra Orchard, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Sangya Pundir, Catherine Rivoire, Elena Speretta, Shyamala Sundaram, Nidhi Tyagi, Kate Warner, Rossana Zaru, Cathy H Wu, Alexander D Diehl, Juancarlos N Chan, Christian Grove, Raymond Y N Lee, Hans-Michael Muller, Daniela Raciti, Kimberly Van Auken, Paul W Sternberg, Matthew Berriman, Michael Paulini, Kevin Howe, Sibyl Gao, Adam Wright, Lincoln Stein, Douglas G Howe, Sabrina Toro, Monte Westerfield, Pankaj Jaiswal, Laurel Cooper, and Justin Elser, "The gene ontology resource: enriching a GOld mine," Nucleic Acids Research **49**, D325–D334 (2020).

[44] Seth Carbon, Amelia Ireland, Christopher J. Mungall, ShengQiang Shu, Brad Marshall, Suzanna Lewis, and and, "AmiGO: online access to ontology and annotation data," Bioinformatics **25**, 288–289 (2008).

[45] Pedro T Monteiro, Jorge Oliveira, Pedro Pais, Miguel Antunes, Margarida Palma, Mafalda Cavalheiro, Mónica Galocha, Cláudia P Godinho, Luís C Martins, Nuno Bourbon, *et al.*, "Yeastract+: a portal for cross-species comparative genomics of transcription regulation in yeasts," Nucleic acids research **48**, D642–D649 (2020).

[46] Timothy L. Bailey and Philip Machanick, "Inferring direct DNA binding from ChIP-seq," Nucleic Acids Research **40**, e128–e128 (2012).

[47] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "MEME SUITE: tools for motif discovery and searching," Nucleic Acids Research **37**, W202–W208 (2009).