Protein overabundance is driven by growth robustness

H. James Choi,[1] Teresa W. Lo,[1] Kevin J. Cutler,[1] Dean Huang,[1] W. Ryan Will,[2] and Paul A. Wiggins[1,3,4,*]

[1]*Department of Physics, University of Washington, Seattle, Washington 98195, USA*
[2]*Department of Laboratory Medicine and Pathology, University of Washington, Seattle, Washington 98195, USA*
[3]*Department of Microbiology, University of Washington, Seattle, Washington 98195, USA*
[4]*Department of Bioengineering, University of Washington, Seattle, Washington 98195, USA*

Protein expression levels optimize cell fitness: Too low an expression level of essential proteins will slow growth by compromising essential processes; whereas overexpression slows growth by increasing the metabolic load. This trade-off naïvely predicts that cells maximize their fitness by sufficiency, expressing just enough of each essential protein for function. We test this prediction in the naturally-competent bacterium *Acinetobacter baylyi* by characterizing the proliferation dynamics of essential-gene knockouts at a single-cell scale (by imaging) as well as at a genome-wide scale (by TFNseq). In these experiments, cells proliferate for multiple generations as target protein levels are diluted from their endogenous levels. This approach facilitates a proteome-scale analysis of protein overabundance. As predicted by the Robustness-Load Trade-Off (RLTO) model, we find that roughly 70% of essential proteins are overabundant and that overabundance increases as the expression level decreases, the signature prediction of the model. These results reveal that robustness plays a fundamental role in determining the expression levels of essential genes and that overabundance is a key mechanism for ensuring robust growth.

Understanding the rationale for protein expression levels is a fundamental question in biology with broad implications for understanding cellular function [1]. Measured expression levels appear to be paradoxically both *optimal* and *overabundant*. For instance, repeated investigations support the idea that gene expression levels optimize cell fitness [2, 3]. Since the overall metabolic cost of protein expression is large [4, 5], fitness optimization would seem to imply that protein levels should satisfy a *Goldilocks condition*: Expression levels should be *just high enough* to achieve the required protein activity [6, 7]. However, a range of approaches suggest that many essential genes are expressed in vast excess of the levels required for function [7–9]. How can expression levels be at once optimal with respect to fitness as well as in excess of what is required for function?

The cell faces a complex regulatory challenge: Even in a bacterium, there are between five and six hundred essential proteins, each of which is required for growth [10]. How does the cell ensure the robust expression of each essential factor? We recently argued that the stochasticity of gene expression processes fundamentally shape the principles of central dogma regulation, including the optimality of protein overabundance [11]. Specifically, we proposed a quantitative model, the Robustness-Load Trade-Off (RLTO) model, which makes a parameter-free prediction of protein overabundance as a function of gene transcription level [11]. The optimality of overabundance can be understood as the result of a highly-asymmetric fitness landscape: the fitness cost of essential protein underabundance, which causes the arrest of essential processes, is far greater than the fitness cost of essential protein overabundance, which leads to slow growth by increasing the metabolic load. However, critical model assumptions and predictions remain untested which is the motivation for the current study. Here, we will quantitatively measure the fitness landscape with repect to protein abundance and determine the level of overabundance for all essential proteins in the bacterium *Acinetobacter baylyi*.

RESULTS

Natural competence facilitates knockout-depletion. To characterize the fitness landscape for essential gene expression, we must deplete the levels of essential proteins. Both degron- and CRISPRi-based approaches have been applied; however, these approaches require careful characterization of protein levels [8, 12–15] and introduce significant cell-to-cell variation on top of the endogenous noise which further obscures the underlying fitness landscape [16]. To circumvent these difficulties, we will use an alternative approach: *knockout-depletion* in the naturally competent bacterium *A. baylyi* ADP1 [17, 18]. In this approach, cells are transformed with a *geneX::kan* knockout cassettes targeting essential gene X, carrying a kanamycin resistance allele Km^R. (See Fig. 1A.) Cells that are not transformed arrest immediately on selective media. The crux of the approach is that transformants remain transiently *geneX*$^+$, due to the presence of already synthesized target protein X, even after the transcription of the target *geneX* stops. Growth can continue, diluting protein X abundance, as long as this residual abundance remain sufficient for function. The success of the knockout-depletion approach is dependent on the extremely high transformation efficiency of *A. baylyi*.

Target proteins are depleted by dilution. A key

* pwiggins@uw.edu

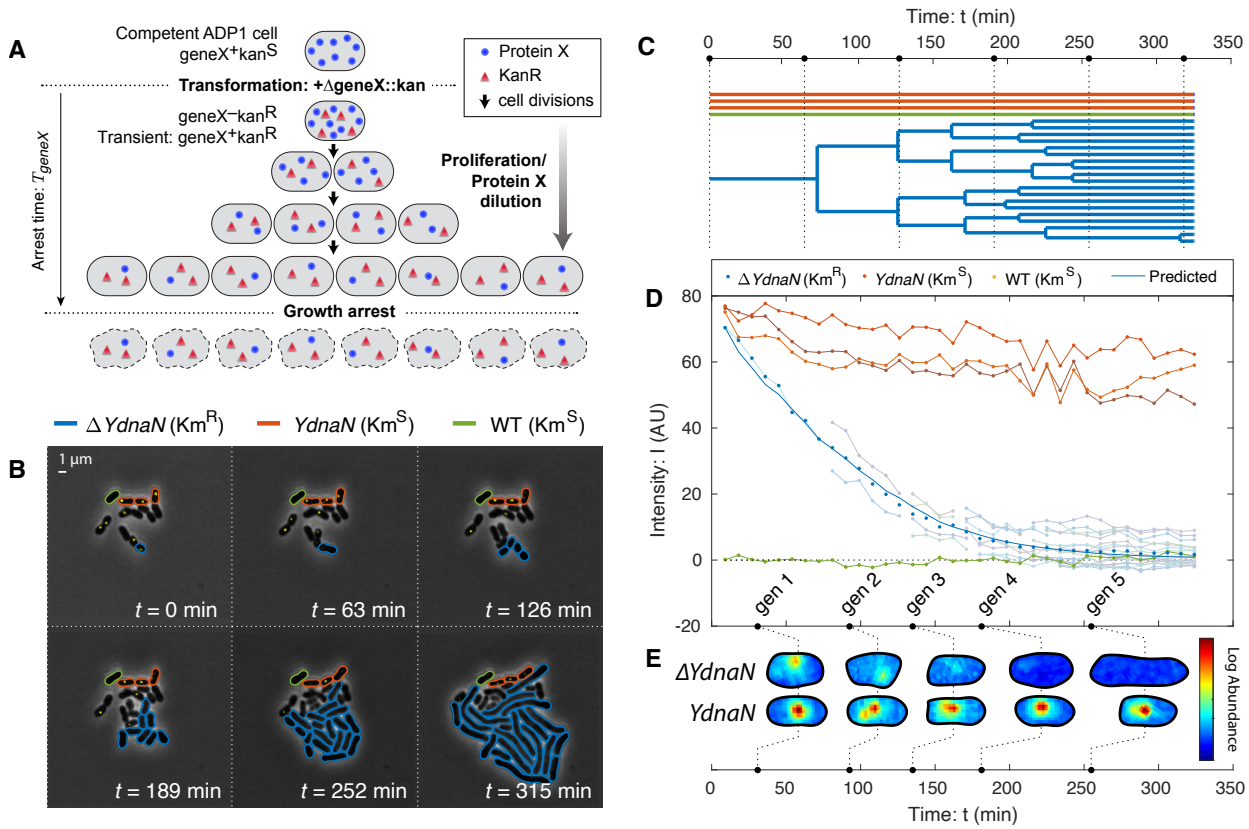

FIG. 1. **Knockout-depletion experiments. Panel A: Experimental schematic.** Competent ADP1 cells are transformed with Δ*geneX::kan*. Untransformed cells arrest immediately on selective media. Transformed cells proliferate, but cease protein X expression (blue circles) while expressing Kan (red triangles). Existing protein X abundance is diluted as cells proliferate. For essential genes, cell growth continues until protein levels are diluted to the threshold level required for growth, after which growth arrests. **Panel B & C: Visualization of knockout depletion.** The fluorescent fusion *YPet-dnaN* to essential gene *dnaN* is knocked outed at $t = 0$. Cell proliferation is visualized using phase-contrast microscopy while protein abundance is measured by fluorescence microscopy (yellow). Transformed cells (Δ*YdnaN*, blue) have a Km^R allele and can proliferate over several generations before arrest; however, untransformed cells (*YdnaN*, orange) and wild-type cells (WT, green) were both kanamycin sensitive and therefore arrested immediately. **Panel C: Lineage tree.** Black dotted lines represent time points shown in Panel B. **Panel D: Target protein is diluted by proliferation.** Protein concentration is measured by integrated fluorescence. Arrested *YdnaN* cells maintain protein abundance, whereas proliferating transformed cells (Δ*YdnaN*, blue) show growth-induced protein depletion. The protein concentration over all transformed progeny (blue points) are consistent with the dilution-model prediction (solid blue). **Panel E: Protein function is robust to dilution.** Representative single-cell images of transformed (Δ*YdnaN*) and untransformed (*YdnaN*) cells are shown for successive time points. The YPet-DnaN fusion shows punctate localization, consistent with function, even as protein abundance is depleted. No puncta are observed in the last generation and the cells form filaments, consistent with replication arrest.

untested assumption in the experimental design of the knockout-depletion approach is that target protein translation stops after transformation, and that the protein abundance is depleted by dilution. The model predicts that the protein concentration is:

$$C(t) = C_0 \cdot V_0/V(t), \tag{1}$$

where C_0 and V_0 are the concentration and volume of the progenitor cell at deletion and $V(t)$ is the total volume of the progeny. To test the predicted protein depletion hypothesis, we designed a knockout-depletion experiment to target a protein we had previously studied that can be visualized using a fluorescent fusion and whose localization is activity dependent: the essential replication gene *dnaN*, whose gene product is the β sliding clamp [19–21]. We constructed a N-terminal fluorescent fusion to *dnaN* using YPet in *A. baylyi* at the endogenous locus. The resulting mutant (YdnaN) had no measurable growth defect under our experimental conditions. We then knocked out the *YPet-dnaN* fusion, yielding Δ*dnaN*, and characterized the protein levels by quantifying YPet-DnaN abundance by fluorescence. The experimentally measured fluorescence intensity is consistent with the dilution model (Eq. 1), as expected. (See Fig. 1D.) We therefore conclude that knockout-depletion experiments are consistent with the

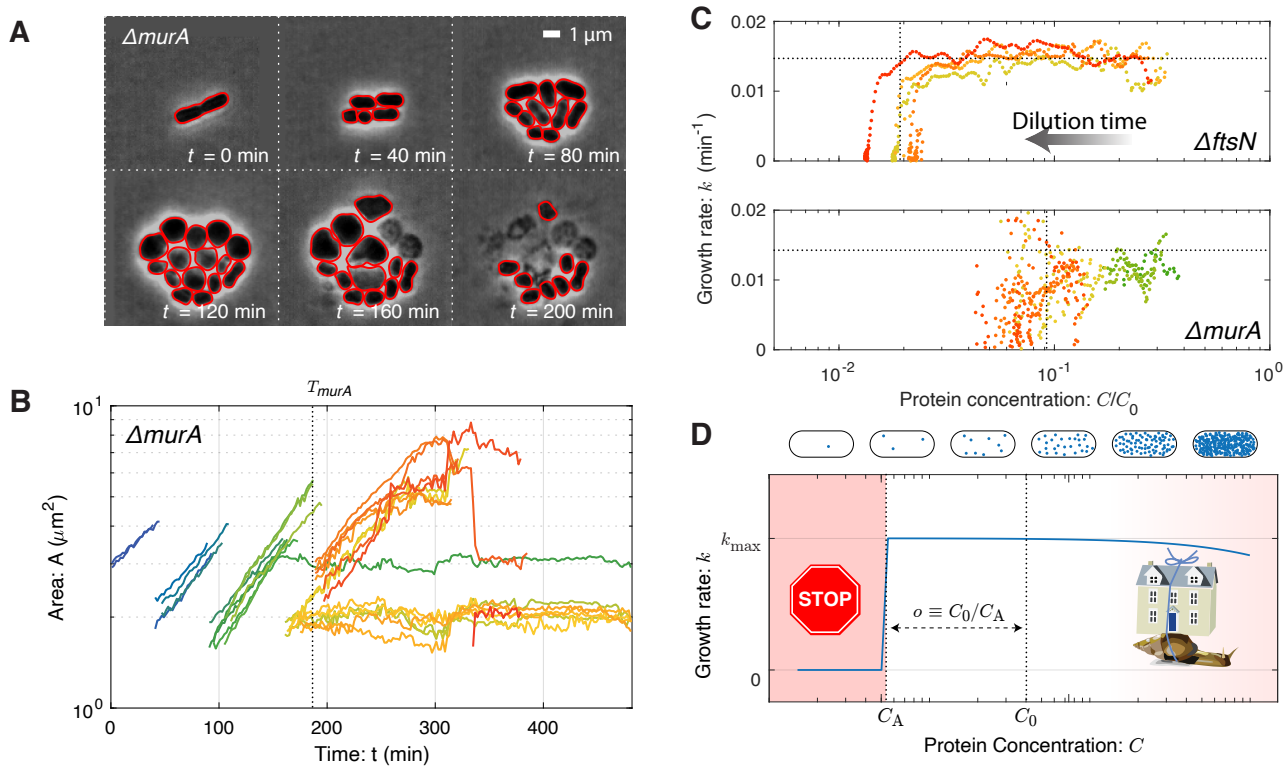

FIG. 2. **The fitness landscape. Panel A: Visualization of growth in a *murA* knockout.** Essential gene *murA* is knocked out at $t = 0$ and cell proliferation is visualized by phase-contrast microscopy. Red outlines represent the Omnipose cell segmentation. Cell proliferation continues for multiple generations after deletion. **Panel B: Quantitative analysis of cell proliferation with single-cell resolution.** Cell area (log scale) as a function of time for the *murA* deletion. The log-slope represents the single-cell growth rate. The vertical dotted line represents the arrest time at which cell growth slows to cell arrest. **Panel C: Growth rate as a function of protein depletion for Δ*ftsN* and Δ*murA*.** In both essential gene deletions, the growth rate is observed to obey the step-like-dependence, transitioning between wild-type growth to arrest at the vertical dotted lines. We define the critical dilution as $o \equiv C_0/C_A$ where C_A is the protein concentration at arrest. **Panel D: The fitness landscape is threshold-like.** Motivated by single-cell growth data, cell fitness is modeled using the Robustness-Load Trade-Off model (RLTO). In the model, there is a metabolic cost of protein expression which favors low expression; however, growth arrests for protein concentration C smaller than the threshold level C_A (red). The relative metabolic cost of overabundance is small relative to the cost of growth arrest due to the large number of proteins synthesized, resulting in a highly asymmetric fitness landscape [11].

experimental design shown schematically in Fig. 1A.

Replication persists during DnaN depletion. A key subhypothesis of the overabundance model for transient growth is that target protein function continues as the target protein abundance is depleted. An alternative hypothesis for transient growth of the Δ*dnaN* strain is a high initial chromosomal copy-number that is partitioned between daughter cells, even after the replication process itself arrests due to target protein depletion [4, 22]. The imaging-based knockout-depletion experiment tests this hypothesis as well. The localization of DnaN is dependent on activity: During ongoing replication, DnaN is localized in puncta corresponding to replisomes, whereas in the absence of active replication, DnaN has diffuse localization [19–21, 23, 24]. During the knockout-depletion experiment, we observed YPet-DnaN puncta persist as the targeted fusion was depleted (Fig. 1DE), consistent with replication activity after dilution. Only after the YPet-DnaN puncta disappear do the cells begin to adopt the Δ*dnaN* phenotype: cell filamentation (Fig. 1BE). We therefore conclude that function (replication) is robust to significant target protein (DnaN) dilution.

Many essential knockouts undergo transient growth. To understand the generic consequences of essential protein depletion, we used the imaging-based knockout-depletion experiments to explore essential genes with a range of functions. We initially targeted four essential genes: the replication initiation regulator gene *dnaA* (movie), the beta-clamp gene *dnaN* (movie), the cell-wall-synthesis gene *murA* (movie), and septation-related gene *ftsN* (movie), as well as a non-essential IS element with no phenotype as a negative control (movie). (Representative frame mosaic images and cytometry appear in Supplementary Material Sec. 4 D.) In each case, transformants continued to pro-

Gene	Annotated gene function:	Message number: μ_m (mRNA molecules/cell cycle)	Log Overabundance: TFNseq Replication $\log_{10} o$	Imaging-based Elongation $\log_{10} o$	Imaging-based Septation $\log_{10} o$	(N_C, N_P)
dnaA	Regulation of replication initiation	30	0.02 ± 0.02	0.7 ± 0.1	0.0 ± 0.2	(4,4)
dnaN	Replication beta sliding clamp	49	1.5 ± 0.1	2.0 ± 3.0	1.4 ± 0.1	(134,8)
ftsN	Essential cell division/septation protein	20	2.6 ± 0.1	1.8 ± 0.2	0.6 ± 0.2	(19,5)
murA	Cell wall precursor synthesis	26	0.7 ± 0.5	1.1 ± 0.1	0.9 ± 0.2	(16,4)

TABLE I. **Measured overabundance for sequencing- versus imaging-based approaches.** The overabundance was determined by both sequencing- and imaging-based approaches. For the imaging-based approach, we show two measurements based on different metrics for arrest: The first is based on the arrest of cell elongation, as defined by Eq. 3, and the second is based on the arrest of the septation process, as visualized by microscopy.

liferate through multiple cell-cycle durations [17] and are therefore consistent with the essential protein overabundance hypothesis. However, in Ref. [17], we were unable to perform a quantitative single-cell analysis of these time-lapse experiments since existing segmentation packages failed to segment the observed morphologies [25]. We therefore developed the *Omnipose* package, which facilitated quantitative analysis of the growth dynamics with single-cell resolution [25]. (See Fig. 2A.)

The fitness landscape is threshold-like. A key input to the RLTO model is the fitness landscape (growth rate) as a function of protein abundance. Omnipose segmentation facilitates the measurement of single-cell growth rates from the time-lapse imaging experiments. We focus first on the single-cell areal growth rate:

$$k(t) = \frac{\mathrm{d}}{\mathrm{d}t} \ln A(t), \tag{2}$$

where $A(t)$ is the area of the cell at time t. This areal growth rate is more convenient than a cell-length based rate since we avoid the necessity of defining cell length for unusual cell morphologies like those observed in the Δ*murA* mutant. Fig. 2B shows representative knockout-depletion dynamics of cell area for the essential-gene target *murA*. The log slope remains constant for multiple generations, consistent with a constant growth rate, even as the gene targeted is depleted over multiple cell cycles. By combining the dilution model (Eq. 1) and the growth rate (Eq. 2), a single knockout-depletion measurement determines the growth rate for a range of protein abundances between wild-type abundance and those realized at growth arrest. This fitness landscape is shown for the MurA and FtsN proteins in Fig. 2C. For all four mutants, the areal growth rate is roughly constant for multiple generations before undergoing a rapid transition to growth arrest.

Protein overabundance. We will define the overabundance as the ratio of protein concentration in wild-type cells (C_0) to the concentration at cell arrest (C_A):

$$o \equiv C_0/C_{\mathrm{A}}, \tag{3}$$

as shown in Fig. 2D. (Supplementary Material Sec. 4 gives a detailed description of the inferred overabundance from single-cell data.) The measured overabundance for the four mutants imaged by microscopy is summarized in Tab. I, using three distinct metrics for growth. We conclude that for each gene, with the exception of *dnaA*, rapid growth continues after the knockout due to the vast overabundance of the target protein.

The RLTO model predicts protein overabundance. The RLTO model explicitly analyzes the trade-off between growth robustness to noise and metabolic load and predicts the optimal central-dogma regulatory principles [11]. Critically, the model incorporates the observed threshold-like dependence of growth rate on protein abundance (Fig. 2CD). The model quantitatively predicts protein overabundance with a signature feature: high-expression genes have low protein overabundance ($o \approx 1$) due to the high metabolic cost of increasing expression and low inherent noise of high expression genes; however, low-expression genes have high overabundance ($o \gg 1$) due to the low metabolic cost of increasing expression and the high inherent noise of low expression genes. (See Supplemental Material Sec. 7 for a more detailed description of the model.)

TFNseq determines overabundances genome-wide. To test the signature expression-dependent overabudance prediction of the RLTO model, we now transition to a genomic-scale analysis. The Manoil lab developed a TFNseq-approach to knockout-depletion experiments for targeting all genes simultaneously in *A. baylyi* [18]. In short: A genomic library was prepared and mutagenized using a transposon carrying the Km^{R} allele. The resulting DNA was then transformed into *A. baylyi*. The transformants were propagated on selective liquid media and fractions collected every two hours from which genomic DNA was extracted. The transposons were then mapped using Tn-seq to generate the relative abundance trajectory for each mutant [18]. (See Fig. 3AB.) We then analyzed each mutant trajectory statistically using three competing growth models: no-effect, sufficiency, and overabundance, using two successive null-hypothesis tests. (See Supplementary Material Sec. 5.) For each mutant i described by the overabundance model, the TFNseq experiment measures a growth arrest time T_i and the corresponding target protein overabundance:

$$o_i = \exp(k_0 T_i), \tag{4}$$

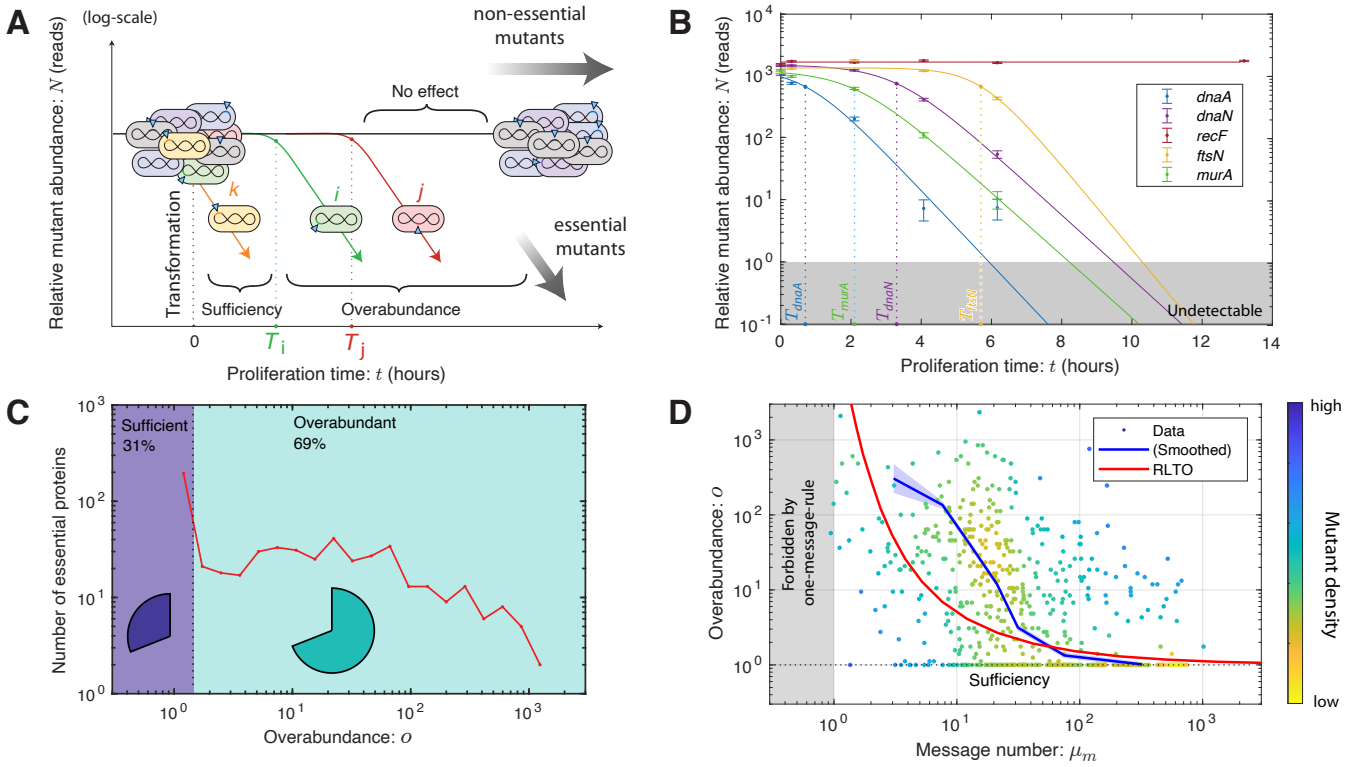

FIG. 3. **A proteome-wide analysis of protein overabundance. Panel A: TFNseq schematic.** A poly-clonal library of knockout mutants is generated by the transformation of ADP1 with DNA mutagenized by transposon insertions. The library is proliferated on selective media and sequential fractions are collected. The relative-abundance trajectories of mutants are determined by mapping transposon insertion sites by sequencing. **Panel B: TFNseq-trajectory analyses for five mutant strains.** Each mutant trajectory is well fit by one of the three trajectory models. As expected, the no-effect model is selected for the non-essential gene *recF*. For the other four essential genes, the overabundance model is selected. The dotted line represents the arrest time for each mutant. **Panel C: Overabundance varies by orders of magnitude between essential proteins.** The protein overabundance is inferred from the arrest time using Eq. 4. Sufficient expression genes have overabundance $o = 1$, while overabundant genes vary from $o > 1$ to very large overabundance ($o > 100$). **Panel D: Overabundance is large for low-expression essential proteins.** The measured message-number-overabundance pairs are shown for essential genes (including estimated gene density.) The smoothed experimental data is shown in blue (with experimental uncertainty.) The RLTO model (red) predicts that overabundance grows rapidly as the transcription level is reduced. The RLTO model qualitatively captures the trend of the data (blue); however, it appears to underestimate the measured overabundance for intermediate expression genes.

where k_0 is the wild-type growth rate. (See Supplementary Material Sec. 5.)

To test the consistency of this TFNseq approach with imaging-based knockout-depletion measurements, we focused first on the analysis of the mutants *dnaA*, *dnaN*, *ftsN*, and *murA*. As shown in Fig. 3B, the trajectories for *dnaA*, *murA*, *ftsN*, and *dnaN* show an unambiguous step-like change in growth dynamics: The no-effect trajectory model (null hypothesis) are rejected with p-values that are below machine precision, and the sufficiency trajectory model is also rejected with $p < 10^{-4}$ for all genes. In Tab. I, we compare protein overabundances determined by imaging- and sequencing-based approaches. These numbers are qualitatively consistent. For instance, the single-cell analysis of *dnaA* mutant shows a nearly immediate phenotype by imaging (*i.e.* cell filamentation). (See Supplementary Sec. 4 D 3.) Likewise, the TFNseq-approach finds an overabundance of 1.0, meaning that protein expression is sufficiency. On the other hand, all three of the other mutants (*murA*, *ftsN*, and *dnaN*) are found to have very large overabundances, and are roughly comparable. Finally, a representative non-essential gene (*e.g. recF*) shows no effect. These results support the use of the TFNseq approach to analyze protein overabundance genome wide.

Many essential proteins have vast overabundance. To determine the protein overabundance genome-wide, we analyzed the knockout-depletion trajectories for all genes in *A. baylyi*. (See Fig. 3BCD.) Our analysis showed that the vast majority (90%) of genes annotated as non-essential were classified as having *no effect* and 10% of non-essential genes had measurable growth defects. (See Supplementary Material Fig. S10.) The most severe growth defect in non-essential annotated genes were observed for the genes *gshA* and *rplI*. For essential genes, all mutants were observed to have growth defects, as

anticipated; however, only 31% of essential proteins were classified as *sufficient*, corresponding to an immediate change in growth rate. Notable genes in this category include ribosomal proteins RpsQ and RpsE, ribonucleotide reductase subunits NrdA and NrdB, and ATP synthase subunits AtpA and AtpD. However, as predicted by the RLTO model, the majority of essential proteins (69%), were classified as *overabundant*, meaning that they required significant dilution before a growth rate change was detected. Fig. 3D shows a histogram of essential gene overabundances.

Low-expression genes are highly overabundant. To understand the overall significance of overabundance in a typical biological process, we determined the median essential protein overabundance: 7-fold. To understand the significance of overabundance from the perspective of the metabolic load, we also determine the mean protein overabundance, weighted by the expression level: 1.6-fold. These two superficially-conflicting statistics emphasize a key predicted regulatory principle: overabundance is high for low-abundance proteins; however, it is close to unity for the high-abundance proteins, which constitute the dominant contribution to the metabolic load.

To explicitly test the predicted relation between protein expression and overabundance, we measured the relative abundance of mRNA messages by RNA-Seq for exponentially growing *A. baylyi* cells. (See Supplementary Material Sec. 6 C.) We computed the message number (transcripts per gene per cell cycle) for each essential gene. (See Supplementary Material Sec. 6 B.) Fig. S9 compares measured message numbers and overabundances for all essential genes with the prediction of the RLTO model.

As predicted, the data shows a clear trend of decreasing overabundance with increasing message number (Fig. 3D). To quantitatively capture this trend, we computed the mean log overabundance over windows of message number (blue curves) to compare the data cloud to the RLTO model predictions. With very few exceptions, high expression genes have extremely low overabundance. At the other extreme, low expression genes typically have large to very large overabundance as shown by the sharp up-turn of the blue curve as the message number approaches the one-message-rule threshold, a lower threshold on transcription that we recently proposed [11].

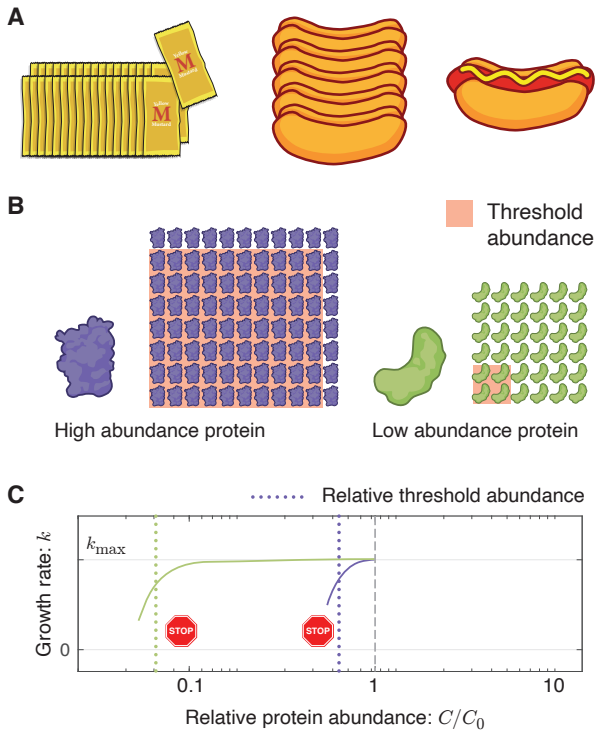

FIG. 4. **How rate-limited kinetics shapes the fitness landscape. Panel A: An analogy for rate-limited kinetics.** The number of sausage sandwiches assembled from the pictured ingredients is limited by a single ingredient, the sausages. A depletion of either bun or mustard abundance does not immediately affect the sandwich number. **Panel B: Protein abundance and threshold.** Two essential protein species with different abundances are pictured schematically. The threshold abundance at which each protein becomes limiting is represented by the pink square and the total cellular abundance is represented by the protein array. **Panel C: Emergent fitness landscape.** A schematic model of the growth rate versus relative protein abundance is shown for the two protein species. The RLTO model predicts that low-abundance proteins (green) have high overabundance, which leads to significant insensitivity to protein depletion. High-abundance protein (purple) are predicted to have small overabundance leads to high sensitivity to protein dilution. The growth rate rapidly decreases with concentration once a species becomes limiting.

DISCUSSION

The shape of the fitness landscape. Despite some large-scale measurements [8, 9, 26, 27], fundamental questions remain about the structure of the fitness landscape and its rationale [7]. Our measurements reveal that most (69%) essential proteins show a step-like transition between wild-type and arrested growth below a critical threshold protein abundance. Although asymmetric landscapes have been observed previously (*e.g.* [3, 26]), the knockout depletion approach is expected to yield more quantitative results. For instance, the use of either CRISPRi (*e.g.* [8]) or inducible promoters (*e.g.* [3]) significantly increases the cell-to-cell variation in protein abundance [16, 28], obscuring the features of the fitness landscape. The sharpness of the protein-abundance threshold is manifest in the single-cell analysis where the progeny begin from a common pool of protein in a single progenitor cell and are therefore not subject to noise (*e.g.* Fig. 2C).

The rationale for a threshold abundance. The observed threshold-like dependence can be rationalized in terms of chemical kinetics: If the protein target is not a rate-limiting reactant in an essential cellular process, then its depletion has no effect on the rate [11, 29]. See Fig. 4. We explicitly demonstrate protein function (*i.e.* replication) is robust to an order-of-magnitude depletion of replisome protein DnaN; however, for most proteins, we must infer this picture from the growth rate.

The rationale for overabundance. Rate-limiting kinetics does not in itself predict vast protein overabundance. The RLTO model predicts that this feature of the fitness landscape is a consequence of a balance between (i) the metabolic cost of protein expression, which favors minimizing protein abundance, and (ii) robustness to the noise in gene expression [30, 31]. The model predicts expression-dependent protein overabundance: large overabundance for low-abundance proteins and small overabundance for high-abundance proteins [11]. We show that this signature prediction is observed (Fig. 3D). In spite of predicting the genomic-scale trend, there are some significant outliers. We discuss their significance as well as evidence for the conservation of overabundance in Supplementary Material Sec. 1

Biological implications. Many important proposals have been made about the biological implications of noise [32]. Our work reveals that noise acts to inflate the optimal expression levels of low-expression proteins and, as a result, significantly increases the metabolic budget for protein, which constitutes 50-60% of the dry mass of the cell [4]. We believe this increased protein budget has cellular-scale implications. For instance, in stress response and stationary phase, the presence of a significant reservoir of overabundant protein provides critical resources, via protein catabolism, to facilitate the adaptation to changing conditions [33, 34]. Protein overabundance may have important implications for individual biological processes as well, including determining which proteins and cellular processes make attractive targets for small molecule inhibitors (*e.g.* antibiotics) [27]. Since overabundance defines the fold-depletion in protein activity required to achieve growth arrest, high-overabundance proteins are predicted to be extremely difficult targets for inhibition.

Conclusion. By combining imaging-, genomic-, and modeling-based approaches, we provide a both a quantitative measurement of the fitness landscape for all essential proteins as well as a clear qualitative and conceptual understanding of the rationale for the observed fitness landscape. The RLTO model fundamentally reshapes our understanding of the rationale for protein abundance. The model predicts, and experiments confirm, that low-abundance proteins are expressed in vast excess of what is required for growth. Despite the limitations of the experiments, the predicted trend is clearly resolved both at a genomic-scale, using sequencing-based approaches, as well as at the single-cell scale, as observed by microscopy. The rationale for the overabundance strategy is intuitive: Growth requires the robust expression of between five to six hundred distinct proteins. The cell contends with this extraordinary complex regulatory challenge by keeping all but the highest-abundance proteins in vast excess.

Data availability. We include source data files and sequencing data from RNA-Seq experiments to quantify transcription levels. Gene Expression Omnibus (GEO) accession number TBA.

Acknowledgments. The authors would like to thank B. Traxler, A. Nourmohammad, J. Mougous, M. Cosentino-Lagomarsino, S. van Teeffelen, C. Manoil, L. Gallagher, J. Bailey, J. Mannik, and S. Murray. H.J.C., T.W.L., D.H., and P.A.W. were supported by NIH grant R01-GM128191 and NSF grant GR046955. K.J.C. was supported by the Molecular Biophysics Training Program (NIH grant T32GM008268). W.R.W. was supported by NIH grant R01-AI150041.

Author contributions: H.J.C., T.W.L., K.J.C., D.H., and P.A.W. conceived the research. H.J.C., W.R.W., and P.A.W. performed the experiments. H.J.C., T.W.L., K.J.C., and P.A.W. performed the analysis. H.J.C., T.W.L., D.H., and P.A.W. wrote the paper.

Competing interests: The authors declare no competing interests.

Supplementary Materials

Supplementary text, Materials and Methods
Figs. S1 to S11.
Tables S1 to S3.
References 35 to 56.
Data files S1 to S9.
Movies S1 to S12.

[1] B. Alberts, *et al.*, *Molecular Biology of the Cell* (Garland, 2002), fourth edn.
[2] E. Dekel, U. Alon, *Nature* **436**, 588 (2005).
[3] J.-B. Lalanne, D. J. Parker, G.-W. Li, *Mol Syst Biol* **17**, e10302 (2021).
[4] J. W. Lengeler, G. Drews, H. G. Schlegel, eds., *Biology of the Prokaryotes* (Georg Thieme Verlag, Rüdigerstrasse 14, D-70469 Stuttgart, Germany, 1998).
[5] M. Kafri, E. Metzl-Raz, F. Jonas, N. Barkai, *FEMS Yeast Res* **16** (2016).
[6] J. Hausser, A. Mayo, L. Keren, U. Alon, *Nat Commun* **10**, 68 (2019).
[7] N. M. Belliveau, *et al.*, *Cell Syst* **12**, 924 (2021).
[8] J. M. Peters, *et al.*, *Cell* **165**, 1493 (2016).

[9] S. Donati, *et al.*, *Cell Syst* **12**, 56 (2021).
[10] T. Baba, H.-C. Huan, K. Datsenko, B. L. Wanner, H. Mori, *Methods Mol Biol* **416**, 183 (2008).
[11] T. W. Lo, H. J. Choi, D. Huang, P. A. Wiggins, *Science Advances (in press).* (2024).
[12] K. E. McGinness, T. A. Baker, R. T. Sauer, *Mol Cell* **22**, 701 (2006).
[13] J. H. Davis, T. A. Baker, R. T. Sauer, *ACS Chem Biol* **6**, 1205 (2011).
[14] D. E. Cameron, J. J. Collins, *Nat Biotechnol* **32**, 1276 (2014).
[15] X. Liu, *et al.*, *Mol Syst Biol* **13**, 931 (2017).
[16] S. N. J. Franks, R. Heon-Roberts, B. J. Ryan, *Biochem Soc Trans* **52**, 539 (2024).
[17] J. Bailey, *et al.*, *PLoS Genet* **15**, e1008195 (2019).
[18] L. A. Gallagher, J. Bailey, C. Manoil, *Proc Natl Acad Sci U S A* **117**, 18010 (2020).
[19] S. M. Mangiameli, C. N. Merrikh, P. A. Wiggins, H. Merrikh, *Elife* **6** (2017).
[20] S. M. Mangiameli, B. T. Veit, H. Merrikh, P. A. Wiggins, *PLoS Genet* **13**, e1006582 (2017).
[21] S. M. Mangiameli, J. A. Cass, H. Merrikh, P. A. Wiggins, *Curr Genet* **64**, 1029 (2018).
[22] G. M. Cooper, *The Cell: A Molecular Approach. 2nd edition* (Sinauer Associates 2000, 2000).
[23] K. P. Lemon, A. D. Grossman, *Science* **282**, 1516 (1998).
[24] R. Reyes-Lamothe, D. J. Sherratt, M. C. Leake, *Science* **328**, 498 (2010).
[25] K. J. Cutler, *et al.*, *Nat Methods* **19**, 1438 (2022).
[26] L. Keren, *et al.*, *Cell* **166**, 1282 (2016).
[27] B. Bosch, *et al.*, *Cell* **184**, 4579 (2021).
[28] A. Novick, M. Weiner, *Proc Natl Acad Sci U S A* **43**, 553 (1957).
[29] D. L. Nelson, M. M. Cox, *Lehninger principles of biochemistry* (W.H. Freeman, New York, NY, 2017), 7th edn.
[30] J. Paulsson, M. Ehrenberg, *Phys Rev Lett* **84**, 5447 (2000).
[31] N. Friedman, L. Cai, X. S. Xie, *Phys Rev Lett* **97**, 168302 (2006).
[32] J. M. Raser, E. K. O'Shea, *Science* **309**, 2010 (2005).
[33] D. Weichart, N. Querfurth, M. Dreger, R. Hengge-Aronis, *J Bacteriol* **185**, 115 (2003).
[34] A. L. Goldberg, A. C. St John, *Annu Rev Biochem* **45**, 747 (1976).
[35] G. Lambert, E. Kussell, *PLoS Genet* **10**, e1004556 (2014).
[36] S. Autret, A. Levine, I. B. Holland, S. J. Séror, *Biochimie* **79**, 549 (1997).
[37] B. Bolognesi, B. Lehner, *Elife* **7** (2018).
[38] I. P. Menikpurage, K. Woo, P. E. Mera, *Front Microbiol* **12**, 662317 (2021).
[39] D. Nevozhay, R. M. Adams, K. F. Murphy, K. Josic, G. Balázsi, *Proc Natl Acad Sci U S A* **106**, 5123 (2009).
[40] V. Barbe, *et al.*, *Nucleic Acids Res* **32**, 5766 (2004).
[41] J. Miller, *Experiments in Molecular Genetics*, Bacterial genetics - E.: Coli (Cold Spring Harbor Laboratory, 1972).
[42] A. C. Chang, S. N. Cohen, *J Bacteriol* **134**, 1141 (1978).
[43] T. W. Lo, *et al.*, SuperSegger-Omnipose (SuperSegger 2) GitHub (2024).
[44] D. R. Cox, D. V. Hinkley, *Theoretical Statistics* (Chapman & Hall, London, England, 1974).
[45] K. P. Burnham, D. R. Anderson, *Socialogical Methods & Research* **33**, 261 (2004).
[46] P. D. Karp, *et al.*, *EcoSal Plus* **11**, eesp (2023).
[47] P. H. Culviner, C. K. Guegler, M. T. Laub, *mBio* **11** (2020).
[48] Y. Taniguchi, *et al.*, *Science* **329**, 533 (2010).
[49] F. Crick, *Nature* **227**, 561 (1970).
[50] J. L. Hargrove, F. H. Schmidt, *FASEB J* **3**, 2360 (1989).
[51] A. L. Koch, H. R. Levy, *J Biol Chem* **217**, 947 (1955).
[52] M. Martin-Perez, J. Villén, *Cell Syst* **5**, 283 (2017).
[53] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, T. Hwa, *Science* **330**, 1099 (2010).
[54] S. S. Wilks, *The Annals of Mathematical Statistics.* **9**, 60 (1938).
[55] G. Casella, R. Berger, *Statistical Inference* (Duxbury Resource Center, 2001).
[56] E. W. Weisstein, *CRC Encyclopedia of Mathematics* (Chapman & Hall/CRC, 2009).

Supplementary material: Protein overabundance is driven by growth robustness

H. James Choi,[1] Teresa W. Lo,[1] Kevin J. Cutler,[1] Dean Huang,[1] W. Ryan Will,[2] and Paul A. Wiggins[1,3,4,*]

[1]*Department of Physics, University of Washington, Seattle, Washington 98195, USA*
[2]*Department of Laboratory Medicine and Pathology, University of Washington, Seattle, Washington 98195, USA*
[3]*Department of Microbiology, University of Washington, Seattle, Washington 98195, USA*
[4]*Department of Bioengineering, University of Washington, Seattle, Washington 98195, USA*

CONTENTS

1. Supplementary Discussion 10
   A. Discussion: Are non-essential proteins overabundant? 10
   B. Discussion: Limitations of knockout-depletion experiments. 10
   C. Discussion: Limitations of the RLTO model. 11
   D. Discussion: Is protein overabundance conserved? 11

2. *Acinetobacter baylyi* strains, manipulation, and culturing 11
   A. Methods: Construction of deletion mutations 11
   B. Methods: *A. baylyi* transformation protocol 11
   C. Methods: Construction of YPet-*dnaN* fusion strain 12

3. Growth models for knockout-depletion experiments 13
      1. No-Effect model 13
      2. Sufficiency model 13
      3. Overabundance model 13

4. Imaging-based knockout-depletion experiments 14
   A. Methods: Experimental protocol 14
      1. Cell preparation for knockout-depletion experiments. 14
      2. Sample/slide preparation. 14
      3. Microscopy. 14
      4. Image processing (cell segmentation) pipeline. 14
   B. Methods: Cytometry data analyses 15
      1. Accessing imaging-based cell cytometry data 15
      2. Protein abundance analysis 15
      3. Areal growth analysis 15
   C. Cell-number growth analysis 16
   D. Results: Imaging-based analyses 17
      1. Some progenitors have heterogenic progeny 17
      2. Wild-type imaging-based analyses 18
      3. *dnaA* imaging-based analysis 18
      4. *dnaN* imaging-based analysis 18
      5. *ftsN* imaging-based analysis 20
      6. *murA* imaging-based analysis 21

5. Statistical analysis of TFNseq trajectories 23
   A. Methods: Time correction 23
   B. Methods: Defining the likelihood 23
   C. Methods: Analysis of overabundance for different Gene Ontologies (GO) 23
   D. Results: Toxicity reduces overabundance. 23
   E. Results: Regulation reduces overabundance. 23

* pwiggins@uw.edu

F. Analysis of overabundance for different gene regulatory controls 24

6. RNA-Seq analysis of transcription 25
A. Methods: RNA-Seq protocol 25
1. RNA extraction 25
2. rRNA Depletion 25
3. Library prep and sequencing 26
B. Methods: Computation of message number 26
C. Results: Comparison of *A. baylyi* and *E. coli* gene expression 27

7. Robustness Load Trade-Off (RLTO) Model 27
A. Methods: Detailed description of the noise model 28
1. Stochastic kinetic model for the central dogma. 28
2. Statistical model for protein abundance. 28
B. Methods: Summary of the RLTO model fitness model 28
1. Metabolic load in the RLTO model 28
2. Growth rate with stochastic arrest 28
3. RLTO growth rate 29
4. Single-gene equation 29
5. Optimization of transcription for bacteria 29
6. Estimate of the relative load in bacterial cells 29
C. Results: The fitness landscape of the RLTO model is highly asymmetric 29
D. Results: The RLTO model predicts overabundance is optimal for low-expression proteins 30
E. Discussion: Does the detailed form of the fitness landscape affect RLTO predictions? 30

8. Methods: Statistical procedures 31
A. Maximum Likelihood Estimation 31
B. Parametric uncertainty estimates 31
C. Null-hypothesis-testing approach 31

9. Distributions and conventions 31
A. Gamma distribution conventions 31
B. Chi-squared distribution conventions 32

10. Description of supplementary data 32
A. Data Tables 32
B. Annotated sequences 32
C. Supplemental movies 32

1. SUPPLEMENTARY DISCUSSION

A. Discussion: Are non-essential proteins overabundant?

We have focused our analysis on essential genes in the model organism *A. baylyi* and demonstrated that most essential proteins are overabundant. To what extent is this mechanism generic to non-essential proteins? Several arguments support a generic applicability to non-essential genes. Our modeling suggests asymmetry rather than explicit growth arrest is the mathematical rationale for the optimality of overabundance [11]. We therefore predict that all proteins that increase cell fitness, not just essential proteins, will be overabundant. In addition, it is important to emphasize that the annotation of genes as *essential* is contextual. For instance, for *E. coli* proliferation on lactose, the gene *lacZ* is essential, although non-essential for other carbon sources. As a result, we predict that when expressed, LacZ should be overabundant, consistent with observation [35]. Finally, The RLTO model also correctly predicts the balance between transcription and translation for all genes, not just essential genes, in eukaryotic cells, suggesting that it should generalize to nonessential genes as well [11].

B. Discussion: Limitations of knockout-depletion experiments.

In spite of the success of the RLTO model in predicting the genomic-scale overabundance trend, there are many significant outliers from this prediction. In considering their significance, it is important to emphasize the flaws both with the knockout-depletion experiments, as well as the RLTO model. With respect to the experiments, the mechanism of growth arrest plays an important role in

determining which growth metric most accurately determines the arrest time. Consider the three arrest times measured for the septation-related essential gene *ftsN* in Tab. I. Due to the absence of strict cell-cycle checkpoints in the bacterial cell, the arrest of the septation process does not immediately arrest cell elongation and replication [36]. Growth arrest is therefore detected first by the cell-number metric, directly dependent on septation, and later in the other two metrics.

C. Discussion: Limitations of the RLTO model.

Likewise, the RLTO model itself has some important limitations. For instance, the model assumes that the dominant contribution to the fitness cost of protein overabundance is metabolic load rather than toxicity [37]. We have already investigated the consequences of a toxicity-based increase in cost from a model perspective: The qualitative behavior of the model is unchanged; however, the optimal overabundance is reduced by toxicity [11]. Motivated by this prediction, we tested whether two classes of proteins, ATPases and enzymes [37], that are expected to exhibit toxicity, have lower overabundance. In Supplementary Material Sec. 5 D, we demonstrate that this predicted trend is observed. Similarly, the low overabundance of DnaA also provides a second clue about a class of genes that is predicted to have low overabundance: *dnaA* is negatively autoregulated [38]. Tight regulation can reduce noise, and therefore we hypothesize that tight regulation, and auto-regulation in particular [39], could therefore reduce the optimal overabundance [11]. In Supplementary Material Sec. 5 E, we demonstrate that this predicted trend is also observed in the data. The putative importance both gene-product toxicity and gene regulation in determining the optimal overabundance emphasizes that the RLTO model describes only part of the biology that determines optimal expression levels.

D. Discussion: Is protein overabundance conserved?

To what extent is the overabundance of essential genes a conserved mechanism from bacteria, to single-cell eukaryotes, to multicellular organisms? As we emphasized above, CRISPRi protein depletions in a wide range of model organisms appear to be consistent with the overabundance hypothesis [8, 12–15]. Furthermore, we have demonstrated elsewhere that the RLTO model also predicts two other principles of central dogma function (the one-message-rule and load balancing in protein expression) that are observed in eukaryotic cells [11]. We therefore expect to observe the overabundance strategy in all organisms for low-expression genes [11].

2. *ACINETOBACTER BAYLYI* STRAINS, MANIPULATION, AND CULTURING

Mutant strains were derived from *Acinetobacter baylyi* ADP1 (MAY101) (the gift of C. Manoil) [40]. Growth media were LB and M9, a minimal-succinate M9 medium [41], supplemented with 15 mM sodium succinate, 2 mM magnesium sulfate, 0.1 mM calcium chloride and 1–3 µM ferrous sulfate (from sterile 5mM stock, made fresh at least once a month). For selective growth, media was supplemented with kanamycin at 20 µg/mL. Cultures were grown at 30°C.

The strains used in the study are summarized in Tab. S1.

A. Methods: Construction of deletion mutations

We generated deletion mutants by transformation of linear DNA fragments, constructed by PCR using extension overlap [17]. A homologous overlap of ∼2 kb flanking target genes was created that either directly joined (for marker-free deletions) or flanked a kanamycin resistance cassette (for kan-selectable deletions). Unmarked deletions were in-frame. Kan deletions were constructed from the *kan* gene from plasmid pACYC177 [42], in an orientation matching the deleted gene [17]. PCR reactions were performed using Q5 Polymerase (New England Biolabs) or Phusion HF polymerase (New England Biolabs) and DNA fragments were purified using Qiaquick columns (Qiagen) before transformation.

B. Methods: *A. baylyi* transformation protocol

DNA fragments were transformed into *A. baylyi* cultures prepared as follows. Cultures were grown overnight in minimal-succinate M9 media with 1 µM ferrous sulfate. The culture was then back diluted 1:5 into fresh medium and grown one hour, shaking at 30°C. The DNA fragment was added at 1 µg/mL, followed by incubation for 2.5 - 3 hours with shaking, and then plated on selective (for *kan*-deletion cassettes) or non-selective media (for marker-free casettes). Marker-free deletion mutants were identified by screening single colonies by PCR using primers flanking targeted genes. Essential gene kan-marked deletion mutations were selected by plating on protective medium supplemented with 20 µg/mL kanamycin. All unmarked and the marked non-essential deletion mutations were verified by PCR. For essential gene deletions, 0.1–1% of the cells were transformed, forming microcolonies of cells carrying the deletion.

Short name:	Lab number:	Organism:	Genotype:	Source:	Stability:	Selectable marker:	Description:
wild-type	#1139	*A. baylyi*	ADP1	Ref. [40]	Stable	—	Wild-type strain.
YdnaN	#1545	*A. baylyi*	ADP1 *dnaN*::*YPet-dnaN*	This study.	Stable	—	The beta clamp (DnaN) is replaced by the fluorescent fusion *YPet-dnaN* at the endogenous locus.
Δ*IS*	N.A.†	*A. baylyi*	ADP1 *dnaN*::*YPet-dnaN* *ACIA0320-0321*::*kan*	This study.	Stable	Km^R	This is a control strain where non-essential genes, corresponding to an IS element, are knocked out from the YdnaN strain. Even through the strain is stable, it is re-transformed in each knockout-depletion experiment. Transformed strain has a wild-type growth phenotype.
Δ*dnaA*	N.A.†	*A. baylyi*	ADP1 *dnaA*::*kan*	This study.	Unstable	Km^R	DnaA is an essential cell-cycle regulator. This strain must be re-transformed in each knockout-depletion experiment. Transformed strain is wild-type.
Δ*dnaN*	N.A.†	*A. baylyi*	ADP1 *dnaN*::*kan*	This study.	Unstable	Km^R	The beta clamp (DnaN) is an essential component of the replisome. This strain must be re-transformed in each knockout-depletion experiment. Transformed strain is wild-type.
Δ*YdnaN*	N.A.†	*A. baylyi*	ADP1 *YdnaN*::*kan*	This study.	Unstable	Km^R	The beta clamp (DnaN) is an essential component of the replisome. This strain must be re-transformed in each knockout-depletion experiment. Transformed strain is YdnaN (not wild-type).
Δ*murA*	N.A.†	*A. baylyi*	ADP1 *murA*::*kan*	This study.	Unstable	Km^R	The gene product of *murA* is UDP-N-acetylglucosamine 1-carboxyvinyltransferase, an essential protein in synthesizing the precursors of cell wall synthesis. This strain must be re-transformed in each knockout-depletion experiment. Transformed strain is wild-type.
Δ*ftsN*	N.A.†	*A. baylyi*	ADP1 *ftsN*::*kan*	This study.	Unstable	Km^R	The gene product of *ftsN* is essential cell division protein FtsN. This strain must be re-transformed in each knockout-depletion experiment. Transformed strain is wild-type.

TABLE S1. **Summary of strains used in this study.** The *short name* describes the nomenclature of the strains as described in the text. †Strain re-created by transformations in each knockout-depletion experiment are not stable and therefore are not assigned a *lab strain number* and, due to their instability, cannot be distributed.

C. Methods: Construction of YPet-*dnaN* fusion strain

In previous work in *Escherichia coli* and *Bacillus subtilis*, we visualized fluorescent fusions to the beta sliding clamp (*dnaN*) to study replication [19–21]. The DnaN protein imaging is a convenient tool for studying replication due to its relatively high abundance and the change in its localization, from diffuse (non-replicating cells) to punctate (replicating cells), which serves as a convenient reporter of activity.

To construct a fluorescent fusion to the *A. baylyi* DnaN protein with a high probability of success, we used the exact same fluorescent protein and linker to that which R. Reyes-Lamothe had used to construct the *E. coli* fusion used in our previous work [24]. In this approach, we inserted the YPet-linker cassette at the 5′ end of the gene. Since the transformation efficiency of *A. baylyi* is so high, we constructed a marker-free fusion. We screened colonies by both PCR and fluorescence local-

ization. We then sequenced the mutant *YdnaN* strain to confirm that the desired construct was achieved. (We provide a supplemental file with the sequence.) Like the original *E. coli* strain, no growth phenotype is observed under experimental conditions.

3. GROWTH MODELS FOR KNOCKOUT-DEPLETION EXPERIMENTS

To quantitatively analyze growth in knockout-depletion experiments, we define three nested growth models: (i) *No-Effect*, (ii) *Sufficiency*, and (iii) *Overabundance* models. In our statistical analysis, we will initially treat the No-Effect model as the null hypothesis and the Sufficiency model as the alternative hypothesis. If the null hypothesis is rejected, we will then adopt the Sufficiency model as the null hypothesis and adopt the Overabundance model as the alternative hypothesis.

1. *No-Effect model*

In the *No-effect model*, the mutant has no effect on the growth rate. The abundance in a log culture will therefore be:

$$N_{\mathrm{N}}(t; N_0) = N_0 e^{k_0 t}, \tag{S1}$$

where k_0 is the wild-type growth rate and N_0 is the abundance at $t = 0$.

For modeling the TFNseq trajectories, it is the relative abundance that is measured and we therefore normalize by wild-type growth of the culture, resulting in the relative abundance:

$$\eta_{\mathrm{N}}(t; \eta_0) = \eta_0, \tag{S2}$$

where η_0 represents the initial relative abundance. (The relative abundance of the No-effect model is independent of t.) Both the abundance N_{N} and relative abundance η_{N} are plotted in Fig. S1. Both models depend on a single model parameter and are therefore dimension 1.

2. *Sufficiency model*

In the *Sufficiency model*, we model the effect of the mutant as immediate. The cell number is assumed to grow at a new unknown rate:

$$N_{\mathrm{S}}(t; N_0, k) = N_0 e^{kt}, \tag{S3}$$

where k is the new growth rate and N_0 is the number of the mutants at $t = 0$. For modeling the TFNseq trajectories, it is the relative abundance that is measured, and we therefore normalize by wild-type growth of the culture, resulting in the relative abundance:

$$\eta_{\mathrm{S}}(t; \eta_0, \Delta k) = \eta_0 e^{-\Delta k t}, \tag{S4}$$

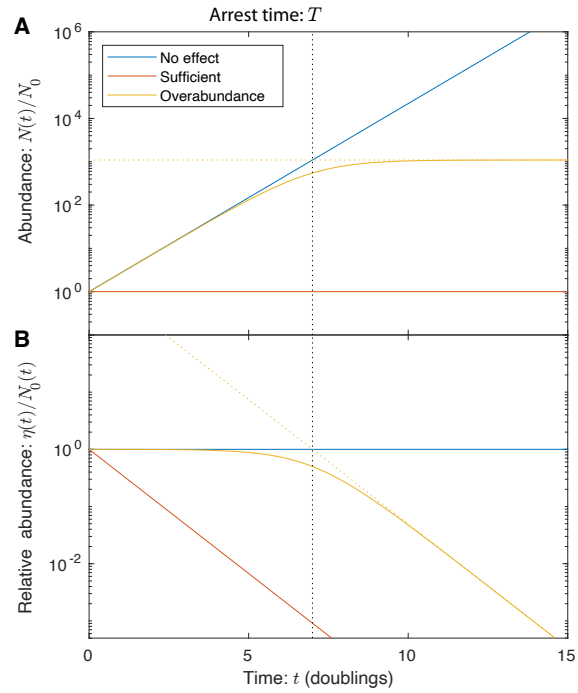

FIG. S1. **Panel A: Mutant abundances for trajectory models.** Mutants described by the *No-effect model* (blue) grow at the wild-type growth rate. Mutants described by the *Sufficient trajectory model* (red) show an immediate change in growth rate after transformation. Mutants described by the *Overabundant trajectory model* (yellow) grow to the arrest time T (black dotted line) with the wild-type growth rate, before adopting a reduced growth rate of $k = 0$. **Panel B: Relative mutant abundances for trajectory models.** Same as above, but abundances are renormalized by wild-type growth.

where $\Delta k \equiv k_0 - k$ is the growth rate reduction of the mutant relative to the wild-type growth rate. Both the abundance N_{S} and relative abundance η_{S} are plotted in Fig. S1. Both models depend on two model parameters and are therefore dimension 2. Note that we might naïvely expect $k = 0$ for essential genes; however, we expect some transient growth due to residual protein levels, and these transients will dominate the fit.

3. *Overabundance model*

In the *Overabundance model*, we model the effect of the mutant with a delayed arrest time, T: the transient growth duration as protein dilutes to the threshold level. For short times, the mutant growth with a wild-type rate:

$$N_{\mathrm{O}} = N_0 e^{k_0 t}, \tag{S5}$$

however, at long times we expect growth with a new unknown growth rate k:

$$N_{\mathrm{O}} = N_0' e^{kt}. \tag{S6}$$

We initially attempted to use a piecewise function to join these two limits; however, the sparsity of the data and discontinuous slope at the boundary appeared to give rise to fitting artifacts. In addition, the cell-to-cell variation in protein expression smooths out the transition time. To fix these shortcomings, we adopted an empirical formula with the correct limits, but with a smooth transition at $t = T$:

$$N_{\mathrm{O}}(t; N_0, k, T) = N_0 e^{k_0 t} \frac{e^{\Delta k T}+1}{e^{\Delta k T}+e^{\Delta k t}}, \tag{S7}$$

where $\Delta k \equiv k_0 - k$ is the loss in growth rate due to the mutation. Modeling the TFNseq trajectories, it is the relative abundance that is measured, and we therefore normalize by wild-type growth of the culture, resulting in the relative abundance:

$$\eta_{\mathrm{O}}(t; \eta_0, k, T) = \eta_0 \frac{e^{\Delta k T}+1}{e^{\Delta k T}+e^{\Delta k t}}, \tag{S8}$$

where $\Delta k \equiv k_0 - k$ is the growth rate reduction of the mutant relative to the wild-type growth rate. Both the abundance N_{O} and relative abundance η_{O} are plotted in Fig. S1. Both models depend on three model parameters and are therefore dimension 3. Note that we might naïvely expect $k = 0$ for essential genes; however, we expect some transient growth due to residual protein levels, and these transients will dominate the fit.

4. IMAGING-BASED KNOCKOUT-DEPLETION EXPERIMENTS

A. Methods: Experimental protocol

For single-cell imaging-based analyses, cells were imaged proliferating in M9 media supplemented with 2% low-melt agarose, and in most cases, kanamycin at 20 µg/mL.

1. *Cell preparation for knockout-depletion experiments.*

The transformation protocol described above was modified as follows: after the 2.5-3 hr incubation with DNA, cells were immediately spotted on selective media pads for imaging. In the knockout-depletion experiments, cells are transformed with knockout cassettes which recombine into the genome, resulting in Km^{R} knockout strains. If transformed cells are transferred to Km^{+} media too quickly, the competent cells do not have sufficient time to integrate the *kan* cassette before growth arrest. If cells are transferred too late, essential proteins are depleted before imaging begins. How do we know transformants after 2.5-3 hr outgrowth are at their initial stages of transient growth? With the 2.5-3 h outgrowth period, many cells still grow slowly (compared to log phase growth) for 10-15 min consistent with the expression of the kanamycin phosphotransferase (the gene product of the *kan* gene) not having reached a sufficiently high level to achieve a resistance phenotype. Furthermore, a significant number of heterogenic progenitors were observed. The presence of these heterogenic progenitor cells is consistent with the 2.5 h outgrowth period representing the typical recombination time for transformants. (See Sec. 4 D 1 for a discussion of heterogenic progenitors.)

2. *Sample/slide preparation.*

Thin pads were fabricated by melting the agarose (Invitrogen UltraPure™ LMP Agarose) and casting it between two slides with two layers of lab tape used as a shim to set the height. After the pad solidified (roughly 10 min), the top slide was carefully removed, and a razor blade was used to trim the pad to form a small square that could be covered with a #1.5 coverslip. For *E. coli* imaging, we typically use a pad that matches the size of the coverslip; however, *for A. baylyi imaging, we trim the pad so it is less than 1 cm in width. This added space allows aerobic growth to continue over multiple hours.* Finally, the coverslip is sealed using a hot glue gun.

3. *Microscopy.*

The samples were imaged using a Nikon Eclipse Ti microscope in phase contrast and fluorescence. We imaged through a Nikon 60× 1.4 NA Phase contrast objective onto a sCMOS camera (Andor Neo). An environmental chamber maintained the sample at 30°C during imaging. For phase imaging, a frame rate of 1 frame / 2 min was used; however, for combined phase and fluorescence imaging we reduced the frame rate to 1 frame / 3 min and 1 frame / 9 min to help reduce bleaching and phototoxicity. (The slowest frame rate was used to resolve the dim YPet-DnaN foci as the protein levels were depleted.) Typically, multiple ($\sim$ 10) fields of view were captured simultaneously in each experiment. For fluorescence-based analysis, we mixed in wild-type cells, in addition to fluorescent-fusion cells (1:2), to determine the autofluorescence levels in each experiment.

4. *Image processing (cell segmentation) pipeline.*

Cell images were processed using the *SuperSegger-Omnipose* package [43] by running the `processExp` command with default settings. Most of the analysis described in the paper was performed from the `clist.mat` files generated for each dataset.

B. Methods: Cytometry data analyses

Imaging-based analysis for protein overabundance was carried out by assessing the transient cell area growth and septation. The three different single-cell analysis approaches are explained below: *protein abundance*, *area*, and *number analysis*.

1. *Accessing imaging-based cell cytometry data*

Most of the analysis described in the paper was performed using the `clist.mat` files generated for each dataset by the *SuperSegger-Omnipose* package. In particular, the `data3D` field provides time-dependent cell descriptors for each cell in each frame, including `Rod Length`, `Area`, `Fluor1 sum`, and `Fluor1 background`. These descriptors were the input for our analyses. To characterize cell progeny area of fluorescence, we would generate cell lineage trees and cell progeny IDs using the `getFamily` command and then sum fluorescence or area over all progeny as a function of time. For instance, this data is shown in Fig. S2.

2. *Protein abundance analysis*

To test the hypothesis that the targeted protein is depleted while protein-associated function continues for multiple generations, we visualized YPet-DnaN abundance and localization after the protein was knocked out as described in the paper. In short, we constructed a fluorescent fusion at the endogenous locus to make the YdnaN strain (Sec. 2 C), in which the endogenous *dnaN* was replaced by the fusion gene *YPet-dnaN*. In the knockout-depletion experiment, we knocked out the *YPet-dnaN* gene with the *kan* cassette to form *YPet-dnaN*::*kan*.

To test the protein dilution hypothesis, we measured total progeny fluorescence (the proxy for protein abundance of YPet-DnaN) as a function of time, as the cell progeny proliferated. The dilution model predicts that the protein abundance should scale with the total progeny area like:

$$C(t) = C(0)\frac{A_0}{A(t)}, \tag{S9}$$

where $C(t)$ is the protein concentration at time t, C_0 is the abundance at time $t = 0$, A_0 is the progenitor area at time $t = 0$, and $A(t)$ is the total area of the progeny at time t. In the context of the fusion experiments, the observable is fusion fluorescence, equivalent to an intensity scaling of:

$$I(t) = I(0)\frac{A_0}{A(t)}, \tag{S10}$$

where $I(t)$ and $I(0)$ are the average pixel intensity of the progeny at time t and the progenitor at $t = 0$. Both area A and intensity I are time-dependent quantities available in the `clist.mat` file. (See Sec. 4 B 1.)

Several successive improvements in the experimental design and analysis were required to test the dilution hypothesis. (i) We initially attempted to image cells at the same frame rate as our phase contrast experiment (1 frame/2 min); however, to resolve YPet-DnaN foci after protein depletion, we had to significantly increase the exposure time of the fluorescence images and decrease the frame rate to avoid phototoxicity and bleaching. Although the predicted scaling (Eq. S10) was immediately observable in the data without corrections at short times, more care was required to observe the depletion at long times. (ii) First, we background subtracted to account for the background fluorescence level, computed as the average intensity in each frame outside the cell masks. This correction significantly improved the agreement with Eq. S10 at intermediate times, but did not yet account for cellular autofluorescence. (iii) Next, we analyzed a mixture of wild-type and *YdnaN* cells, using the intensity of the wild-type cell in the same microcolony for the background subtraction. This method led to good agreement with Eq. S10 even at long times (Fig. 1).

Why was a mixture of wild-type and *YdnaN* cells preferable to imaging the two strains independently? A detailed analysis of single cell intensities revealed that wild-type cells in close proximity to *YdnaN* cells in the microcolony had higher pixel intensity, due to the diffuse halo created by the bright *YdnaN* cells. The use of wild-type cells in the same field of view helped correct for the diffuse fluorescent light necessary for the analysis of protein abundance at large depletion times. Cell fluorescence intensities at $t = 0$ are used to differentiate between wild-type and *YdnaN* cells.

3. *Areal growth analysis*

In this section, we develop the statistical model for the analysis of cell-area based growth assays to determine both the model parameters and the statistical uncertainty of parameters on a per-experiment basis. We provide this development for completeness; **however, cell-to-cell variation will dominate the reported errors**.

Statistical procedure. For the imaging-based analyses, we define the following statistical procedure: For the analysis of essential genes, we will fix the asymptotic growth rate $k = 0$. Therefore, the Sufficiency model is now considered the null hypothesis since it is the lowest dimensional model. The first alternative hypothesis is the No-effect model, where the wild-type growth rate k_0 is fit in each analysis. If the Sufficiency model is rejected, we then adopt the No-effect model as the new null hypothesis and adopt the Overabundance model as the new alternative hypothesis.

Areal growth models. This growth metric is sensitive to

cell elongation (rather than septation). Let $A(t)$ be the observed area of all cells sharing a single progenitor cell. For the areal growth model, we substitute cell area $A(t)$ for the abundance $N(t)$ and A_0 for N_0 in Eqs. S1, S3, and S7. The models are:

$$\ln A_{\mathrm{S}}(t; A_0) = \ln A_0, \tag{S11}$$

$$\ln A_{\mathrm{N}}(t; A_0, k_0) = \ln A_0 + k_0 t, \tag{S12}$$

$$\ln A_{\mathrm{O}}(t; A_0, k_0, T) = \ln A_0 + k_0 t + ... + \ln \frac{e^{k_0 T}+1}{e^{k_0 T}+e^{k_0 t}}, \tag{S13}$$

where we have substituted $k = 0$.

Statistical model for areal growth analysis. We will model the error associated with determining the area of the cells as proportional to cell number or area:

$$\sigma_A \propto A(t). \tag{S14}$$

This model is consistent with many mechanisms. Rather than fitting a model with a variable error, it is more convenient to introduce a new variable, a, with constant error:

$$a(t) \equiv \ln A(t). \tag{S15}$$

Since $\mathrm{d}a = \mathrm{d}A/A$, then $\sigma_a = \sigma_A/A$ which leads to an analysis with constant error.

The Shannon information (minus log likelihood) for the log area in frame i is:

$$h(a_i|\boldsymbol{\theta}) = \frac{1}{2}\ln 2\pi\sigma_a^2 + \frac{1}{2\sigma_a^2}[a_i - \mu_a(t_i;\boldsymbol{\theta})]^2, \tag{S16}$$

where $\boldsymbol{\theta}$ represents the parameter vector, μ_a is the time-dependent mean log area defined by the growth models (Eqs. S11-S13). For a time series with $i = 1...N$ frames, the total Shannon information is:

$$h(\{a_{i=1...N}\}|\boldsymbol{\theta}) = \frac{N}{2}\ln 2\pi\sigma_a^2 + \frac{1}{2\sigma_a^2}S^2, \tag{S17}$$

which can be formulated as a least-squares minimization where:

$$\Delta a_i \equiv a_i - \mu_a(t_i;\boldsymbol{\theta}), \tag{S18}$$

$$S^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \Delta a_i^2, \tag{S19}$$

where i is the frame index.

Estimate of error for areal growth analysis. We will statistically estimate the relative area uncertainty (σ_a) from the wild-type growth data. The expression for the MLE for σ_a^2 is:

$$\hat{\sigma}^2_{a,\mathrm{MLE}} = \frac{1}{N}S^2(\hat{k}_0, \hat{a}_0), \tag{S20}$$

where Eq. S19 is evaluated at the MLE values of the other parameters for the No-effect model. There is one additional improvement to this estimate which is straight forward to implement. It is well known that Eq. S20 is biased from below. We can construct an unbiased estimator by correcting for the complexity of the model for the mean (dimension two) [44]:

$$\hat{\sigma}_a^2 = \frac{1}{N-2}S^2(\hat{k}_0, \hat{a}_0), \tag{S21}$$

which we will use for our variance estimator. Note that if only a single mean were fit, the prefactor would be $(N-1)^{-1}$ accounting for the one model dimension; however, since we fit both the slope and the offset, the prefactor is $(N-2)^{-1}$ accounting for the two model dimensions [44, 45].

From the wild-type growth data, the unbiased estimator for the error for log area (Eq. S21) is:

$$\sigma_a = 1.5 \times 10^{-3}, \tag{S22}$$

or alternatively, this result can be stated in a more intuitive form: There is a 0.15% error in the cell area.

Application to observed data. To determine the model parameters (Eq. S67), we will minimize the Shannon information (Eq. S34) numerically, by a least-squares minimization of Eq. S18. We estimate the Fisher information using the resulting Jacobian from the least-squares minimization:

$$\hat{I} \equiv \frac{1}{\sigma_a^2}JJ^T, \tag{S23}$$

where the Jacobian matrices J are contracted over the frame index and σ_a is given by Eq. S22. The parameter uncertainties are then estimated from the Fisher information (Eq. S68). Although Eq. S68 accounts for the statistical uncertainty in the parameters, it does not account for the cell-to-cell variation. We found that this cell-to-cell variation was dominant. We therefore cite this cell-to-cell variation-based uncertainty. For the p-value calculations (Eq. S71), we compute the test statistic λ (Eq. S69) from the differences between residual norms for the null and alternative hypotheses:

$$\lambda = \frac{1}{2\sigma_a^2}(S_0^2 - S_1^2), \tag{S24}$$

where σ_a is given by Eq. S22, and the residual norms for model I (the null (0) or the alternative (1) hypotheses) are defined in Eq. S19.

C. Cell-number growth analysis

In this section, we develop the statistical model for the analysis of cell-number based growth assays to determine both the model parameters and the statistical uncertainty of parameters on a per-experiment basis. We provide this development for completeness; **however, cell-to-cell variation will dominate the reported errors**.

Statistical procedure. For the imaging-based analyses, we define the following statistical procedure: For the analysis of essential genes, we will fix the asymptotic growth

rate $k = 0$. Therefore, the Sufficiency model is now considered the null hypothesis since it is the lowest dimensional model. The first alternative hypothesis is the No-effect model where the wild-type growth rate k_0 is fit in each analysis. If the Sufficiency model is rejected, we then adopt the No-effect model as the new null hypothesis and adopt the Overabundance model as the new alternative hypothesis.

Cell-number growth models. For the cell-number growth model, we use Eqs. S1, S3, and S7. The statistical models depend on the growth rates as function of time for model I, which we define as:

$$k_I = \frac{\partial}{\partial t} \ln N_I(t; \theta_I), \tag{S25}$$

where N_I is the cell abundance in model I at time t. The growth rates for the respective models are:

$$k_N(t; k_0) = k_0, \tag{S26}$$
$$k_S(t) = 0, \tag{S27}$$
$$k_O(t; k_0, T) = k_0 \cdot [1 + e^{k_0(t-T)}]^{-1}, \tag{S28}$$

where Eq. S28 interpolates between the initial growth rate k_0 and final growth rate $k = 0$ at time T.

Deriving the Shannon information. Consider an experiment in which images are taken with a high frame rate, where the time duration between frames is δt. Let the frame number be denoted $I = 1...m$ and the number of cells in each frame N_I. Let the model for cell growth be formulated such that the growth rate at time t_I is:

$$k_I = k(t_I; \theta), \tag{S29}$$

where θ represents a parameter vector. In this analysis, we with model cell division as a Markovian process where:

$$\dot{N} = kN, \tag{S30}$$

which is to say that we will ignore the internal state of cells. For instance, at time t, cells have the same rate of division, irrespective of cell age.

In this model, the number of cell divisions n_I that occur over the short time interval δt is Poisson distributed:

$$q(n_I|\mu_I) = \frac{\mu_I^{n_I}}{n_I!} e^{-\mu_I}, \tag{S31}$$

where

$$\mu_I \equiv \delta t N_I k(t_I; \theta), \tag{S32}$$

is the mean number of divisions.

We now compute the Shannon information associated with the entire experiment:

$$h(\{N_I\}_{I=1...m}|\theta) = -\sum_{I=1}^{m} \ln q(n_I|\mu_I). \tag{S33}$$

Substituting Eqs.S31 and S32, the equation is simplified to:

$$h = \sum_{I=1}^{m} \delta t N_I k_I - \sum_{I \in \text{Div}} n_I \ln \delta t N_I k_I + \sum_{I \in \text{Div}} \ln n_I!, \tag{S34}$$

where Div represents the frames immediately preceding division. For instance, if there is one cell at frame 5 and two cells at frame 6, Div $= \{5\}$.

Application to observed data. To determine the model parameters (Eq. S67), we will minimize the Shannon information (Eq. S34) numerically, and determine the Hessian at the optimal parameter values to estimate the Fisher information:

$$\hat{I}_{ij} = H_{ij}, \tag{S35}$$

where H is the Hessian matrix. The parameter uncertainties are then estimated from the Fisher information (Eq. S68). Although Eq. S68 accounts for the statistical uncertainty in the parameters, it does not account for the cell-to-cell variation. We found that this cell-to-cell variation was dominant. We therefore cite this cell-to-cell variation-based uncertainty. For the p-value calculations (Eq. S71), we compute the test statistic λ (Eq. S69) from the differences in the Shannon information (Eq. S34).

D. Results: Imaging-based analyses

1. *Some progenitors have heterogenic progeny*

Heterogenic progenitors are progenitor cells that are observed to have progeny with two distinct heritable phenotypes: the Km^{R} knockout phenotypes and the Km^{S} wild-type phenotype. For instance, in the Δ*murA* knockout-depletion experiments, progenitors were observed with one daughter whose progeny proliferated for multiple generations on Km^{+} media before lysing, the knockout phenotype, and whose other daughters proliferated for a short period but maintained wild-type morphology. The maintenance of the wild-type morphology suggested that the cells were *murA*$^{+}$ Km^{S}. How were these cells able to proliferate while other Km^{S} cells immediately arrested?

We hypothesize that since both cells had the same progenitor, recombination occurred in the mother cell, after the *murA* gene was replicated, leading to one wild-type chromosome and one Δ*murA* chromosome. The transient growth of the wild-type cells was the result of overabundance of the *kan* gene product APH(3′)II being expressed before cell division in the original mother cell.

Heterogenic progenitor cells appeared frequently for *dnaN* knockout-depletion experiments, presumably because of the location of *dnaN* in the immediate vicinity of the origin, resulting in early replication. In these

experiments, an additional test of the heterogenic progenitor hypothesis was possible due to the fluorescent labeling of the target protein. Cells that arrested early with the wild-type morphology showed no protein depletion; whereas cells that displayed the mutant phenotype (filamentation) showed depleted YPet-DnaN levels.

2. *Wild-type imaging-based analyses*

We analyzed two different strains with wild-type growth phenotypes: wild-type cells (*Acinetobacter baylyi* ADP1) and ACIA0320-0321::*kan*.

IS::kan. To generate a reference wild-type growth phenotype, we choose a non-essential gene with no reported phenotype, genes ACIA0320-0321, corresponding to an IS element. The deletion was performed on the YdnaN strain, which shows no growth phenotype under the experimental conditions. We will abbreviate this strain Δ*IS*. We constructed this deletion and measured its growth relative to wild-type on Km$^-$ media, and no growth phenotype was observed. However, even though this strain can be stably maintained (since ACIA0320-0321 is non-essential), we transformed this cassette using the same protocol in knockout-depletion experiments. As expected, a comparable number of transformants were observed using this construct to those targeting essential genes.

A typical transformant from a knockout-depletion experiment targeting *IS* is shown in Fig. S2 for which six generations of growth are captured. Both the areal (cell-elongation-dependent) and cell-number (septation-dependent) analyses are consistent with the null hypothesis, the *No-effect model*, as expected. The growth rate was observed to be $k = 0.925 \pm 0.005$ hr^{-1} for the areal analysis and $k = 1.04 \pm 0.14$ hr^{-1} for the cell-number analysis.

Qualitative phenomenology. A typical knockout-depletion experiment is shown in Fig. S2. Panel A shows a frame mosaic. The cells in this dataset show the log-phase growth phenotype of wild-type cells. Both cell number and area show exponential growth. The step-like growth of the cell number reflects the desynchronization of cell division events of the ancestors for a single progenitor.

Quantitative analysis. The null hypothesis (*Sufficiency model*) was rejected in favor of the No-effect model for both the area and cell-number analysis (both p-values under machine precision). The growth rate was observed to be $k = 1.04 \pm 0.14$ hr^{-1} for the areal analysis and $k = 0.925 \pm 0.005$ hr^{-1} for the cell-number analysis.

3. *dnaA imaging-based analysis*

Annotated gene function. DnaA is an essential regulator of the cell cycle and DNA replication initiation in particular.

Qualitative phenomenology. A typical knockout-depletion experiment is shown in Fig. S3. Panel A shows a frame mosaic. The cells in this dataset show the onset of the phenotype, cell filamentation, without undergoing significant growth-induced protein dilution. As a result, the cell number, shown in Panel B, is constant since no divisions are observed. However, as shown in Panel C, cell elongation continues for roughly 100 min before it begins to arrest. We interpret the metric that shows the earliest arrest to define the overabundance. In this case, since septation is not observed again after transformation, DnaA abundance is consistent with the Sufficiency model.

Quantitative analysis. The null hypothesis (*Sufficiency model*) was rejected in favor of the *Overabundance model* (p-value under machine precision) for the areal analysis. The initial growth rate was observed to be $k = 1.25 \pm 0.02$ hr^{-1} with an arrest time of $T = 1.24 \pm 0.10$ hr. In case of the cell number analysis, we fail to reject the null hypothesis (*Sufficiency model*), indicating that there is no statistical significance to support the alternative hypothesis *No-effect model* (p = 1.0). We used the Δ*IS* wild-type growth rate ($k = 0.925 \pm 0.005$ hr^{-1}) to fit the arrest time: $T = 0.0 \pm 0.3$ hr.

4. *dnaN imaging-based analysis*

Annotated gene function. The gene product of *dnaN* is the β sliding clamp (DnaN), which is an essential component of the replisome complex.

Qualitative phenomenology. A typical knockout-depletion experiment is shown in Fig. S4. Panel A shows a frame mosaic. The cells in this dataset show the onset of the phenotype, cell filamentation, at about 220 min, after multiple rounds of cell division. As a result, the cell number, shown in Panel B, plateaus shortly after the filamentation is observed since the filamentation is a consequence of the failure of the cells to efficiently septate. However, as shown in Panel C, cell elongation continues, although slowing slightly, throughout the experiment. In this case, since arrest is observed first with respect to septation, we use the arrest of this process to define overabundance.

Quantitative analysis. The null hypothesis (*Sufficiency model*) was rejected for both the area ($p = 8.9 \times 10^{-140}$) and cell-number analysis ($p = 6.0 \times 10^{-19}$). The initial growth rate was observed to be $k = 1.02 \pm 0.05$ hr^{-1} with an arrest time of $T = 4.5 \pm 7.7$ hr for the areal analysis.

Gene:	**Area** (Cell elongation dependent) Model selected:	Growth rate: k (hr^{-1})	Arrest time: T (hr)	Overabundance: $\log_{10} o$	**Cell-number** (Cell septation dependent) Model selected:	Growth rate: k (hr^{-1})	Arrest time: T (hr)	Overabundance: $\log_{10} o$	Number of cells: N_C	Number of progenitors: N_P
IS(Wild-type)	No-effect	0.925 ± 0.005	NA	NA	No-effect	1.04 ± 0.14	NA	NA	60	1
dnaA	Overabundance	1.25 ± 0.02	1.2 ± 0.1	0.7 ± 0.1	Sufficiency	1.04	0.0 ± 0.3	0.0 ± 0.2	4	4
dnaN	Overabundance	1.02 ± 0.05	4.5 ± 7.7	2.0 ± 3.0	Overabundance	0.88 ± 0.07	3.8 ± 0.1	1.4 ± 0.1	134	8
ftsN	Overabundance	0.78 ± 0.06	5.2 ± 0.3	1.8 ± 0.2	Overabundance	1.12 ± 0.25	1.3 ± 0.4	0.6 ± 0.2	19	5
murA	Overabundance	0.70 ± 0.08	3.6 ± 0.4	1.1 ± 0.1	Overabundance	0.96 ± 0.24	2.0 ± 0.3	0.9 ± 0.2	16	4

TABLE S2. **Detailed results from fitting imaging-based knockout-depletion experiments.** The table summarizes the analysis of cell proliferation by two complementary metrics: area and cell-number analyses. These two metrics depend on distinct cellular processes: Growth in cell area is dependent on cell elongation, whereas the proliferation of cell number is dependent on the septation process. We give two metrics for sample size: the number of progenitors (N_P) and the total number of cells analyzed (N_C), corresponding to progenitor and progeny. The estimated standard error is provided for parameter fits.

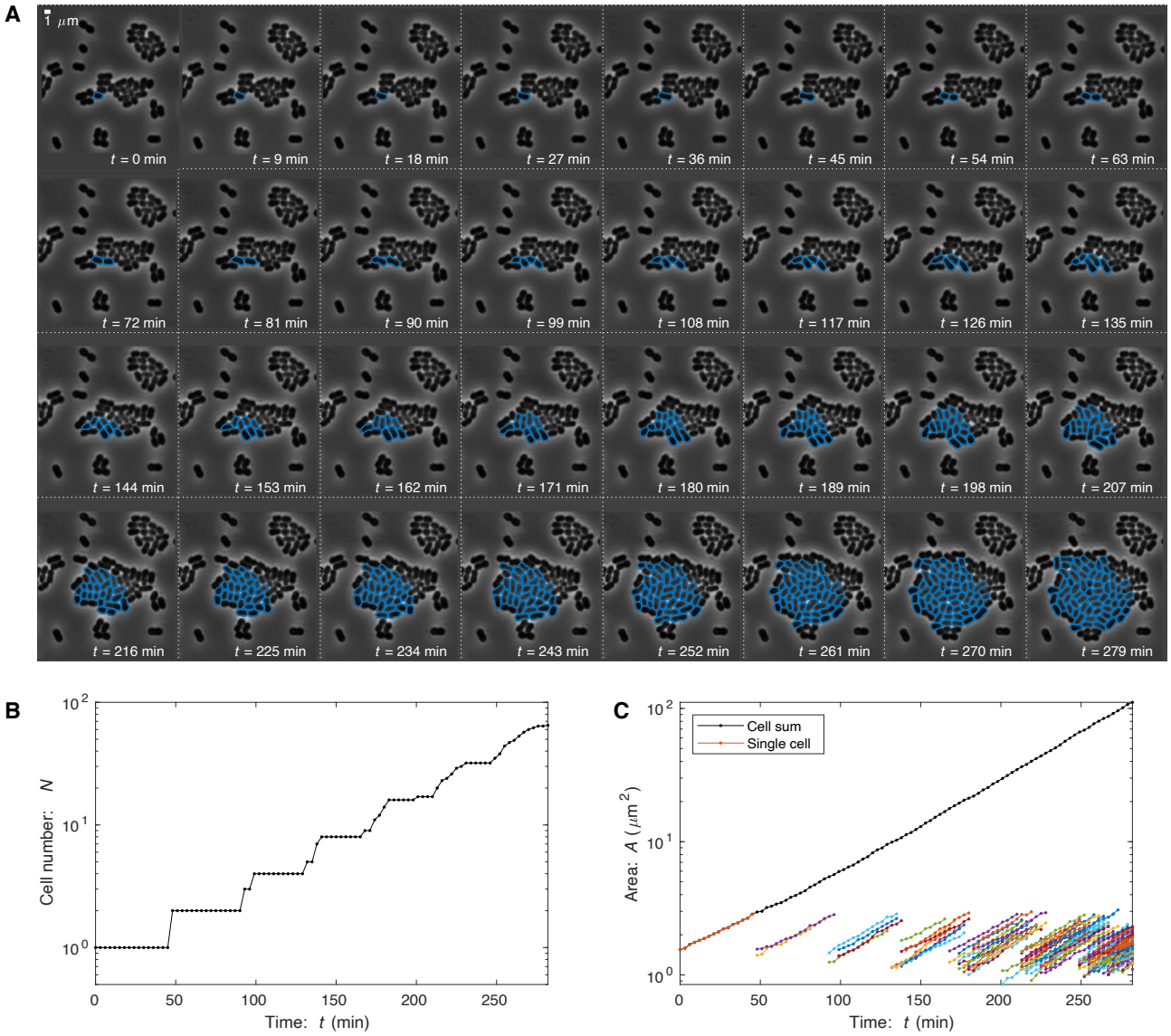

FIG. S2. **Knockout-depletion experiment: IS element (Non-essential). Panel A: Frame mosaic.** In the knockout-depletion experiment, the majority of cells are not transformed and immediately arrest on media supplemented with kanamycin. The lone transformant (*IS*::*kan* (Km^R), blue) proliferates normally. Cells were segmented using SuperSegger-Omnipose for quantitative analysis. **Panel B: Cell number.** The number of transformant progeny as a function of time. **Panel C: Progeny area.** Total progeny-cell area as a function of time. Total cell area is plotted with the black-dotted line, while individual cell areas are plotted with color.

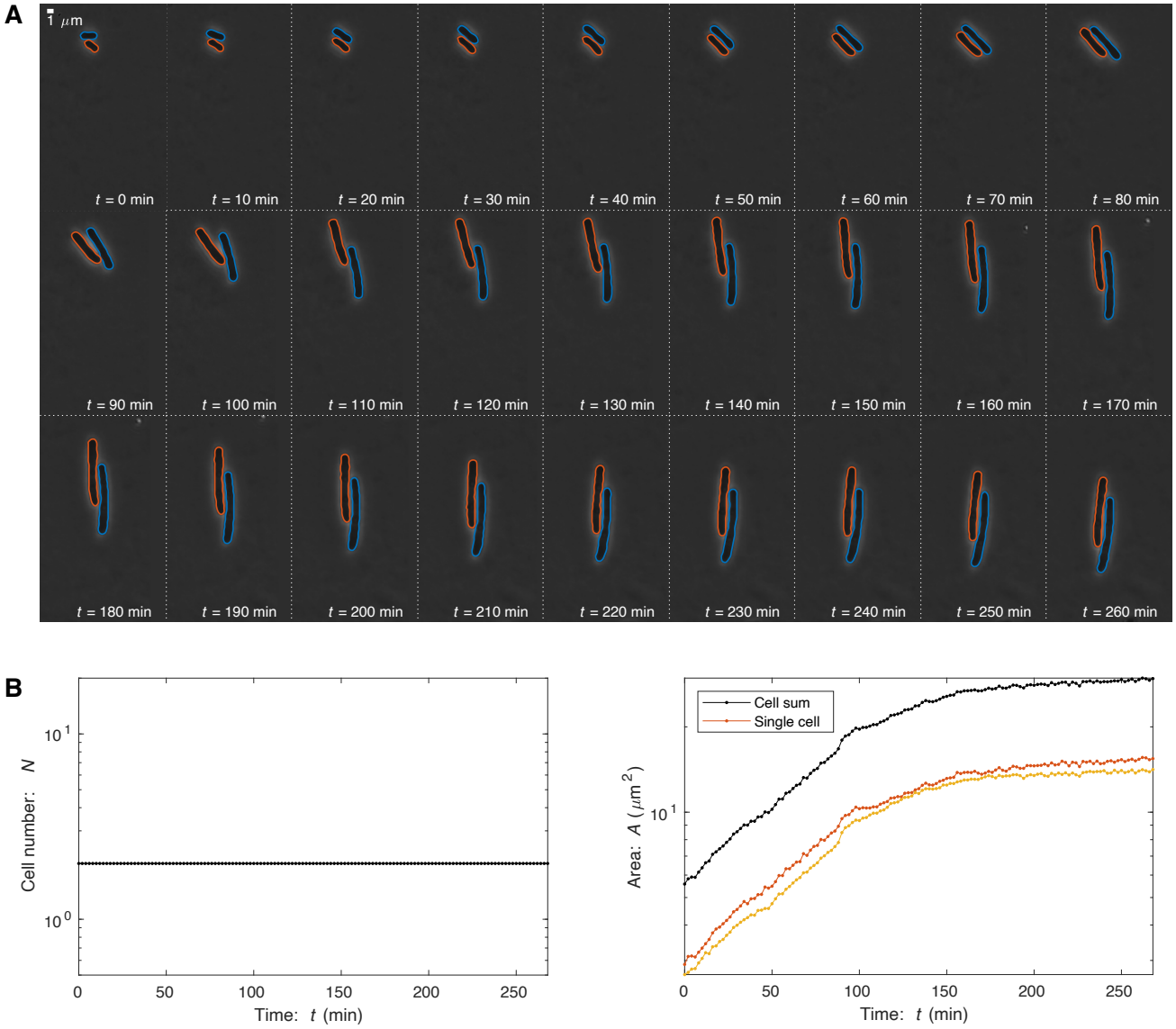

FIG. S3. **Knockout-depletion experiment:** Δ*dnaA*. **Panel A: Frame mosaic.** Two transformants (*dnaA*::*kan*(Km^R), blue, orange) proliferate. DnaA is an essential regulator of replication initiation. Its depletion leads to a failure of the chromosome to replicate, and therefore results in cell filamentation. Cells were segmented using SuperSegger-Omnipose for quantitative analysis. **Panel B: Cell number.** The number of transformant progeny as a function of time. After transformation, cells fail to divide, consistent with DnaA expression being sufficient rather than overabundant. **Panel C: Progeny area.** Total progeny-cell area as a function of time. In spite of the arrest of septation/division, cell areal elongation persists for roughly 120 minutes.

For cell-number analysis, the initial growth rate was observed to be $k = 0.88 \pm 0.07$ hr^{-1} with an arrest time of $T = 3.8 \pm 0.1$ hr.

5. *ftsN imaging-based analysis*

Annotated gene function. The gene product of *ftsN* is essential cell division protein FtsN.

Qualitative phenomenology. A typical knockout-depletion experiment is shown in Fig. S5. Panel A shows a frame mosaic. The cells in this dataset show the onset of the phenotype: the failure to septate, at roughly 150 minutes, after several rounds of division. As a result, the cell number, shown in Panel B, plateaus shortly after 150 min as a consequence of the failure of the cells to efficiently septate. However, as shown in Panel C, cell elongation continues, although slowing slightly, to roughly 220 min.

Quantitative analysis. The null hypothesis (*Sufficiency model*) was rejected for both the area (p-value under ma-

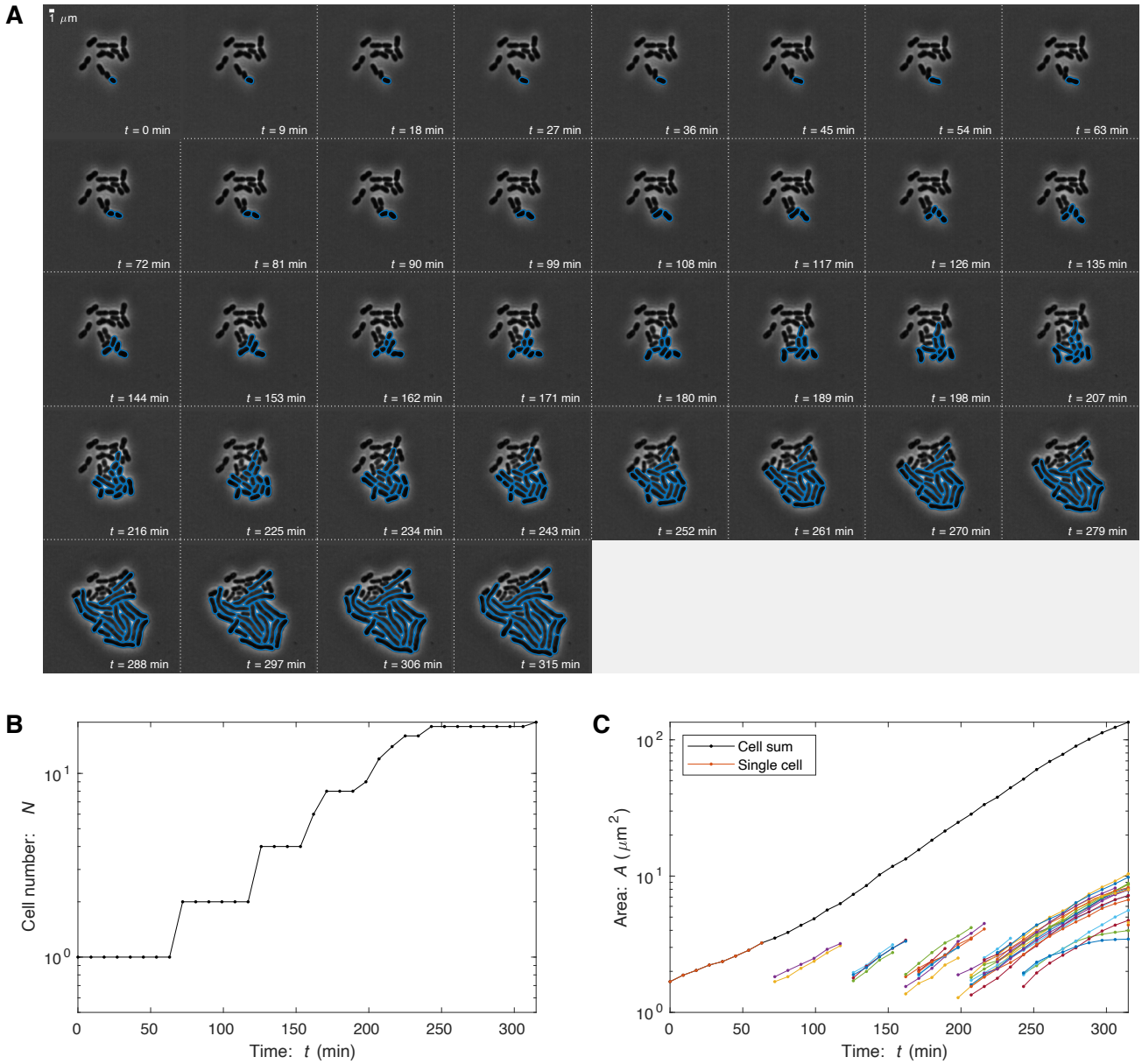

FIG. S4. **Knockout-depletion experiment:** Δ*dnaN*. **Panel A: Frame mosaic.** One transformant (*dnaA*::*kan*(Km^R), blue) proliferates. The frame mosaic shows a typical imaging-based knockout-depletion experiment. DnaN is the sliding beta clamp, an essential DNA replication protein and a core component of the replisome. Its depletion leads to a failure of the chromosome to replicate and therefore results in cell filamentation. Cells were segmented using SuperSegger-Omnipose for quantitative analysis. **Panel B: Cell number.** The number of transformant progeny as a function of time. After transformation, normal growth persists for roughly 240 min, consistent with DnaN expression being overabundant. **Panel C: Progeny area.** Total progeny-cell area as a function of time. The areal elongation dynamics persists even are cell division arrests.

chine precision) and cell-number analysis ($p = 1.7 \times 10^{-7}$). The initial growth rate was observed to be $k = 0.78 \pm 0.06$ hr^{-1} with an arrest time of $T = 5.2 \pm 0.3$ hr for the areal analysis. For cell-number analysis, the initial growth rate was observed to be $k = 1.12 \pm 0.25$ hr^{-1} with an arrest time of $T = 1.3 \pm 0.4$ hr.

6. *murA imaging-based analysis*

Annotated gene function. The gene product of *murA* is UDP-N-acetylglucosamine 1-carboxyvinyltransferase, an essential protein in synthesizing the precursors of cell wall synthesis.

Qualitative phenomenology. A typical knockout-depletion

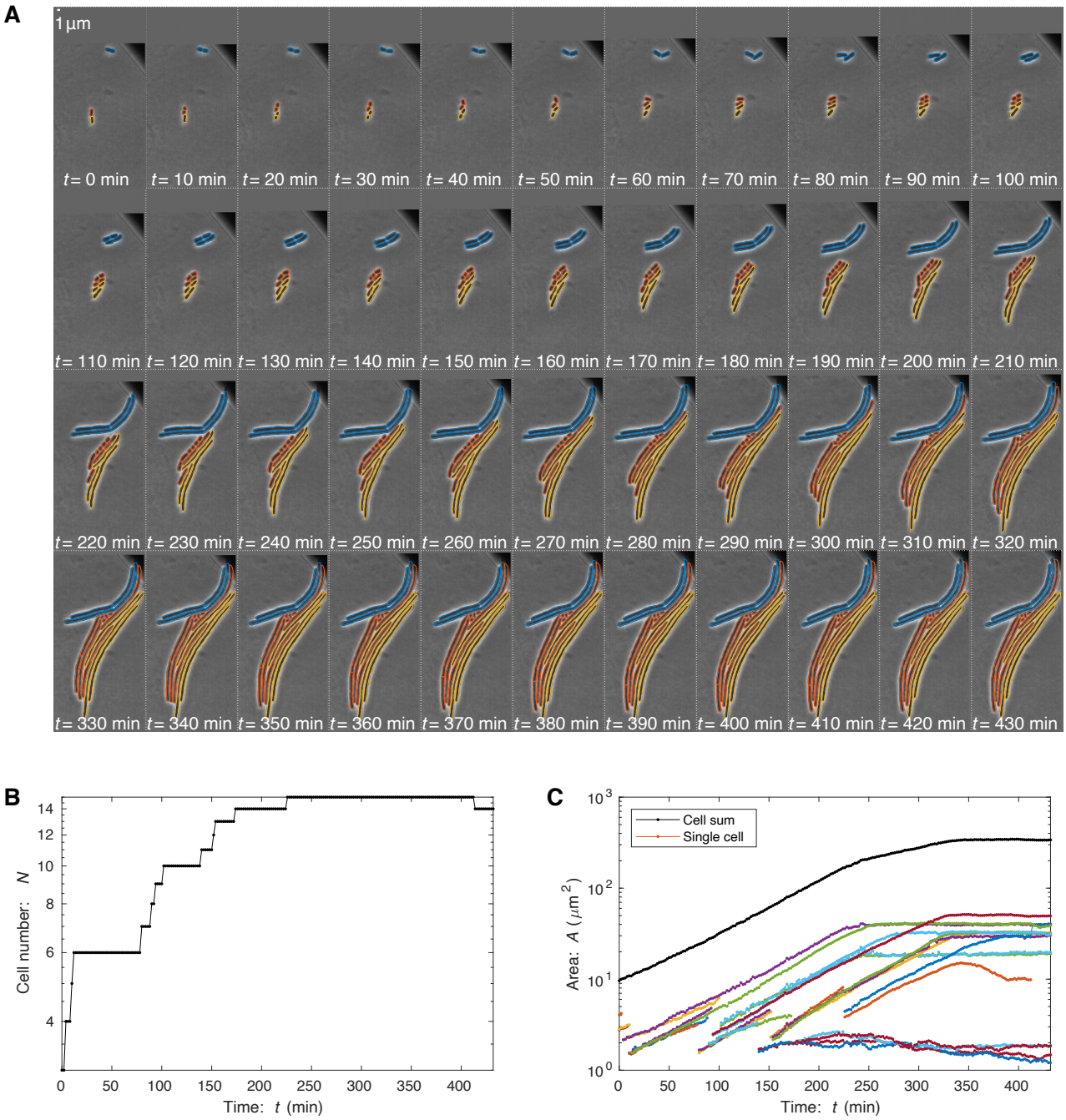

FIG. S5. **Knockout-depletion experiment:** Δ*ftsN*. **Panel A: Frame mosaic.** Three transformants (*ftsN*::*kan*(Km^R), blue, yellow, orange) proliferate. FtsN is an essential cell division protein. Its depletion leads to a failure of the cells to septate. Cells were segmented using SuperSegger-Omnipose for quantitative analysis. **Panel B: Cell number.** The number of transformant progeny as a function of time. After transformation, normal growth persists for roughly 150 min, consistent with FtsN expression being overabundant. **Panel C: Progeny area.** Total progeny-cell area as a function of time. The areal elongation persists even after cell division arrests.

experiment is shown in Fig. S6. Panel A shows a frame mosaic. The cells in this dataset show the onset of the phenotype: the loss of cell wall integrity, and therefore first the loss of wild-type cell morphology and then cell lysis. Cells begin to lose their wild-type morphology at roughly 120 min, after multiple rounds of cell division. As a result, the cell number, shown in Panel B, plateaus shortly after 150 min as a consequence of the

failure of the cells to efficiently septate. However, as shown in Panel C, cell elongation continues, although slowing slightly, to roughly 200 min.

Supplemental approach. For this analysis, we did not want to explicitly model cell lysis. Therefore, in our fitting of the cell-number and areal growth curves, we locked the individual cell area at the last value taken immediately preceding lysis. Similarly, we treated cells that had lysed as arrested, not absent. (This fitting-refined data is *not* shown in Fig. S6. The resulting refined data for Panels B and C plateau rather than decease after growth arrest.)

Quantitative analysis. The null hypothesis (*Sufficiency model*) was rejected for both the area (p-value under machine precision) and cell-number analysis ($p = 1.2 \times 10^{-6}$). The initial growth rate was observed to be $k = 0.70 \pm 0.08$ hr^{-1} with an arrest time of $T = 3.6 \pm 0.4$ hr for the areal analysis. For cell-number analysis, the initial growth rate was observed to be $k = 0.97 \pm 0.24$ hr^{-1} with an arrest time of $T = 2.0 \pm 0.3$ hr.

5. STATISTICAL ANALYSIS OF TFNSEQ TRAJECTORIES

A. Methods: Time correction

Since the mutants transition from lag phase to log phase after transformation, we used a log phase equivalent time for the TFNseq-approach analysis. The corrected sampling times (t_s) are estimated from the number of doublings (D_s) for the non-essential mutants obtained from TFNseq experiment([18]):

$$t_s = D_s * [doubling\ time]. \tag{S36}$$

For our experiment, the doubling time for ADP1 in M9 at 30°C is 37 min.

B. Methods: Defining the likelihood

We assume that deep-sequencing is well modeled by a Poisson process for which the probability mass function is:

$$p(n|\mu) = \frac{\mu^n}{n!} e^{-\mu}, \tag{S37}$$

where n is the number of reads and μ is the mean-number parameter. For large n, we use the normal-distribution approximation:

$$p(n|\mu) \approx \frac{1}{\sqrt{2\pi\mu}} \exp\left[-\frac{(n-\mu)^2}{2\mu}\right]. \tag{S38}$$

The total likelihood for sequential observations $n_{1\ldots m}$ at time $t_{1\ldots m}$ is therefore:

$$q(n_{1\ldots m}|\theta_I) = \prod_{i=1}^{m} p(n_i|\mu)|_{\mu=N_I(t_i|\theta_I)}, \tag{S39}$$

where N_I is one of the trajectory models and θ_I is the parameter vector for model I. The Shannon information is:

$$h(n_{1\ldots m}|\theta_I) \equiv -\ln q(n_{1\ldots m}|\theta_I), \tag{S40}$$

$$= -\sum_{i=1}^{m} \ln p(n_i|\mu)|_{\mu=N_I(t_i|\theta_I)}. \tag{S41}$$

C. Methods: Analysis of overabundance for different Gene Ontologies (GO)

To classify genes, the gene ontology classifications and terms summarized in Tab. S3 were used.

D. Results: Toxicity reduces overabundance.

A second key assumption in the RLTO model is that the metabolic cost of transcription and translation are the dominant fitness costs of protein overabundance (*i.e.* there is no toxicity) [11]. To explore the potential role of toxicity, we generated groups of essential ATPases and enzymes, hypothesizing that these proteins would have higher cost due to excessive activity when overabundant, and a group of DNA-Binding Proteins (DBP), which we hypothesized would have low cost when overabundant. We find that the median overabundance for ATPase genes is 2-fold, and for enzymes more generally 5-fold, compared to 7-fold for all essential genes and 13-fold for DBP. These results are consistent with the hypothesis that toxicity, and in particular ATPase activity, is also a key determinant of overabundance. (See Fig. S9B.)

E. Results: Regulation reduces overabundance.

Instead, we adopted a hypothesis-driven approach and attempted to construct subgroups of essential genes that violate the underlying assumptions used to formulate the RLTO model. A key assumption is that gene expression noise is a consequence of the message number only and is otherwise independent of regulation [11]. Precise control of expression could lead to a reduction in the optimal overabundance. To explore the regulatory hypothesis, we generated three lists of essential genes: autoregulatory, highly regulated (top 10% of genes ranked by number of regulators), and unregulated. If regulation can obviate the need for overabundance, we would expect lower median overabundances in both regulated groups and potentially higher overabundances for the un-regulated group. Consistent with this hypothesis, we find that the median overabundance for autoregulatory genes is 1-fold and for highly-regulated, 3-fold, compared with 7-fold for all essential genes, and 12-fold for un-regulated genes, strongly

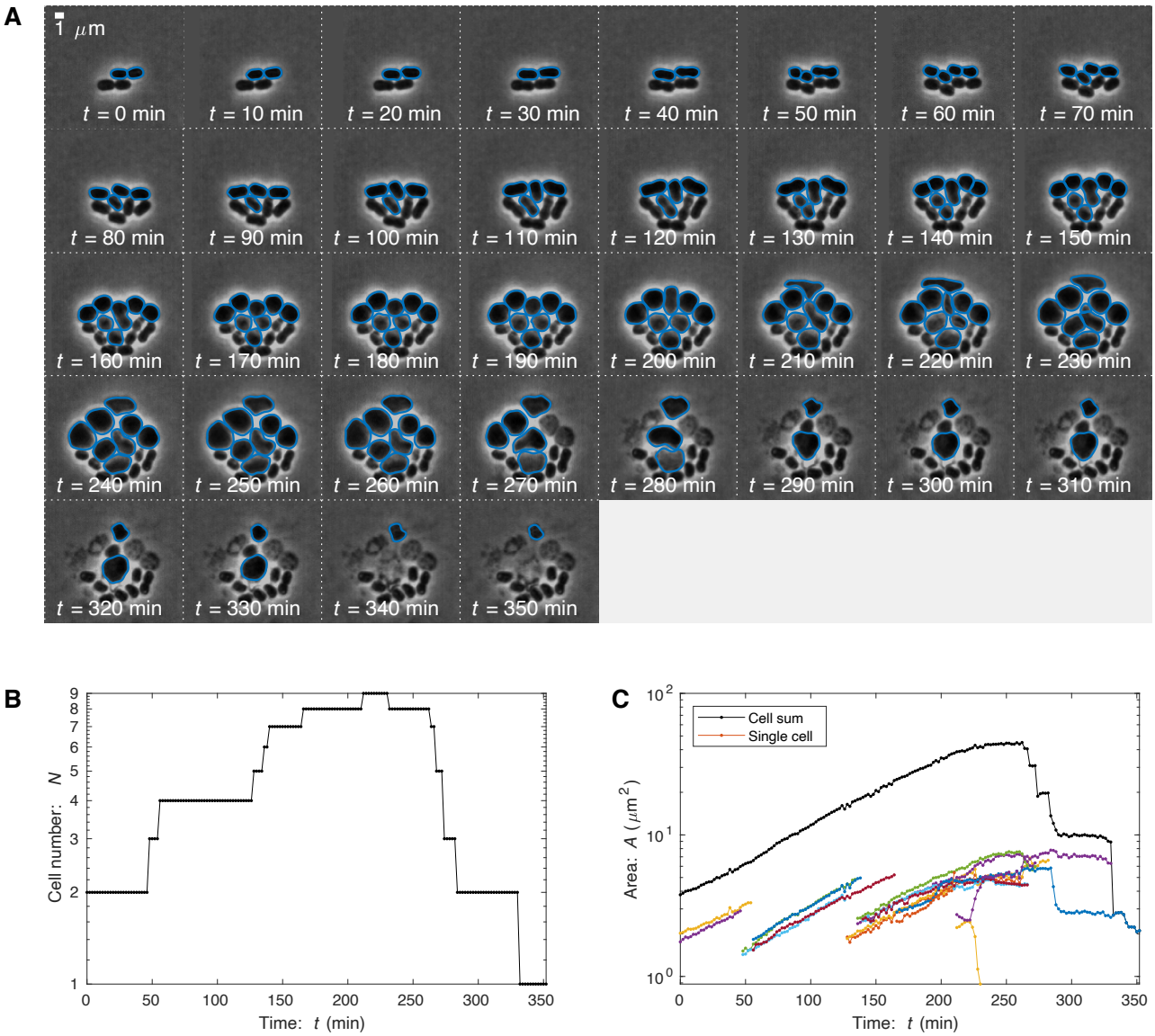

FIG. S6. **Knockout-depletion experiment:** Δ*murA*. **Panel A: Frame mosaic.** Two transformants (*murA*::*kan*(Km^R), blue) proliferate. MurA is an essential enzyme responsible for cell wall precursor synthesis. Its depletion leads to the loss of cell wall integrity, and therefore first the loss of wild-type cell morphology and then cell lysis. Cells were segmented using SuperSegger-Omnipose for quantitative analysis. **Panel B: Cell number.** The number of transformant progeny as a function of time. After transformation, normal growth persists for roughly 200 min, consistent with MurA expression being overabundant. **Panel C: Progeny area.** Total progeny-cell area as a function of time. The areal elongation dynamics are largely consistent with the cell number dynamics: Normal growth persists for roughly 200 min.

supporting the hypothesis that tight regulation could reduce the need for overabundance. (See Fig. S9B.)

We analyzed only genes in *A. baylyi* that had homologues in *E. coli*. The *E. coli* classifications were downloaded from EcoCyc database [46].

F. Analysis of overabundance for different gene regulatory controls

To investigate the effect of transcriptional regulation in determining protein overabundance, we assumed that the regulatory network in *A. baylyi* is roughly equivalent to that in *E. coli* which has been much more extensively studied. We used the EcoCyc database [46] to generate a list for each gene i of the list of direct regu-

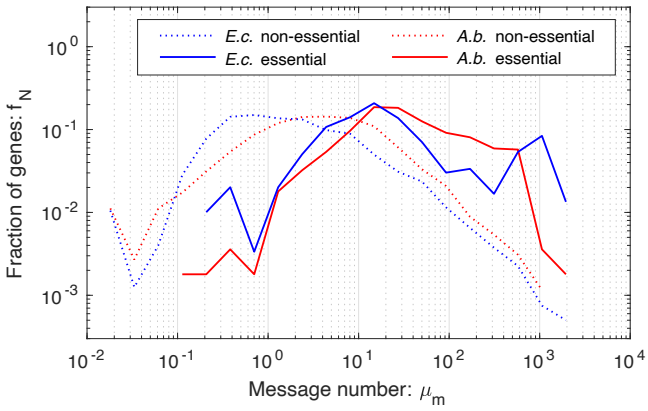

FIG. S7. **Message number distributions for essential and non-essential genes in *E. coli* and *A. baylyi*.** Nearly all *A. baylyi* essential genes are expressed above the one-message-per-cell-cycle threshold. This distribution of both non-essential and essential genes in *A. baylyi* is qualitatively similar to that in *E. coli*, as predicted [11].

Class	Terms
GO:0006260	DNA replication
GO:0051301	Cell division
GO:0008610	Lipid biosynthetic process
GO:0009252	Peptidoglycan biosynthetic process
GO:0008643	Carbohydrate transport
GO:0006355	Regulation of DNA-templated transcription
GO:0003824	Catalytic activity
GO:0003677	DNA binding

TABLE S3. **Gene ontology classifications and terms.** A summary of the gene ontology classifications and terms used in the study.

lators. For each gene, we counted the direct regulators of each gene, then ranked the genes in term of regulator number, and finally we defined the top 10% of the genes as *highly regulated*. We also generated a list of genes directly inhibited by each gene. If a gene directly inhibited itself, we defied the gene as *autoregulatory*.

6. RNA-SEQ ANALYSIS OF TRANSCRIPTION

A. Methods: RNA-Seq protocol

1. *RNA extraction*

ADP1 RNA was harvested through methods developed by Culviner et al. [47] Total RNA was harvested by mixing 1ml of *A. baylyi*(0.5 OD) with 110ul of ice-cold stop solution (95% ethanol and 5% acid-buffered phenol) and spinning in a tabletop centrifuge for 30 s at 13,000 rpm. The supernatant was flash-frozen and stored at -80°C until RNA extraction is ready. To start RNA extraction, 1ml of heated 65°C was added to the sample. The mixture was shaken at 65°C for 10 min and flash-frozen at -80°C for at least 10 min. The pellets were thawed at room temperature and spun at top speed in a benchtop centrifuge at 4°C for 5 min. The supernatant was collected and added to 400 µl of 100% ethanol. The mixture was passed through DirectZol spin column (Zymo). The column was washed twice with RNA prewash buffer and once with RNA wash buffer (Zymo). RNA was eluted from the column with 90 µl diethyl pyrocarbonate (DEPC)-H2O. Genomic DNA was removed with 4 µl of Turbo DNase I (Invitrogen) and supplemented with 10µl of 10x Turbo DNase I buffer to a final volume of 100µl. The solution was heated to 37°C for 40 min. Then RNA was diluted with 100 µl DEPC-H2O, extracted with 200 µl buffered acid phenol-chloroform, followed by ethanol precipitation at -80°C for 4 h with 20 µl of 3 M sodium acetate (NaOAc), 2 µl GlycoBlue (Invitrogen), and 600 µl ice-cold ethanol. To pellet RNA, the samples were centrifuged at 4°C for 30 min at 21,000 × g. The pellets were washed twice with 500 µl of ice-cold 70% ethanol, followed by centrifugation at 4°C for 5 min. RNA pellets were then air dried and resuspended in 50 µl DEPC-H2O. The yield and integrity of RNA was verified with NanoDrop spectrophotometer, and by running 50 ng of total RNA on a Novex 6% Tris-buffered EDTA (TBE)-urea polyacrylamide gel (Invitrogen).

2. *rRNA Depletion*

rRNA was depleted through the DIY method developed by Culviner et al. [47] as well. We used their 21 biotinylated oligonucleotides for *E. coli*. The selected biotinylated oligonucleotides were synthesized by IDT and resuspended to 100 µM in TE buffer (Qiagen). An oligonucleotide mixture was made by mixing equal volumes of each 16S and 23S primers and double volumes of 5S primers. The pooled mixture was diluted with DEPC treated H2O based on the total RNA, using their Excel-based calculator. Using the Excel-based calculator, the calculated volume of Dynabeads MyOne streptavidin C1 beads (ThermoFisher) were washed three times in equal volume of 1x B&W buffer, resuspended in 30 µl of 2× B&W buffer and supplemented with 1µl of SUPERase-In RNase inhibitor (ThermoFisher). The beads were set aside in room temperature until the probes were ready to be pulled down. To collect rRNA, 2 to 3µg total RNA and 1 µl of the diluted biotinylated probe mix were combined on ice into a final annealing reaction mixture of 1xSSC and 500 µM EDTA. All the appropriate volumes were computed using the Excel-based calculator. The RNA and probe mixture was incubated at 70°C for 5 min, and slowly cooled to 25°C at a rate of 1°C per 30 s. The annealed mixture was then added to 30 µl of beads that were resuspended in 2× B&W buffer. The mixture was mixed by pipetting and vortexing at medium speed, and followed by incubating for 5 min at room temperature. The reaction mix-

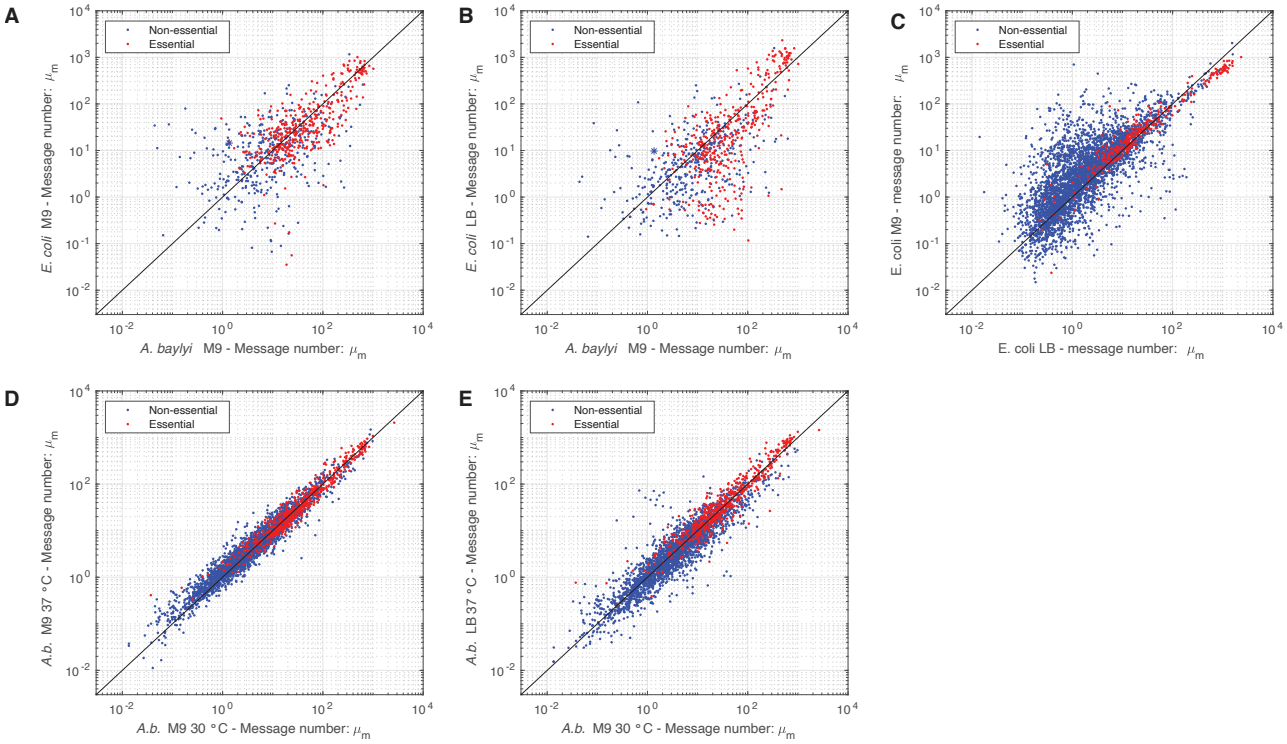

FIG. S8. **Transcriptome comparisons. Panel A: *E. coli* on M9 versus *A. baylyi* on M9. Panel B: *E. coli* on LB versus *A. baylyi* on M9. Panel C: *E. coli* on M9 versus on LB. Panel D: *A. baylyi* on M9 at 37°C versus at 30°C. Panel E: *A. baylyi* on LB at 37°C versus on M9 at 30°C .** Throughout, there is broad consistency between the expression levels (message number) of genes, both between organisms and between conditions. These observations suggest a consistent overall transcriptional program governs gene expression both between organisms and growth conditions.

tures were then vortexed, and incubated at 50°C for 5 min. To pull down the biotinylated probes, the reaction mixtures were placed immediately placed on the magnetic rack. The supernatant was carefully pipetted, placed on ice, and diluted to 200 µl in DEPC-H2O. The RNA was purified through ethanol precipitation with 20 µl of 3 M NaOAc, 2 µl GlycoBlue (Invitrogen), and 600 µl ice-cold ethanol at -20°C for at least 1 h. To pellet RNA, the samples were centrifuged at 4°C for 30 min at 21,000 × g. The pellets were washed twice with 500 µl of ice-cold 70% ethanol, followed by centrifugation at 4°C for 5 min. RNA pellets were then air dried and resuspended in 10 µl DEPC-H2O. The yield and rRNA depletion effectiveness was verified with NanoDrop spectrophotometer, and by running 50 ng of total RNA on a Novex 6% Tris-buffered EDTA (TBE)-urea polyacrylamide gel (Invitrogen). The yield and integrity of the library was checked by running the samples in qPCR using NEBNext Library Quant Kit for Illumina(NEB) and the Bioanalyzer.

3. *Library prep and sequencing*

The RNA library was prepared with NEBNext® Multiplex Oligos for Illumina(NEB) and NEBNext ultra II RNA Library Prep Kit for Illumina(NEB). For the library prep protocol, we followed section 4 of the kit's provided protocol: Protocols for use with Purified mRNA or rRNA Depleted RNA. The quality of the final library was verified by running the samples on high sensitivity Bioanalyzer chip. The samples were pooled to a final concentration of 8.5nM, and were sequenced with NextSeq 150 cycle kit.

B. Methods: Computation of message number

To estimate the message number for gene i, defined as the total number of mRNA molecules transcribed per cell cycle, from the RNA-Seq data, we use the approach we described earlier [11]. Let the relative number of reads for gene i be r_i:

$$r_i = \frac{N_i}{N_{\text{tot}}}, \tag{S42}$$

where N_i is the reads per kilobase (rpk) for gene i and N_{tot} is the rpk for all genes. We apply two different

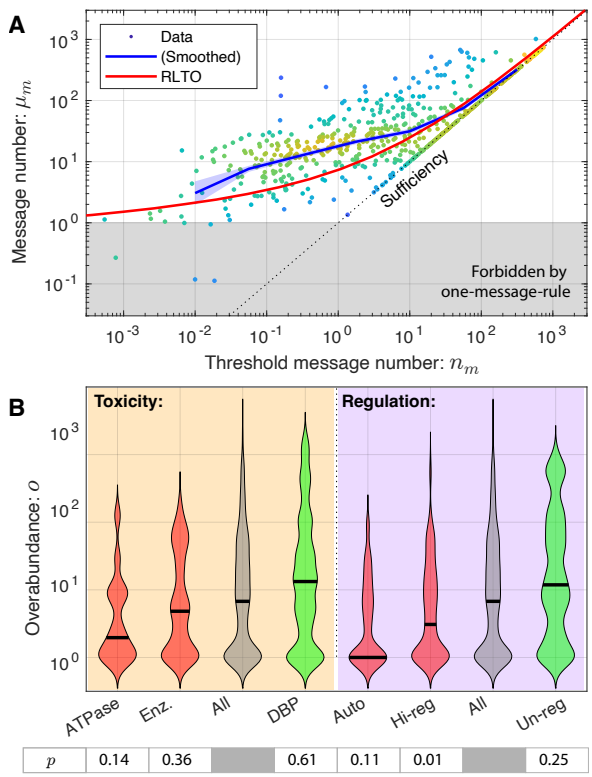

FIG. S9. **Panel A: Threshold and message number.** Alternatively, the message number μ_m is shown as a function of the inferred threshold message number: $n_m \equiv \mu_m/o$. Although it is μ_m and o that are most directly measurable, the quantity n_m is a more intuitive quantity from a modeling perspective since μ_m is optimized to maximize fitness at fixed n_m in the RLTO model. **Panel B: Toxicity and regulation are determinants of overabundance.** We compared the overabundance measurements for six essential gene subgroups to determine whether toxicity and regulation could affect overabundance. Red groups were predicted to decrease overabundance while green groups were expected to increase it. The p-values for the consistency of each distribution with the all gene group is shown below each category. As hypothesized, the data is consistent with both toxicity and regulation decreasing overabundance.

re-scalings: First we re-scale the relative message abundance to reflect the cellular abundance of the message, and then we scale this number by the ratio of cell cycle duration to mRNA lifetime to estimate the number of times a gene is transcribed per cell cycle. For *A. baylyi*, we use the same scaling factor as *E. coli*:

$$\mu_{m,i} = 9.4 \times 10^4 \cdot r_i, \tag{S43}$$

where $\mu_{m,i}$ is the estimated message number (number of mRNA molecules transcribed per cell cycle).

To check the consistency of this estimate, we generated histograms for message number for essential and non-essential genes, and compared them to the histograms for *E. coli*. We expect the distribution of essential message numbers to abut 1 message per cell cycle,

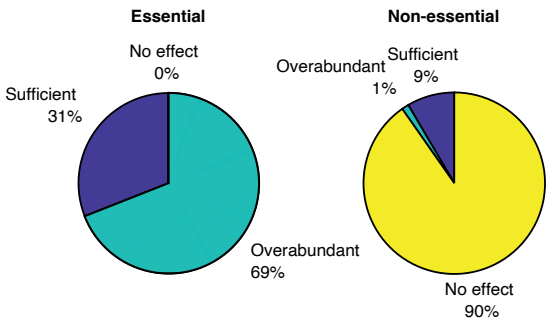

FIG. S10. **Proteome-wide analysis of proliferation dynamics.** For genes classified as essential, 31% were best fit by the sufficiency model, while 69% were best fit by the overabundance model. For genes classified as non-essential, 90% were best fit by the no-effect model, while 10% showed a detectable reduction in growth rate.

while non-essential genes can be expressed at significantly lower levels. The observed distribution are consistent with this expectation. (See Fig. S7.)

C. Results: Comparison of *A. baylyi* and *E. coli* gene expression

Knockout-depletion experiments are not tractable in *E. coli* and many other model systems. It is therefore difficult to directly test the overabundance hypothesis in these other systems. However, it is possible to determine if *E. coli* expression patterns are consistent with overabundance.

If overabundance were specific to *A. baylyi*, we would expect to see higher relative transcription of lower abundance essential genes in *A. baylyi*, where overabundance is large, relative to *E. coli* if its expression levels were sufficient. Fig. S8 compares the message number between homologues in the two organisms and between growth conditions within a particular organism for all genes.

7. ROBUSTNESS LOAD TRADE-OFF (RLTO) MODEL

We have provided a detailed description of the Robustness Load Trade-Off (RLTO) Model in Ref. [11]; however, in the interest of making this paper self-contained, we provide a concise summary of key elements and results from that paper in this supplementary section.

A. Methods: Detailed description of the noise model

1. *Stochastic kinetic model for the central dogma.*

The canonical steady-state noise model for the central dogma describes multiple steps in the gene expression process [30, 31, 48]: Transcription generates mRNA messages. These messages are then translated to synthesize the protein gene products [49]. Both mRNA and protein are subject to degradation and dilution [50]. At the single cell level, each of these processes are stochastic. We will model these processes with the stochastic kinetic scheme [49]:

$$\begin{array}{ccccc} \text{DNA} & \xrightarrow{\beta_m} & \text{mRNA} & \xrightarrow{\beta_p} & \text{Protein} \\ & & \gamma_m \big\downarrow & & \gamma_p \big\downarrow \\ & & \varnothing & & \varnothing, \end{array} \tag{S44}$$

where β_m is the transcription rate (s^{-1}), β_p is the translation rate (s^{-1}), γ_m is the message degradation rate (s^{-1}), and γ_p is the protein effective degradation rate (s^{-1}). The message lifetime is $T_m \equiv \gamma_m^{-1}$. For most proteins in the context of rapid growth, dilution is the dominant mechanism of protein depletion and therefore γ_p is approximately the growth rate [48, 51, 52]: $\gamma_p = T^{-1} \ln 2$, where T is the doubling time.

2. *Statistical model for protein abundance.*

Consistent with previous reports [30, 31], we find that the distribution of protein number per cell (at cell birth) was described by a gamma distribution [11]:

$$N_p \sim \Gamma(a, \theta), \tag{S45}$$

where N_p is the protein number at cell birth and Γ is the gamma distribution, which is parameterized by a scale parameter θ and a shape parameter a. (See Sec. 9 A.) We refer to this distribution as the *canonical steady-state noise model*; The relation between the four kinetic parameters and these two statistical parameters has already been reported, and have clear biological interpretations [31]: The scale parameter:

$$\theta = \varepsilon \ln 2, \tag{S46}$$

is proportional to the translation efficiency:

$$\varepsilon \equiv \frac{\beta_p}{\gamma_m}, \tag{S47}$$

where β_p is the translation rate and γ_m is the message degradation rate. ε is understood as the mean number of proteins translated from each message transcribed. The shape parameter a can also be expressed in terms of the kinetic parameters [31]:

$$a = \frac{\beta_m}{\gamma_p}; \tag{S48}$$

however, we will find it more convenient to express the scale parameter in terms of the cell-cycle message number:

$$\mu_m \equiv \beta_m T = a \ln 2, \tag{S49}$$

which can be interpreted as the mean number of messages transcribed per cell cycle. Forthwith, we will abbreviate this quantity *message number* in the interest of brevity.

B. Methods: Summary of the RLTO model fitness model

1. *Metabolic load in the RLTO model*

To produce a minimal model to study the trade-off between robustness and metabolic load, we must consider both the metabolic cost of transcription and translation. We will write that the metabolic load (in protein equivalents) associated with gene i is:

$$\delta N_i = \lambda \mu_{m,i} + \mu_{p,i}, \tag{S50}$$

where λ is the message cost, the metabolic load associated with an mRNA molecule relative to a single protein molecule of the gene product.

$$\ln \frac{k}{k_0} = -\frac{(\lambda + \varepsilon_i)\mu_{m,i}}{N_0}. \tag{S51}$$

This equation has an intuitive interpretation: growth slows in proportion to the relative added metabolic load. In resource allocation models [53], the capacity of the cell for growth can increase as protein sectors increase in size. In our context, this does not occur since we consider the uncoordinated changes in the levels of single proteins. *I.e.* we assume some other protein of factor is rate limiting. See the detailed discussion in Ref. [11].

2. *Growth rate with stochastic arrest*

As discussed in Ref. [11], we idealize the slow growth associated with essential proteins falling below threshold as growth arrest. This arrest model has phenomenology consistent with more detailed and realistic models where cells experience a significant growth slowdown rather than true growth arrest [11].

In the idealized growth arrest model, if all essential proteins are above threshold, the cell cycle duration τ is determined by the metabolic load predictions (Eq. S51); however, if any essential protein is below threshold, the cell cycle duration is infinite. The probability mass function for the cycle-cycle duration T interpreted as a random variable is therefore:

$$p_T(t) = \begin{cases} P_+, & t = \tau \\ (1 - P_+), & t \to \infty \end{cases}. \tag{S52}$$

As we show in Ref. [11], the growth rate can be computed exactly:

$$k = \tau^{-1} \ln(2P_+). \tag{S53}$$

As expected, the growth rate goes down as the probability of growth P_+ decreases, stopping completely at $P_+ = \frac{1}{2}$. We can then compute the ratio of the growth with (k) and without arrest (k_0):

$$\frac{k}{k_0} = 1 + \frac{1}{\ln 2} \ln P_+, \tag{S54}$$

where k_0 is computed by evaluating Eq. S53 at $P_+ = 1$.

3. RLTO growth rate

In the RLTO model, we will assume the probability of growth is the probability that all essential protein numbers are above threshold. We will further assume that each protein number is independent, and therefore:

$$P_+ = \prod_{i \in \mathscr{E}} \Pr\{N_{p,i} > n_{p,i}\}, \tag{S55}$$

where $\mathscr{E}$ is the set of essential genes. Clearly, this assumption of independence fails in the context of polycistronic messages. We will discuss the significance of this feature of bacterial cells elsewhere, but we will ignore it in the current context. As we will discuss, the probability of arrest of any protein i to be above threshold is extremely small. It is therefore convenient to work in terms of the CDFs, which are very close to zero:

$$\ln P_+ \approx -\sum_{i \in \mathscr{E}} \gamma^-\left(\frac{\mu_{m,i}}{\ln 2}, \frac{n_{p,i}}{\varepsilon_i \ln 2}\right), \tag{S56}$$

where γ^- is the regularized lower incomplete gamma function (Eq. S77) and represents the probability of arrest.

4. Single-gene equation

By summing the fitness losses from the metabolic load and cell arrest (Eqs. S51, S54, and S56), we can write an expression for the growth rate including contributions from essential gene i:

$$\ln \frac{k}{k_0} = -\frac{\lambda+\varepsilon_i}{N_0}\mu_{m,i} - \frac{1}{\ln 2}\gamma^-\left(\frac{\mu_{m,i}}{\ln 2}, \frac{n_{p,i}}{\varepsilon \ln 2}\right), \tag{S57}$$

where the first term on the RHS represents the fitness loss due to the metabolic load and the second term represents the fitness loss due to stochastic cell arrest due to protein i falling below threshold.

5. Optimization of transcription for bacteria

The growth rate is:

$$\ln \frac{k}{k_0} = -\left(\Lambda + \frac{\varepsilon}{N_0}\right)\mu_m - \frac{1}{\ln 2}\gamma^-\left(\frac{\mu_m}{\ln 2}, \frac{n_p}{\varepsilon \ln 2}\right), \tag{S58}$$

where γ^- is the regularized lower incomplete gamma function (Eq. S77), which is the CDF of the gamma distribution and represents the probability of arrest due to gene i. For bacteria, we consider the special case of optimizing the message number only at fixed translation efficiency [11, 48]. To determine the optimal transcription level, we set the partial derivative of Eq. S58 with respect to μ_m to zero. The optimum message number $\hat{\mu}_m$ satisfies the equation:

$$\frac{(\lambda+\varepsilon)\ln 2}{N_0} = -\left[\partial_{\hat{\mu}_m}\gamma(\hat{\mu}_m, \hat{n}_m)\right]_{\hat{n}_m = \frac{\hat{\mu}_m}{\delta}}. \tag{S59}$$

We define the relative load:

$$\Lambda \equiv \frac{(\lambda+\varepsilon)}{N_0}, \tag{S60}$$

and substitute this into the optimum message number equation:

$$\Lambda \ln 2 = -\left[\partial_{\hat{\mu}_m}\gamma(\hat{\mu}_m, \hat{n}_m)\right]_{\hat{n}_m = \frac{\hat{\mu}_m}{\delta}}, \tag{S61}$$

which is solved numerically.

6. Estimate of the relative load in bacterial cells

In bacterial cells, we will assume a constant translation efficiency model. We therefore use the modified relative load formula (Eq. S60) to estimate Λ. We will assume that the load is dominated by proteins and messages:

$$N_0 = \sum_i (\lambda+\varepsilon)\mu_{m,i} = (\lambda+\varepsilon)N_m, \tag{S62}$$

where N_m is the total number of messages. We can then solve this equation for Λ:

$$\hat{\Lambda} = \frac{\lambda+\varepsilon}{N_0} = \frac{1}{N_m} \approx 10^{-5}, \tag{S63}$$

based on the total message number estimate for *E. coli* [11].

C. Results: The fitness landscape of the RLTO model is highly asymmetric

In the RLTO model, the fitness landscape for a single cell is determined by an asymmetric fitness landscape: protein underabundance is extremely costly due to the risk of growth arrest, while the cost of protein overabundance is only associated with an increase in metabolic load. (See Fig. S11A.) Naïvely, this tradeoff predicts

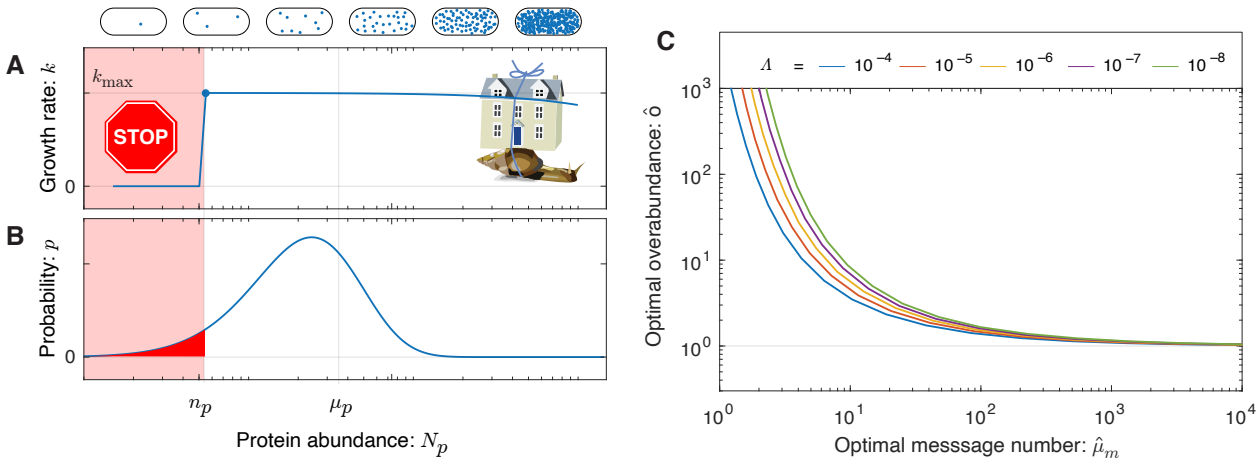

FIG. S11. **Panel A: The fitness landscape is asymmetric in the RLTO model.** Motivated by single-cell growth data, cell fitness is modeled using the Robustness-Load Trade-Off model (RLTO). In the model, there is a metabolic cost of protein expression which favors low expression; however, growth arrests for protein number N_p smaller than the threshold level n_p (red). The relative metabolic cost of overabundance is small relative to the cost of growth arrest due to the large number of proteins synthesized, resulting in a highly asymmetric fitness landscape [11]. **Panel B: The gene expression process is stochastic.** There is significant cell-to-cell variation in protein abundance (N_p) around the mean level (μ_p). Even for mean expression levels significantly above the threshold level n_p, some cells fall below threshold (red). The distribution in protein number is modeled using a gamma distribution [48]. **Panel C: Overabundance is predicted to optimize cell fitness.** The asymmetry of the fitness landscape drives the optimal protein expression level to be overabundant ($\mu_p > n_p$). The RLTO model makes a quantitative prediction of the optimal overabundance ($\hat{o} \equiv \hat{\mu}_p/n_p$) as a function of the message number μ_m and a global parameter, the relative load $\Lambda \approx 10^{-5}$ (red curve). Overabundance is predicted to be extremely high ($o \gg 1$) for low expression genes ($\mu_m \approx 1$) and much closer to sufficiency ($o \approx 1$) for high expression genes ($\mu_m \gg 10$). Although the optimal overabundance depends on the relative load Λ, its qualitative dependence is unchanged over orders of magnitude in variation of the parameter.

that the cell maximizes its fitness by simply expressing just above the minimum protein threshold for function [7]. However, achieving growth robustness at a population level is nontrivial. Gene expression is stochastic [32], leading to significant cell-to-cell variation in protein numbers, which we model with a gamma distribution (Fig. S11B) [30, 31]. Therefore, the strong asymmetry of the fitness landscape predicts protein overabundance.

D. Results: The RLTO model predicts overabundance is optimal for low-expression proteins

The optimal regulatory program for transcription and translation (μ_m and ε values) can be predicted analytically. The values are determined by a single global parameter, the relative load Λ, and the gene-specific threshold number n_p. The threshold number is not directly observable experimentally; instead we predict the optimal overabundance o, defined as the ratio of the mean protein number to the threshold number:

$$o \equiv \mu_p/n_p. \tag{S64}$$

As shown in [11], by taking partial derivatives of the relative growth rate (Eq. S58) with respect to message number and translation efficiency, respectively, we can define the optimal overabundance:

$$\hat{o} \equiv \frac{\hat{\mu}_p}{n_p} = \frac{\hat{\varepsilon}\hat{\mu}_m}{n_p}, \tag{S65}$$

in the large multiplicity limit where the overall metabolic load is much smaller than the metabolic load for a single gene: $N_0 \gg (\lambda + \hat{\varepsilon})\hat{\mu}_m$. The optimal overabundance can be rewritten to find the optimization condition for message number:

$$\Lambda \ln 2 \;=\; -\partial_{\hat{\mu}_m}\gamma\left(\frac{\hat{\mu}_m}{\ln 2}, \frac{\hat{\mu}_m}{\hat{o}\ln 2}\right). \tag{S66}$$

As seen in Fig. S11C, the RLTO model generically predicts that for a range of relative loads, the optimal protein fraction is overabundant ($o > 1$); however, overabundance is not uniform for all proteins, but rather depends on transcription. For highly-transcribed genes ($\mu_m \gg 1$), the overabundance is predicted to be quite small ($o \approx 1$); however, for lowly-transcribed genes (message numbers approaching unity), the overabundance is predicted to be extremely high ($o \gg 1$).

E. Discussion: Does the detailed form of the fitness landscape affect RLTO predictions?

It is important to emphasize that the detailed mathematical form of the RLTO model is not essential to generate the predicted phenomenology. For instance, changes

in the functional form of the protein expression noise, the metabolic load, or a more realistic model of the fitness landscape do not significantly change the predictions of the model. It is the strong asymmetry of the fitness landscape that is required to predict protein overabundance [11].

8. METHODS: STATISTICAL PROCEDURES

In this section, we provide a summary of statistical approaches that are common to the analyses in the paper.

A. Maximum Likelihood Estimation

The maximum likelihood (*i.e.* minimum information) estimates (MLE) of the parameters are defined:

$$\hat{\theta}^i = \arg\min_{\theta^i} h(\text{data}|\theta^i). \tag{S67}$$

In all instances, these optimizations are performed numerically, either by direct minimization of the Shannon information (h), or for normal models, by least-squares minimization.

B. Parametric uncertainty estimates

To estimate the parameter uncertainty in the analysis of datasets, we use the Cramer-Rao bound to estimate of the uncertainty from the Fisher information [44]:

$$\sigma_{\theta^i} = \sqrt{[\hat{I}^{-1}]^{ii}}, \tag{S68}$$

where σ_{θ^i} is the estimate of the standard error for parameter θ^i, $\hat{I}$ is the estimator of the Fisher information, and $[\hat{I}^{-1}]^{ii}$ is the ii component of the inverse Fisher information matrix. For each statistical model, we describe how the Fisher information is estimated in detail (Hessian or Jacobian *etc*).

C. Null-hypothesis-testing approach

For null-hypothesis testing, we define two sequential null-hypothesis tests of nested statistical models. If the initial null hypothesis is rejected, we then interpret the initial alternative hypothesis as the updated null hypothesis and adopt the remaining model as the alternative hypothesis. For each test, we will use a Likelihood Ratio Test (LRT) where we define the test statistic λ in terms of the Shannon information:

$$\lambda = h_0 - h_1, \tag{S69}$$

where h_0 and h_1 are the Shannon information for the null and alternative hypotheses respectively. We will assume the Wilks' theorem: *I.e.* the test statistic Λ under the null hypothesis will have a chi-squared distribution [54, 55]:

$$\Lambda \sim \tfrac{1}{2}\chi^2_{\Delta K}, \tag{S70}$$

where the degrees-of-freedom $\Delta K = 1$ is equal to the difference between the dimension of the alternative and null models. (The factor of $1/2$ appears in this equation, since the test statistic is defined by the Shannon information difference rather than the deviance [44].) The p-value can then be computed:

$$p = \Pr\{\Lambda > \lambda\} = \gamma^+(\tfrac{1}{2}\Delta K, \lambda), \tag{S71}$$

where γ^+ is the upper regularized incomplete gamma function (Eq. S87), $\Delta K = 1$ is the difference in model dimensions, and λ is the test statistic [44].

9. DISTRIBUTIONS AND CONVENTIONS

A. Gamma distribution conventions

There are a number of conflicting conventions for the gamma function and distribution arguments. We will use those defined on Wikipedia and the CRC Encyclopedia of Mathematics [56].

The gamma distributed random variable X will be written:

$$X \sim \Gamma(a, \theta), \tag{S72}$$

where a is the shape parameter and θ is the scale parameter. The PDF of the distribution is:

$$p_X(x|a,\theta) \equiv \tfrac{x^{a-1}}{\theta^a \Gamma(a)} e^{-x/\theta}, \tag{S73}$$

where $\Gamma(a)$ is the gamma function. The CDF is therefore:

$$\begin{aligned} P_X(x|a,\theta) &\equiv \Pr\{X < x|a,\theta\}, && \text{(S74)}\\ &= \int_0^x \mathrm{d}x'\, p_\Gamma(x'|a,\theta), && \text{(S75)}\\ &= \int_0^{x/\theta} \mathrm{d}x''\, \tfrac{x''^{a-1}}{\Gamma(a)} e^{-x}, && \text{(S76)}\\ &= \gamma^-(a, x/\theta), && \text{(S77)} \end{aligned}$$

where γ^- is the regularized lower incomplete gamma function. The survival function is:

$$\begin{aligned} \Pr\{X > x|a,\theta\} &= 1 - P_X(x|a,\theta), && \text{(S78)}\\ &= \int_{x/\theta}^{\infty} \mathrm{d}x''\, \tfrac{x''^{a-1}}{\Gamma(a)} e^{-x}, && \text{(S79)}\\ &= \gamma^+(a, x/\theta), && \text{(S80)} \end{aligned}$$

where γ^+ is the regularized upper incomplete gamma function.

B. Chi-squared distribution conventions

In statistical null hypothesis testing, the chi-squared distribution arises in the context of the Likelihood Ratio Test (LRT). Let Y be distributed like a chi-squared with k degrees of freedom:

$$Y \sim \chi^2_k, \tag{S81}$$

where the PDF is:

$$p_Y(y|k) = \frac{1}{2^{k/2}\Gamma(k/2)} y^{k/2-1} e^{-y/2}, \tag{S82}$$

where Γ is the gamma function. The CDF is therefore:

$$\begin{aligned} P_Y(y|k) &\equiv \Pr\{Y < y|k\}, & \text{(S83)}\\ &= \int_0^y \mathrm{d}y'\, p_Y(y'|k), & \text{(S84)}\\ &= \int_0^x \mathrm{d}x'\, p_X(x'|\tfrac{k}{2}, 2), & \text{(S85)}\\ &= \gamma^-(\tfrac{k}{2}, \tfrac{y}{2}), & \text{(S86)} \end{aligned}$$

where p_X is the PDF of the gamma distribution (Eq. S73) and γ^- is the regularized lower incomplete gamma function. The survival function is:

$$\Pr\{Y > y|k\} = \gamma^+(\tfrac{k}{2}, \tfrac{y}{2}), \tag{S87}$$

where γ^+ is the regularized upper incomplete gamma function.

10. DESCRIPTION OF SUPPLEMENTARY DATA

A. Data Tables

Data S1: Overabundance for all genes as measured by TFNseq analysis. The original TFNseq data was previously generated by the Manoil lab [18]. Format: Open Document Format (ODS).

Data S2: A list of essential genes ranked by overabundance. Format: Open Document Format (ODS).

Data S3: Representative single-cell imaging-based cell cytometry data for wild-type *A. baylyi* proliferating on minimal media (Km$^-$) from a single progenitor cell (Sec. 4 D 2). Format: Open Document Format (ODS).

Data S4: Representative single-cell imaging-based cell cytometry data for *A. baylyi* Δ*IS* proliferating on minimal media (Km$^+$) from a single progenitor cell in a knockout-depletion experiment (Sec. 4 D 2). Format: Open Document Format (ODS).

Data S5: Representative single-cell imaging-based cell cytometry data for *A. baylyi* Δ*dnaA* proliferating on minimal media (Km$^+$) from a single progenitor cell in a knockout-depletion experiment (Sec. 4 D 3). Format: Open Document Format (ODS).

Data S6: Representative single-cell imaging-based cell cytometry data for *A. baylyi* Δ*dnaN* proliferating on minimal media (Km$^+$) from a single progenitor cell in a knockout-depletion experiment (Sec. 4 D 4). Format: Open Document Format (ODS).

Data S7: Representative single-cell imaging-based cell cytometry data for *A. baylyi* Δ*murA* proliferating on minimal media (Km$^+$) from a single progenitor cell in a knockout-depletion experiment (Sec. 4 D 6). Format: Open Document Format (ODS).

Data S8: Representative single-cell imaging-based cell cytometry data for *A. baylyi* Δ*ftsN* proliferating on minimal media (Km$^+$) from a single progenitor cell in a knockout-depletion experiment (Sec. 4 D 5). Format: Open Document Format (ODS).

B. Annotated sequences

Data S9: The annotated sequence of the DnaN fluorescent fusion *YPet-dnaN*. Format: Genbank file.

C. Supplemental movies

Movie S1: Wild-type *A. baylyi* proliferating on minimal media (Km$^-$). Frame rate: 1 frame/2 min. (Sec. 4 D 2.) Raw images. Youtube.

Movie S2: Wild-type *A. baylyi* proliferating on minimal media (Km$^-$). Frame rate: 1 frame/2 min. (Sec. 4 D 2.) Annotated/segmented images. Youtube.

Movie S3: *A. baylyi* Δ*IS* proliferating on minimal media (Km$^+$) in a knockout-depletion experiment. Frame rate: 1 frame/3 min. (Sec. 4 D 2.) Raw images. Youtube.

Movie S4: *A. baylyi* Δ*IS* proliferating on minimal media (Km$^+$) in a knockout-depletion experiment. Frame rate: 1 frame/3 min. (Sec. 4 D 2.) Annotated/segmented images. Youtube.

Movie S5: *A. baylyi* Δ*dnaA* proliferating on minimal media (Km$^+$) in a knockout-depletion experiment. Frame rate: 1 frame/2 min. (Sec. 4 D 3.) Raw images. Youtube.

Movie S6: *A. baylyi* Δ*dnaA* proliferating on minimal media (Km$^+$) in a knockout-depletion experiment. Frame rate: 1 frame/2 min. (Sec. 4 D 3.) Annotated/segmented images. Youtube.

Movie S7: *A. baylyi* Δ*dnaN* proliferating on minimal media (Km$^+$) in a knockout-depletion experiment. Frame rate: 1 frame/9 min. (Sec. 4 D 4.) Raw images. Youtube.

Movie S8: *A. baylyi* Δ*dnaN* proliferating on minimal media (Km$^+$) in a knockout-depletion experiment. Frame rate: 1 frame/9 min. (Sec. 4 D 4.) Annotated/segmented images. Youtube.

Movie S9: *A. baylyi* Δ*murA* proliferating on minimal media (Km^+) in a knockout-depletion experiment. Frame rate: 1 frame/2 min. (Sec. 4 D 6.) Raw images. Youtube.

Movie S10: *A. baylyi* Δ*murA* proliferating on minimal media (Km^+) in a knockout-depletion experiment. Frame rate: 1 frame/2 min. (Sec. 4 D 6.) Annotated/segmented images. Youtube.

Movie S11: *A. baylyi* Δ*ftsN* proliferating on minimal media (Km^+) in a knockout-depletion experiment. Frame rate: 1 frame/2 min. (Sec. 4 D 5.) Raw images. Youtube.

Movie S12: *A. baylyi* Δ*ftsN* proliferating on minimal media (Km^+) in a knockout-depletion experiment. Frame rate: 1 frame/2 min. (Sec. 4 D 5.) Annotated/segmented images. Youtube.