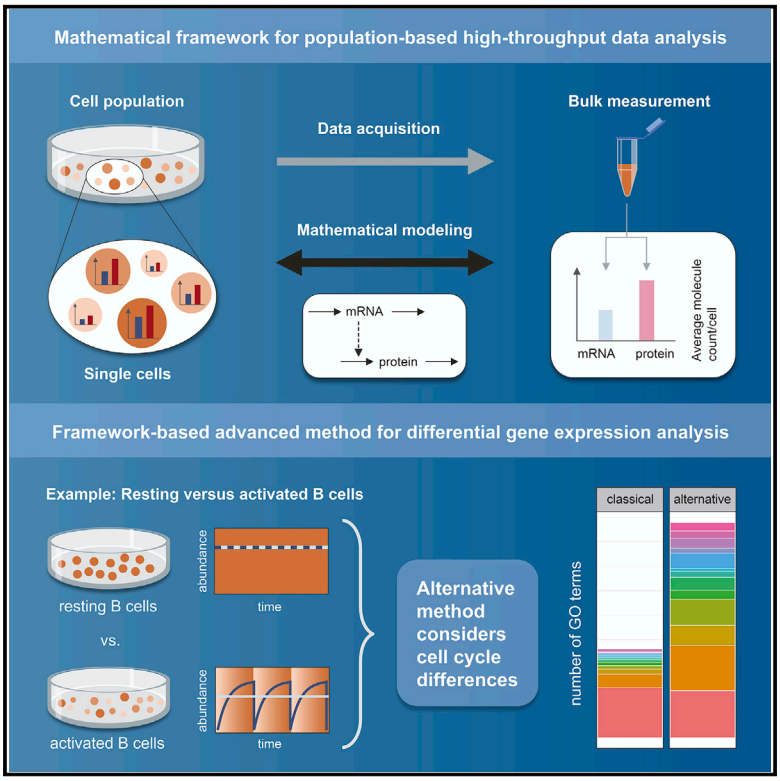
Cell Systems

Report

Of Gene Expression and Cell Division Time: A Mathematical Framework for Advanced Differential Gene Expression and Data Analysis

Graphical Abstract



Highlights

- Advanced differential gene expression analysis for populations of dividing cells
- Quantitative framework for the integration of multiple omics data
- Easy-to-use formulas link single-cell gene expression to population-level data
- Application to gene expression data of B cell activation shows a very high hit rate

Authors

Katharina Baum, Johannes Schuchhardt, Jana Wolf, Dorothea Busse

Correspondence

katharina.baum@mdc-berlin.de (K.B.), jana.wolf@mdc-berlin.de (J.W.), dorotheabusse@yahoo.de (D.B.)

In Brief

We provide an easy-to-use quantitative framework that links rates of single-cell gene expression to population-level data such as abundances measured by RNA sequencing or mass spectrometry. For populations of dividing cells, this framework integrates multiple layers of omics data for differential gene expression analysis and predicts when cell division is critical in this analysis. Using published human B cell data, we show that the sensitivity of differential gene expression analysis improves noticeably when comparing rates of gene expression instead of abundances.

Baum et al., 2019, Cell Systems 9, 1–11
September 25, 2019
https://doi.org/10.1016/j.cels.2019.07.009

CellPress

Please cite this article in press as: Baum et al., Of Gene Expression and Cell Division Time: A Mathematical Framework for Advanced Differential Gene Expression and Data Analysis, Cell Systems (2019), https://doi.org/10.1016/j.cels.2019.07.009

Cell Systems
Report

Of Gene Expression and Cell Division Time: A Mathematical Framework for Advanced Differential Gene Expression and Data Analysis

Katharina Baum,[1,2,*] Johannes Schuchhardt,[3] Jana Wolf,[1,5,*] and Dorothea Busse[1,4,*]
[1]Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Robert-Rössle-Str. 10, 13125 Berlin, Germany
[2]Luxembourg Institute of Health, 1A-B rue Thomas Edison, L-1445 Strassen, Luxembourg
[3]MicroDiscovery GmbH, Marienburgerstr. 1, 10405 Berlin, Germany
[4]Integrative Research Institute for the Life Sciences, Humboldt University Berlin, Philippstr. 13, 10115 Berlin, Germany
[5]Lead Contact
*Correspondence: katharina.baum@mdc-berlin.de (K.B.), jana.wolf@mdc-berlin.de (J.W.), dorotheabusse@yahoo.de (D.B.)
https://doi.org/10.1016/j.cels.2019.07.009

SUMMARY

Estimating fold changes of average mRNA and protein molecule counts per cell is the most common way to perform differential expression analysis. However, these gene expression data may be affected by cell division, an often-neglected phenomenon. Here, we develop a quantitative framework that links population-based mRNA and protein measurements to rates of gene expression in single cells undergoing cell division. The equations we derive are easy-to-use and widely robust against biological variability. They integrate multiple "omics" data into a coherent, quantitative description of single-cell gene expression and improve analysis when comparing systems or states with different cell division times. We explore these ideas in the context of resting versus activated B cells. Analyzing differences in protein synthesis rates enables to account for differences in cell division times. We demonstrate that this improves the resolution and hit rate of differential gene expression analysis when compared to analyzing population protein abundances alone.

INTRODUCTION

Gene expression is a central process in living organisms and its characterization reveals important insights on cellular regulation. High-throughput measurement techniques, such as RNA sequencing and mass spectrometry, are very suitable to study gene expression on genomic scale. Therefore, the extent of these data has steadily increased over the past years, also due to new biological insights, improved analysis tools, and technological progress (Aebersold and Mann, 2016; Lowe et al., 2017). Population-based high-throughput data are the basis of many recent innovations in biomedical and pharmaceutical research (Macarron et al., 2011). Hence, they are a valuable source of information and remain, also in the arising era of single-cell multi-omics, the state-of-the-art approaches for new drug discovery and perturbation screens (Janzen, 2014). Methods that enhance the degree of information derived from these data are still desirable.

Here, we introduce a differential gene expression analysis method that delivers a complementary, more comprehensive picture of the changes in gene expression than the classical method of comparing abundances does. It is especially suitable to compare large gene expression datasets gathered from cellular states or cell types that strongly differ in their cell cycle duration. This method can therefore be applied in particular to populations of proliferating cells, e.g., mammalian cell cultures. It enables but also requires the combination of population-based high-throughput datasets.

Population-based high-throughput techniques rely on bulk measurements, but gene expression is taking place on the level of the individual cells forming the population. Therefore, the question arises how data generated as average molecule counts of mRNAs and proteins per cell relate meaningfully to gene expression in single cells. In order to illustrate the challenges that have to be met for deducing this relation in populations of proliferating cells, we first focus on the gene expression in individual cells. In proliferating cells, the molecule counts of mRNAs and proteins increase during the cell cycle before the cellular transcriptome and proteome is distributed between the two daughter cells during cell division (Figure 1A, left). The cell cycle sets the time frame for this recurrent process of increase and distribution of cellular mRNA and protein abundances. Therefore, the permanent change of single-cell gene expression over time depends on the duration of the cell cycle.

As the molecule counts per cell depend on cell proliferation also the population averages do. This becomes evident if considering a population of resting, non-proliferating cells (Figure 1A, right). The mRNA and protein molecule count does not change over time. Synthesis and degradation are balanced and subsequently gene expression is in steady state. Assuming a population of identical cells, its average molecule counts equal the cellular steady-state abundances of mRNAs and proteins (shown for mRNA; Figure 1A, right). However, as soon as cells proliferate, mRNA and protein abundances are not in steady state anymore, the population average molecule count of mRNA and protein has no corresponding cellular parameter (shown for mRNA; Figure 1A, left).

Please cite this article in press as: Baum et al., Of Gene Expression and Cell Division Time: A Mathematical Framework for Advanced Differential Gene Expression and Data Analysis, Cell Systems (2019), https://doi.org/10.1016/j.cels.2019.07.009

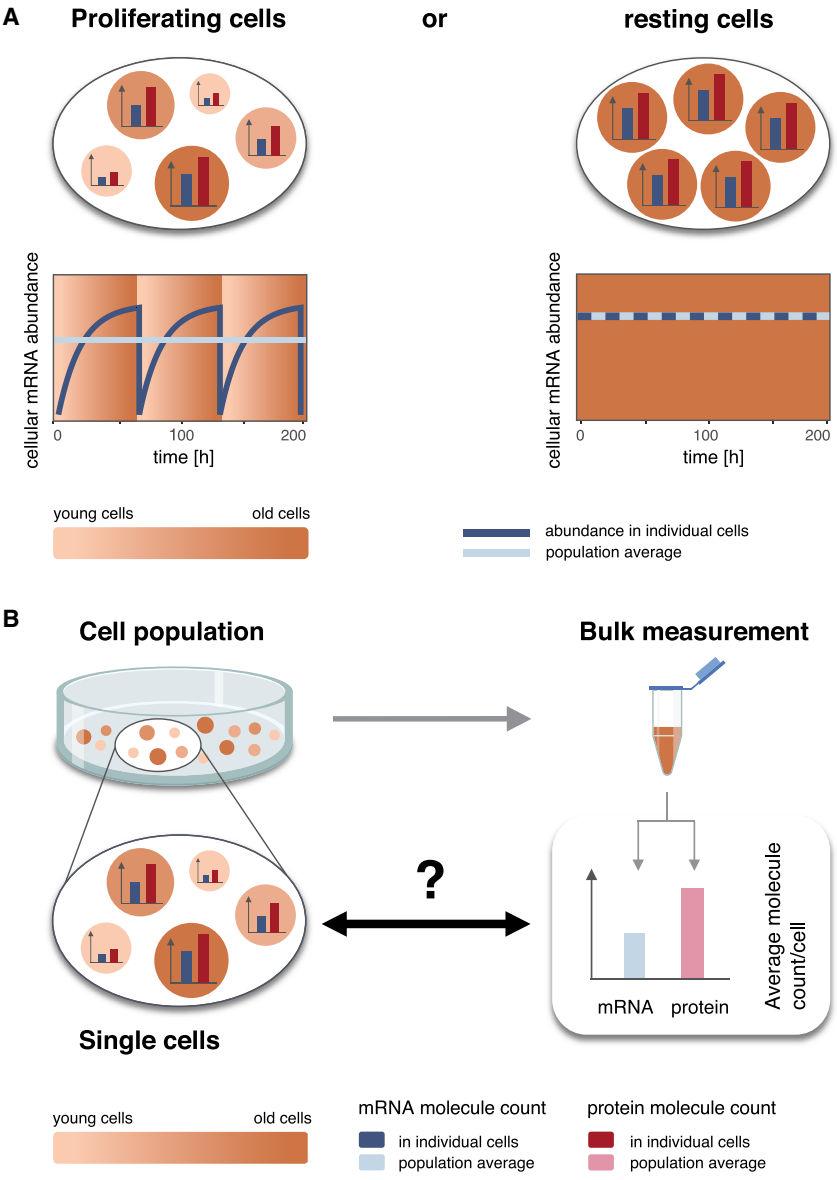


Figure 1. Relation between Population Average mRNA and Protein Measurements and Single-Cell Gene Expression

(A) Snapshot of mRNA and protein abundances in a proliferating (top left) or resting (top right) cell population. Populations of proliferating cells (left) contain cells of different age (shades of brown) and age-dependent mRNA (blue) and protein (red) counts. In a proliferating single cell, gene expression varies over time as shown here schematically for the mRNA abundance (Figure 1A, bottom left, dark blue); additionally, the corresponding population average (Figure 1A, bottom left, light blue) is given. In contrast, in resting cells, mRNA and protein counts do not vary within a population (Figure 1A, bottom right). Here, abundances in single cells coincide with the population average. This illustrates that the cell division time has a strong influence on the interplay between population average and single-cell abundances.

(B) Populations of proliferating cells (left) contain cells of different ages and subsequently age-dependent mRNA and protein counts. Bulk measurements of such a population deliver population average mRNA or protein abundances measured in molecule count per cell. In this study, we focus on the question how gene expression, a cellular process, can be characterized by population average measurements.

Consider now a snapshot of a growing population, e.g., at the time point of harvesting cells for gene expression analysis based on high-throughput measurements (Figure 1B). The cells differ in their molecule counts of mRNA and protein since the population is composed of cells in different stages of the cell cycle or rephrased at different ages. This information about the distribution of mRNA and protein abundances among the individual cells (Figure 1B, left) is crucial for deducing gene expression characteristics in single cells, i.e., rates of gene expression; however, it is lost in the process of mRNA and protein extraction required for RNA sequencing and mass spectrometry measurements (Figure 1B, right).

In summary, the above quoted question, "How can the measured population averages be related to gene expression in single cells?" leads to the following more specific questions: (1) How can the highly dynamic process of gene expression be characterized in single cells, thereby including not only the kinetic parameters which govern the rates of gene expression but also the cell cycle duration? (2) How can the age distribution be described in proliferating cell systems to identify the contribution of each single cell to the population average mRNA and protein count per cell?

The effect of the cell division on gene expression has been taken into account in earlier studies by us and others (Schwanhäusser et al., 2011; Miller et al., 2011; Eden et al., 2011). Here, we build on and further develop these studies by considering mRNA and protein dynamics together, clearly distinguishing between intracellular gene expression dynamics and population growth and accounting for the age distribution of cells in a population. We focus our study on populations of resting or steadily growing cells that are otherwise in a stable state, e.g., a differentiated condition, or an activated state. We assume that regulations are steady and settled and, in particular, that the parameters of gene expression are constant for a certain state or condition of the population.

We start by representing gene expression in single cells. We use a basic, well-established, linear ordinary differential equation model, which incorporates four different rates: transcription, translation, and mRNA and protein degradations (Hargrove and Schmidt, 1989; Alon, 2006; Legewie et al., 2008; Schwanhäusser et al., 2011). It is formulated in terms of mRNA and protein molecules per cell in order to enable a direct relationship to the molecules-per-cell-based high-throughput measurement output. In addition, when using this formulation,

Please cite this article in press as: Baum et al., Of Gene Expression and Cell Division Time: A Mathematical Framework for Advanced Differential Gene Expression and Data Analysis, Cell Systems (2019), https://doi.org/10.1016/j.cels.2019.07.009

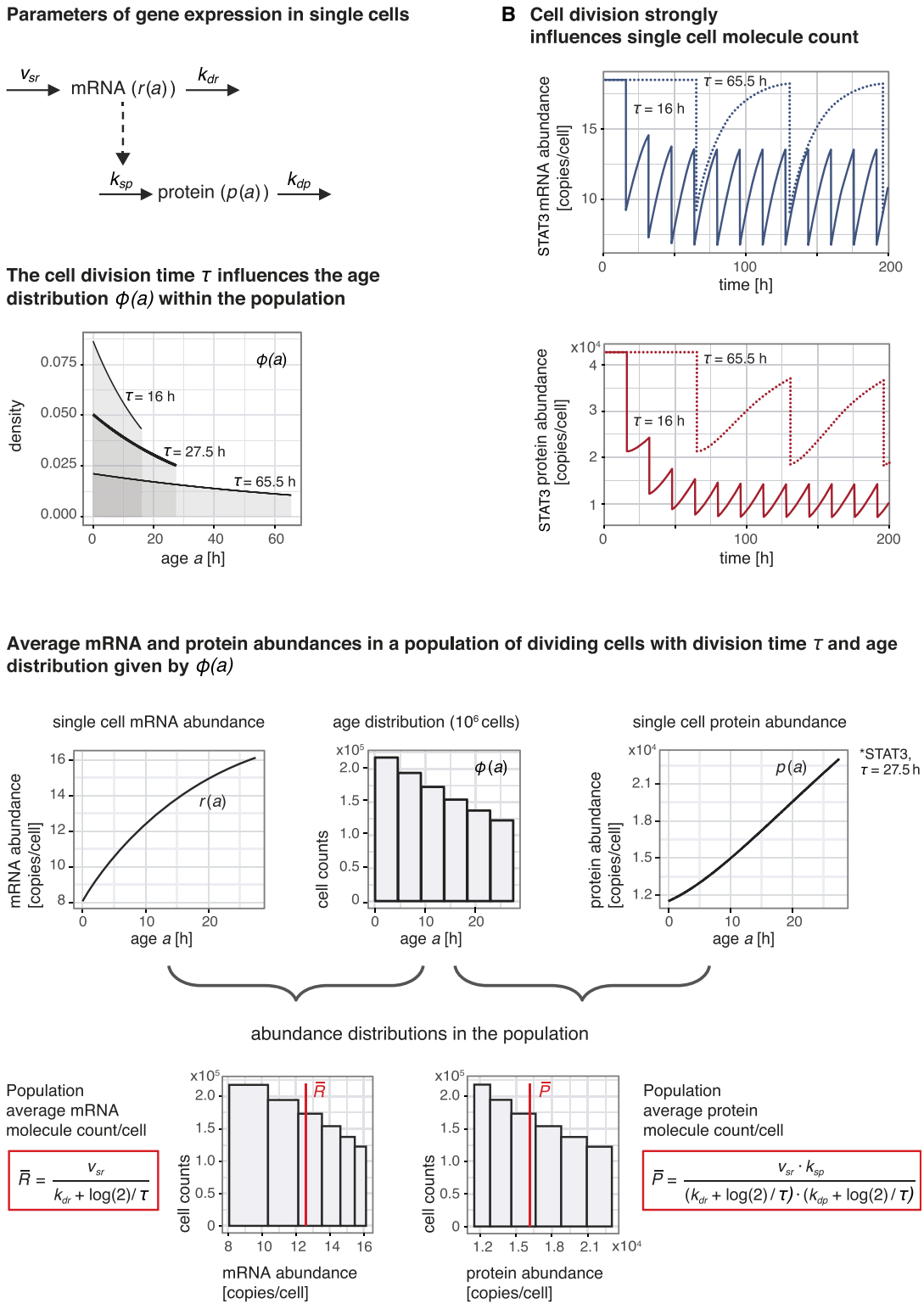


Figure 2. From Dynamic Single-Cell Gene Expression to Population Average of mRNA and Protein Abundance

(A) Scheme of a basic gene expression model realized by ordinary differential equations. The mRNA (r) is translated with rate v_{sr} and degraded with the rate constant k_{dr}. The protein (p) is translated proportionally to the mRNA abundance with the rate constant k_{sp} and degraded with the rate constant k_{dp}. The age of the cell is denoted by a. The system is formulated to deliver mRNA and protein molecule counts and can be solved analytically (STAR Methods).

(B) Simulation of STAT3 mRNA and protein dynamics (top and bottom, respectively) according to the model in (A) for a cell cycle duration τ of 16 h (human embryonic stem cells [Becker et al., 2006], solid lines) and 65.5 h (human precursor B cells [Cooperman et al., 2004], dotted lines) for a single cell undergoing

(legend continued on next page)

Please cite this article in press as: Baum et al., Of Gene Expression and Cell Division Time: A Mathematical Framework for Advanced Differential Gene Expression and Data Analysis, Cell Systems (2019), https://doi.org/10.1016/j.cels.2019.07.009

assumptions on temporal or compartmental volume dependencies in mammalian cells are not necessary. On this basis, we derive easy-to-use formulas, which enable concluding straightforwardly from population averages on single-cell kinetic gene expression parameters of growing cell populations. We show that the derived formulas are widely robust also for populations of non-identical cells. In addition, we find that it is important to incorporate cell division to quantitatively link population average to single-cell gene expression, especially when mRNA and protein half-lives and cell division time are in the same order of magnitude. This is often the case in mammalian cell culture systems and disease state. We therefore dedicate this study especially to this kind of systems.

Our proposed framework allows for differential gene expression analysis of population-based high-throughput data by comparing the condition- or state-specific parameters of the single-cell gene expression instead of population average abundances. This alternative method can detect changes in gene expression even if mRNA or protein population averages remain unaltered as illustrated in an example comparing protein expression in resting and activated human memory B cells (Rieckmann et al., 2017). Differences in synthesis rates between the resting and activated state reveal many changes in key cellular processes such as the immune system regulation that are not disclosed by analyzing changes in population-based abundances alone, and which provide additional targets for pharmaceutical and biomedical research.

RESULTS

Population Averages of mRNAs and Proteins Relate to Mammalian Single-Cell Gene Expression

The developed framework and subsequently the mathematical description of gene expression in a single cell are based on the classical idea of hierarchical organization of gene expression (Figure 2A). The presented scheme is translated into an ordinary differential equation system (STAR Methods). In the model, the mRNA (r) is produced by a constant rate v_{sr} and degraded proportionally to its molecule count with the degradation rate constant k_{dr}. The mRNA is translated into a protein (p) with the rate constant k_{sp}, and the protein is degraded proportionally to its molecule count with the rate constant k_{dr}. We assume that these parameters are constant and specific for each cellular condition and a given mRNA-protein pair but can differ between conditions, e.g., between the resting and activated state of immune cells. Due to the simplicity of the mathematical model, several biological processes are condensed in one parameter. For example, the transcription rate constant (v_{sr}) summarizes all processes from transcription initiation to mRNA processing in the cytoplasm. For this ordinary differential equation system, an analytical solution can be derived (STAR Methods). For the purpose of this analysis, we consider the temporal progression of gene expression within a cell as a function of its age (a). The single-cell mRNA and protein abundances are therefore denoted as $r(a)$ and $p(a)$, respectively. At age zero cell division has just occurred and at an age that equals the cell cycle duration τ the cell divides. We represent cell division as instantaneous process, in which the cellular mRNA and protein amount is set to half (scheme Figure 1A, left). To ensure that cellular mRNA and protein levels can be maintained, the solutions $r(a)$ and $p(a)$ fulfill the additional assumptions that the mRNA and protein abundances double over the duration of one cell cycle.

To illustrate the influence of the cell cycle duration we consider, as an example, the kinetics of STAT3 (signal transducer and activator of transcription 3) mRNA and protein (Schwanhäusser et al., 2011) expression for one cell undergoing several cell divisions (Figure 2B, top and bottom, respectively). Cell division has profound consequences on the mRNA and protein dynamics as the simulation of STAT3 shows for two different cell cycle durations, which represent the cell division time of human embryonic stem cells (16 h) (Becker et al., 2006) and human precursor B cells (65.5 h) (Cooperman et al., 2004) but otherwise fixed kinetic parameters of gene expression (Table S1). Starting the simulation at the steady state defined by the kinetic parameters, STAT3 mRNA and STAT3 protein abundances do not change during the first cell cycle. As soon as the cell divides, the steady state is left and is not reached again. The cellular dynamics develop toward the solution of the system that fulfills the assumption of doubling mRNA and protein abundances during one cell cycle ($2r(0) = r(\tau)$ and $2p(0) = p(\tau)$) and keeps this dynamic state (see also Figure S1 for other mRNA-protein pairs). Thereby, the shorter the cell division time, the more distant this dynamic state is from the steady state (Figure 2B). This is especially true for changes in protein abundances because the protein half-life is often longer than mRNA half-lives (here, STAT3 mRNA 12.8 h and STAT3 protein 22.1 h (Schwanhäusser et al., 2011); see also Cambridge et al., 2011; Tani et al., 2012).

To be able to relate the gene expression dynamics of a dividing single cell as derived above to measurements of population average mRNA and protein abundance, we consider the situation of a cell culture following exponential growth. Each culture dish contains a population of cells of all cell cycle phases or, rephrased, of all ages (compare Figure 1B). The age distribution of an exponentially growing population has been studied by Powell

multiple divisions. Steady state abundances are assumed during the first cell cycle, abundances are halved at cell division. Due to differences in the cell division times, the dynamics differ even though the kinetic parameters are identical. Kinetic parameter values are given in Table S1.

(C) The mathematically derived age distributions of exponentially growing cell populations (Powell, 1956) for different cell cycle durations τ are shown (human embryonic stem cells: τ = 16 h, NIH3T3 cells: τ = 27.5 h, human precursor B cells: τ = 65.5 h).

(D) The population abundance distributions of mRNA and protein can be derived from the single-cell dynamics given a certain age distribution. STAT3 expression serves as an example. The single-cell mRNA $r(a)$ and protein $p(a)$ dynamics (top left and right, respectively) and the age distribution as in (C) of a population of 10^6 cells with a cell division time of 27.5 h (top middle) gives the STAT3 mRNA and protein distributions (bottom). Red lines indicate the population averages of mRNA and protein expression, $\overline{R}$ and $\overline{P}$, respectively. These population average mRNA and protein abundances are derived as the expected values of the distributions (equations in red boxes, bottom left and right, respectively) and are linked to the single-cell kinetic parameters of gene expression (v_{sr}, k_{dr}, k_{sp}, and k_{dp}) and the cell division time (τ). The analytical derivations are given in the STAR Methods.

See also Figures S1, S2, and S3 and Tables S1 and S2.

(Powell, 1956). He found the age distribution to be stable, i.e., not changing in time, in exponentially growing cell cultures. Powell also derived an analytical description of the age distribution relying solely on the assumption of exponential population growth (Figure 2C; a simulation showing the convergence to this age distribution is shown in Figure S2). The age distribution is not homogeneous; there are more young cells than old cells. The heterogeneity of the age distribution depends on the duration of the cell cycle, the distribution of fast dividing cells being more heterogeneous than of cells with a very long cell cycle (Figure 2C). For the latter, it is close to a uniform age distribution, with equal percentages of cells of each age. The heterogeneous age distribution results in heterogeneity of cellular mRNA and protein abundances in a growing population (Figure 2D).

To link the single-cell gene expression to population measurements of mRNA and protein levels, the stable age distribution and its analytical description are crucial. For cells from an exponentially growing cell culture, i.e., under steady growth, the age distribution and the mRNA and protein kinetics of STAT3 (Figure 2D, top) determine the distributions of STAT3 mRNA and protein abundances in this cell population (Figure 2D, bottom). The abundance distributions are again stable. Please note that due to the non-linear cellular mRNA and protein kinetics, the bin widths of the mRNA and protein histograms are not equal when maintaining equal bin width of the age histogram. Adding all mRNA or protein abundances weighted by their percentage of occurrence, or more precisely integrating over their product as a function of the continuous variable age, gives us the expected values of these distributions, i.e., the average population abundances (Equations, Figure 2D, red boxes). The average mRNA and protein abundances of an exponentially growing cell population (Equations, Figure 2D) are given by the product of rate constants of the producing reactions divided by those of the depleting reactions. Thus, they resemble closely their respective steady-state equations, and without cell division (i.e., $\tau \rightarrow \infty$), they even converge to the latter (STAR Methods). However, owing to cell division, in general the mRNA and protein loss not only depends on the degradation rate constant but also on the inverse of the cell cycle duration ($\log(2)/\tau$), the so-called dilution rate constant (Eden et al., 2011). The latter accounts for the redistribution of mRNA and protein between the two daughter cells during cell division.

Up to now, we have considered an idealized situation in which all cells in the population are identical in the parameters of gene expression and cell division time. Since variability is often large in biological systems, we estimated the robustness of the derived quantitative link between population average mRNA and protein abundances and single-cell kinetic parameters of gene expression (Equations, Figure 2D) toward biological variability of cells of the population. We found that even strong variability of 50% in the cell division times and the associated change in the age distribution of the population introduces errors in our derivation of below 10% (STAT3, see Figure S2E). In addition, we examined the effect of combined variations as large as 30% in the kinetic parameters of gene expression and initial conditions together with variations of the cell division time of 15% between cells of a considered population (see Figures S3A and S3B; Table S2). On average, these variations led to deviations from our estimates for populations of identical cells of less than 6% for mRNA and 13% for protein. Deviations up to 32% were only observed for the highly unstable MDM2 (E3 ubiquitin-protein ligase). This indicates a high robustness of our derivations to variability between cells of the population.

In summary, in a cell population of dividing but otherwise identical cells, an easy-to-use analytical solution exists that describes the relationship between population average mRNA and protein abundances, the kinetic parameters of single-cell gene expression and cell division. The link could be made because the age distribution in a steadily dividing cell population is stable and can be described analytically (Powell, 1956).

The Influence of the Cell Division Time on Gene Expression Analysis

The cell cycle duration enters the above-derived equations (Figure 2D) in the denominator within the sum of its inverse and the degradation rate constants of either mRNA or protein; its influence on the average mRNA and protein population abundance is therefore non-linear, dependent on the magnitude of the kinetic parameters of gene expression and subsequently difficult to predict.

The influence of the cell division time on the interplay of single-cell gene expression dynamics and the average population abundance measurements can be assessed by analyzing the deviation of the average population mRNA and protein abundances under exponential population growth from their respective steady-state values, e.g., in resting cells. In the latter case, the population average equals the single-cell steady-state abundance (compare Figure 1A, right) and is therefore by definition independent of the cell cycle duration. The ratio between the steady-state value and the average population abundance under exponential population growth depends on the cell division time. In addition, it depends on the half-life of the mRNA for the mRNA abundance or on both mRNA and protein half-lives for the protein abundance (Figure 3A; STAR Methods). This ratio is close to one if the average abundance under exponential population growth and the average abundance in steady state are approximately equal; in this case, the influence of the cell division on gene expression is negligible (Figure 3A, gray). With increasing ratio, the influence of cell division increases (Figure 3A, red and dark red).

It becomes evident that the influence of cell division increases with increasing half-life of either mRNA (Figure 3A, top row) or mRNA and protein (Figure 3A bottom row). MDM2, STAT3, and RPS3 (40S ribosomal protein S3) are chosen as examples that have small, medium, and large half-lives, respectively, for both mRNA and protein (Figure 3 left, middle, and right column, respectively). For the very unstable MDM2 mRNA and protein, the importance of incorporating cell division is small regardless of the considered cell cycle duration (gray area in Figure 3A, left). For the very stable RPS3, the influence of cell division is pronounced (dark red to red area in Figure 3A, right). The same holds true for STAT3 protein, which has an intermediate half-life (Figure 3 middle, bottom). For STAT3 mRNA, the importance of considering cell division strongly depends on the actual cell cycle duration. For fast cell division, the cell cycle duration strongly governs gene expression of both mRNA and protein, whereas in slowly proliferating cells, its influence is negligible (Figure 3 middle, top).

Please cite this article in press as: Baum et al., Of Gene Expression and Cell Division Time: A Mathematical Framework for Advanced Differential Gene Expression and Data Analysis, Cell Systems (2019), https://doi.org/10.1016/j.cels.2019.07.009

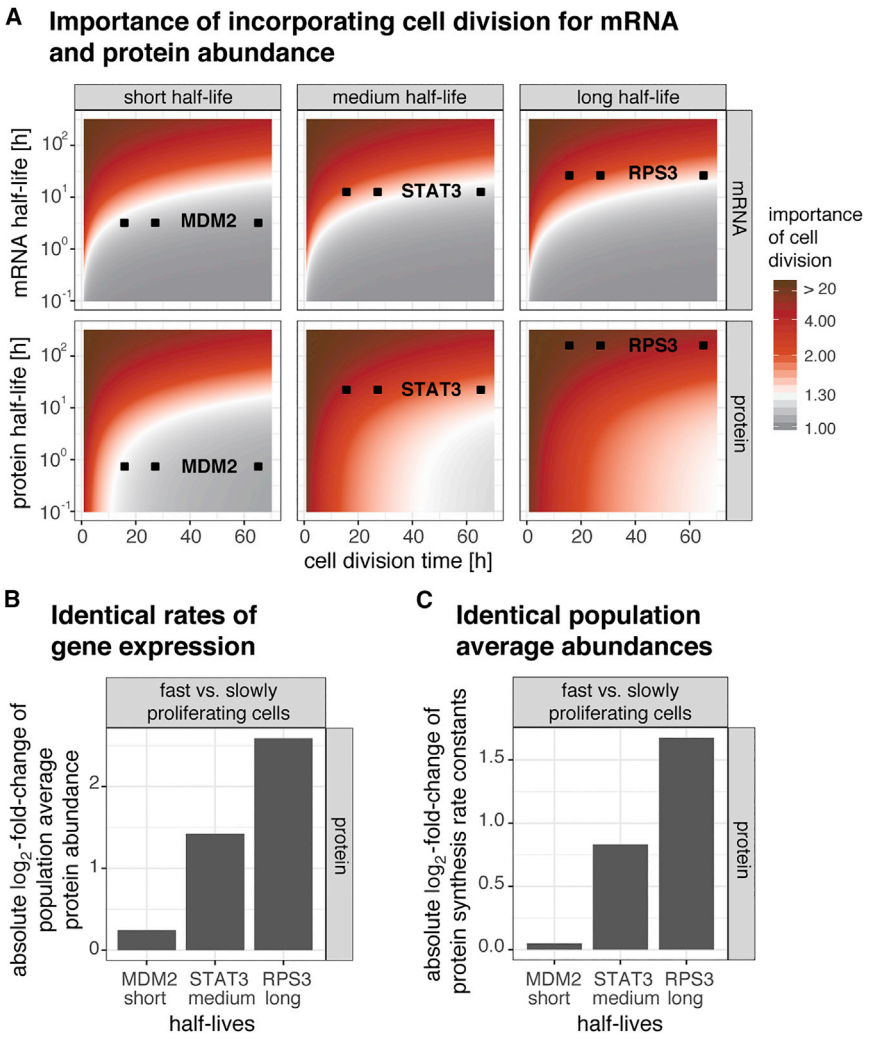


Figure 3. The Importance of Considering Cell Division in Gene Expression Analysis Depends on the Half-Lives

(A) To illustrate the influence of cell division on the population average abundance of mRNA (top) and protein (bottom), we computed the ratio of (1) neglecting cell division, i.e., assuming single cells are in steady state, and (2) incorporating the effect of cell division according to our formulas (Figure 2D, see also STAR Methods, Equations 19 and 20). Kinetic parameters are considered to be identical for both approaches. The fold changes of the average abundances calculated as in (1) and (2) are given in color code and in dependence on the cell division time and the half-lives. For proteins, we show the combinations with short, intermediate or long mRNA half-lives (left, middle, and right, respectively). Gray indicates strong similarity between (1) single-cell steady state versus (2) incorporating cell division time, red denotes strong differences. Therefore, red areas mark the cases where an incorporation of cell division is highly important. As sample mRNAs and proteins (black symbols) we present STAT3 (left), MDM2 (middle) and RPS3 (right) for the three cell division times of 16 h, 27.5 h, and 65.5 h. Only for the very unstable MDM2, the mRNA and protein average abundances incorporating cell division is less important. Kinetic parameter values are given in Table S1.

(B) Absolute $\log_2$-fold changes of population average protein abundances between two cell systems with different cell division times (τ = 16 h versus τ = 65.5 h) for otherwise identical rates of gene expression. Population averages were computed for MDM2, STAT3, and RPS3 (see Table S1) according to equation in Figure 2D, right. Population averages can differ strongly even if the gene expression characteristics of the underlying single cells, i.e., their kinetic parameters, are the same.

(C) Absolute $\log_2$-fold changes of protein synthesis rate constants between two cell systems with different cell division times (τ = 16 h versus τ = 65.5 h) for identical mRNA and protein population averages and identical rate constants of protein degradation. Rate constants were computed for MDM2, STAT3, and RPS3 (see Table S1) according to the transformed equation Figure 2D, right (STAR Methods Equation 17; see also Equation 26). Even if population averages are the same between conditions, rates of gene expression of the underlying single cells can differ strongly between cells with different cell division times.

In general, if the degradation rate constants are much larger than the dilution rate constant, $\log(2)/\tau$, the influence of cell division is small. These conditions are satisfied if (1) the cells do not divide (the cell cycle duration approaches infinity) or (2) the half-lives are much smaller than the cell division time. Conversely, if the cell division time is comparable to or smaller than the half-lives, its influence is strong. Therefore, within a transcriptome or proteome of a cell the relative importance of degradation and dilution by cell division on gene expression varies (Eden et al., 2011).

The dependence of the average population mRNA and protein molecule counts on the cell cycle duration has severe consequences for differential gene expression data analysis, especially when comparing data gathered from cell systems with different cell division times. In this case, it is not possible to directly conclude from altered population average molecule counts on altered gene expression characteristics, i.e., rates of gene expression in the single cells, or vice versa. To illustrate this fact, we use again the sample proteins presented in Figure 3A and estimate the $\log_2$-fold changes in their population average abundances for two cell systems with different cell division times (fast, 16 h; slow, 65.5 h) but otherwise identical parameters of gene expression (Figure 3B). Consequently, the observed fold changes result exclusively from differences in the cell cycle duration. Depending on the stability of mRNA and protein, the absolute $\log_2$-fold change ranges from small difference (0.24, MDM2) up to 2.6 (RPS3).

One can also consider the situation from a different perspective. Observing similar or identical population average abundances between cell systems with different cell cycle durations does not necessary imply similar gene expression characteristics in the single cells. The loss of mRNA and protein molecules by cell division is enhanced in fast dividing cells compared to slowly dividing or resting cells and has to be compensated to maintain population average mRNA and protein molecule counts. A possibility to compensate for enhanced cellular

Box 1. Gene Expression in Resting versus Activated Human B Cells

Our easy-to-use mathematical expressions (Figure 2D) provide a powerful tool for advanced data analysis, as we demonstrate in the following example. For this purpose, we combined a dataset published by Rieckmann and colleagues (Rieckmann et al., 2017), quantifying the transcriptome and proteome from healthy human donors before and after activation of memory B cells, with protein half-life measurements in resting B cells (Mathieson et al., 2018). Resting B cells are very long lived and cell division occurs in the order of years, whereas activated B cells divide rapidly within hours (Jones et al., 2015; Milo et al., 2010). Therefore, this example is especially suitable to demonstrate the strong link between cellular gene expression, the population average protein molecule count per cell and the cell cycle duration (Figures 2 and 3), and the resulting consequences for differential gene expression analysis.

For our analysis, we used 2,438 genes for which protein abundances, mRNA abundances and protein half-lives were available (STAR Methods). From this gene set, we performed two differential gene expression analyses comparing resting and activated B cells. In the first, we determined the subset of proteins, which showed significantly different protein abundances between the conditions, termed the classical method. In the second, alternative method, we applied the derived mathematical expressions (Figure 2D) to estimate changes in the single-cell translation rates (Figure 4A, left; see STAR Methods). Approximately 15% (355/2438) of the proteins had significantly changed abundances after activation; the vast majority is up-regulated. For the protein expression dynamics, we find that approximately 73% (1768/2438) of the translation rate constants significantly changed; again, enhanced synthesis predominates (Figure 4A middle and Figure S4A). Almost all genes with changes in protein abundance also show altered protein synthesis. However, the majority of proteins with altered synthesis exhibit no significant change in abundance (Figure 4A, right). This suggests that the altered translation rates compensate protein loss due to enhanced cell division in activated state. Therefore, if cell populations differ in cell division time, comparing their population average protein or mRNA abundances provides only a limited picture of changes in the gene expression of the underlying single cells. In this example, we identified numerous cases of altered translation and thereby extended the possibilities for experimental validation and targeted drug manipulation by approximately 60% of all considered proteins.

To further illustrate the value of estimating differences in the rates of gene expression between resting and activated B cells, we performed gene ontology (GO) enrichment analyses using PANTHER (Mi et al., 2017). We identified the sets of significantly enriched GO terms for the subsets of genes with significantly altered protein abundances (classical) or translation rates (alternative), respectively (Figure 4B, left; for details see STAR Methods). Both sets were clustered separately using a semi-automated procedure and were aggregated into categories, which are related but not identical to parental GO terms (Tables S3 and S4; for details see STAR Methods). In line with the differences in numbers observed in Figure 4A, more GO terms are characteristic for changed synthesis rates than for altered protein abundances (Figure 4B, left). However, all categories - except for "actin polymerization and cell size regulation" - occur in both enrichment analyses. The increase in GO terms therefore results from a decisively larger number of terms per category for the genes exhibiting changes in their protein translation rate. We look in greater detail at the category "immune system." Therein, two new sub-categories appear for the alternative analysis, namely "Fc receptor (B cell receptor) signaling" and "antigen processing" (Figure 4B, right). Therefore, both sets of enriched GO terms are similar with respect to their represented processes, but almost all processes are resolved in greater detail when considering differential single-cell protein expression dynamics instead of bulk protein abundances. Consequently, a more comprehensive picture of the changes in the cellular machinery arises.

In summary, we demonstrated in this example (1) the importance of considering differences in the cell division time for differential gene expression analysis and (2) that shifting the focus from mRNA and protein expression data alone to additionally estimating single-cell synthesis rates can greatly enhance the information output of population-based high-throughput gene expression data. We therefore encourage the estimation of translation or transcription rates either experimentally or as suggested in this study by combining population-based high-throughput datasets.

mRNA and protein dilution is to increase synthesis rates. We illustrate this in Figure 3C. Therein, we show the estimated change in protein synthesis rate constants for the sample proteins from Figure 3A assuming identical population average abundances for two cell systems with different cell division times. The $\log_2$-fold change of the increase in synthesis rates compensating protein loss by dilution in fast dividing compared to slowly dividing cells is again close to zero for the unstable MDM2 and highest for the stable RPS3. As shown in this example, equal population average abundances can conceal a strong change in protein synthesis.

To sum up, accounting for the non-uniform age distribution in a growing population allows determining the rate constants of gene expression from population-based high-throughput data. Cell division time and half-lives are important parameters that define this relationship. Only under certain conditions, observations of differential gene expression on the population level transfer directly to differential gene expression on the single-cell level. For an example of published data, we further showed that differential gene expression analyses based on protein synthesis rate constants can be more sensitive especially when comparing cell systems with different cell division times (Box 1).

DISCUSSION

The complex influence of the cell division time on gene expression is often overlooked while performing differential gene expression analysis based on high-throughput mRNA and

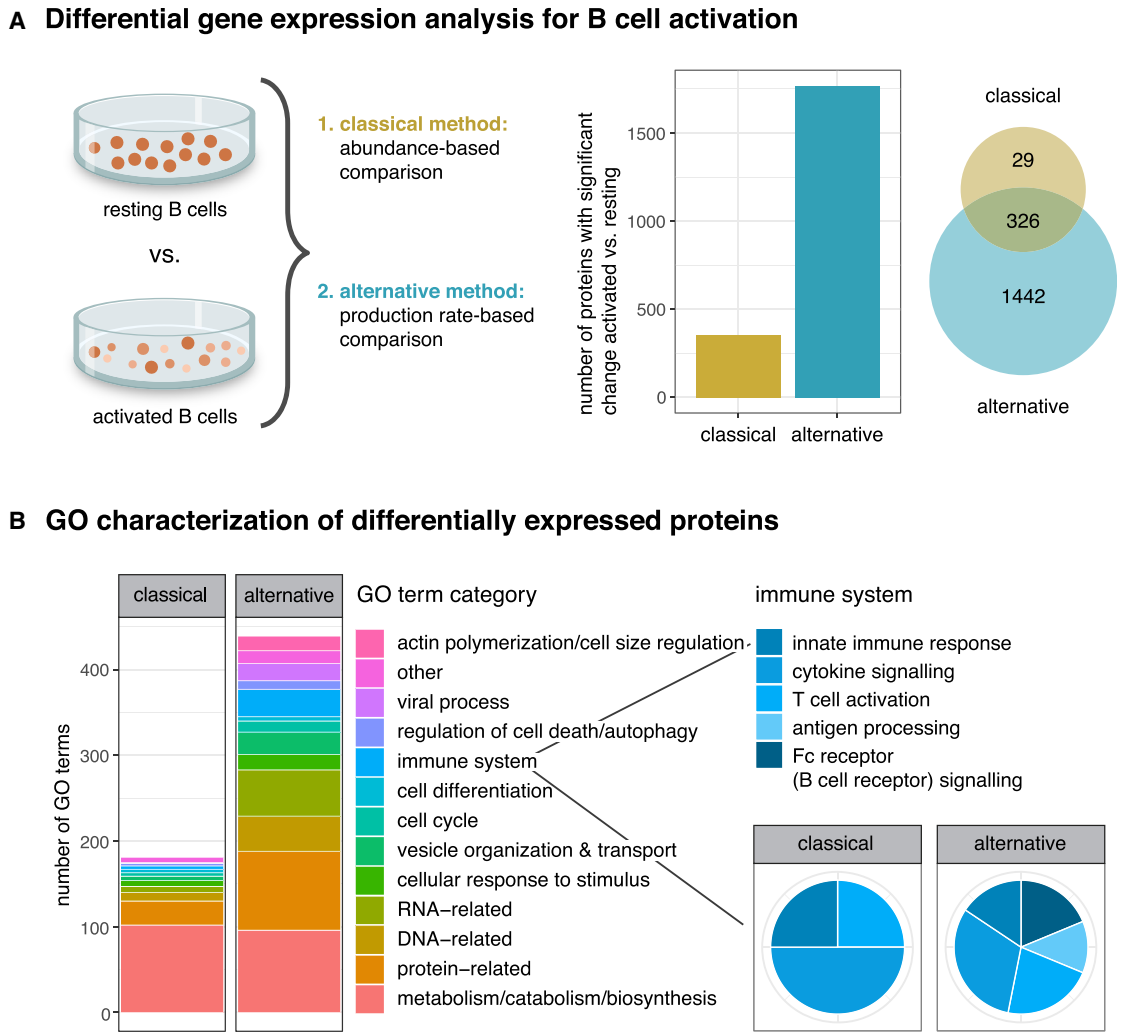


Figure 4. Differential Gene Expression in Resting versus Activated Human B Cells

(A) Comparison of the protein expression between resting and activated memory B cells, either via the classical approach based on comparing protein abundance measurements of cell cultures, or via the alternative approach of comparing protein synthesis rates of the cells in the population derived here. Abundances of 2,438 proteins were used (Rieckmann et al., 2017). We obtained 355 proteins with significantly altered abundance between the two states (Welch's test FDR < 0.05, fold change [FC] > 50%, yellow bar). For the alternative approach, protein translation rate constants were calculated for both resting and activated B cells using the transformed equation Figure 2D right (STAR Methods, Equation 17) and the following data: mRNA and protein expression (Rieckmann et al., 2017) and protein half-lives (Mathieson et al., 2018). 1,768 proteins reveal a significantly altered protein translation rate constant (Welch's test FDR < 0.05, FC > 50%, blue bar). The overlap between the sets of differentially expressed proteins detected by the classical or alternative approach are shown in a Venn diagram (right). Details are given in the STAR Methods.

(B) Categories of enriched GO terms among the proteins with significantly altered protein abundance or significantly altered protein translation rate constants as in (A) according to PANTHER analysis (Mi et al., 2017) and semi-automated clustering. All categories except for "actin polymerization and cell size regulation" occur in both protein sets, but almost all processes are resolved in greater detail when considering the alternative differential expression approach. The relative compositions (number of GO terms) of subcategories of the immune system-related terms are shown in pie charts. Details are given in the STAR Methods. See also Figure S4 and Tables S3 and S4.

protein measurements. The reason for this may be that cell division does not appear as a core process of gene expression (Figure 2A). By clearly distinguishing between intracellular processes of gene expression and population growth, we were able to demonstrate via mathematical modeling its influence on the link between single-cell level and population-level mRNA and protein expression (Figure 2D). The cell division time (1) sets the time frame for the single-cell gene expression dynamics and (2) is the parameter that determines the inhomogeneous age distribution in a growing cell population. Therefore, the cell division time is an additional parameter complementing the single-cell kinetic parameters of gene expression and linking them to the population average mRNA and protein molecule counts. Our stringently derived, easy-to-use mathematical expressions capture this relationship in exponentially growing cell populations.

To derive the presented formulas, assumptions have been made, which are listed in the STAR Methods. For example, we

assumed a constant mRNA transcription over the course of the cell cycle, which is supported by recent hints on mechanisms of dosage compensation in mammalian cells (Padovan-Merhar et al., 2015; Skinner et al., 2016) and has been used to model transcription before (Miller et al., 2011; Schwanhäusser et al., 2011). Another important assumption is that the model does not take into account any extracellular or intracellular feedback or feed-forward regulation on the processes involved (Alon, 2006; Vogel and Marcotte, 2012; Braun and Young, 2014).

Because of the assumption of condition-specific, fixed kinetic parameters of gene expression, transient changes in the state of a population cannot be described with this approach. In turn, our approach is very suitable for using population-based methods to study gene expression of unperturbed systems or cellular processes that result in a new stable state, e.g., comparing undifferentiated and the terminal differentiated state of a differentiation process, the immune system in its activated compared to its resting state, or the unperturbed versus perturbed state of a perturbation screen.

As an example for differential gene expression analysis between cell types that have very different cell division times, we compared gene expression between resting and activated states of B cells using publicly available data of human B cells (Rieckmann et al., 2017; Mathieson et al., 2018) (Box1, Figure 4). Because of the strong differences in cell division, fold changes of bulk protein measurements do not directly reveal all changes in gene expression characteristics in the underlying single cells. We presented an alternative method to perform differential expression analysis based on changes in rates of gene expression instead of relying on changes in protein abundances. Comparing rates of gene expression, e.g., synthesis rates, takes the effect of the cell division time into account. We show that this approach can increase the information output of differential gene expression analysis. We observe a strong increase in the number of proteins with changed synthesis rates, by almost 5-fold, compared to analyzing protein abundance changes only.

Among the additional processes revealed by differential gene expression analysis based on synthesis rates are some that are highly relevant for biomedical and pharmaceutical research. We find that processes that are strongly represented only in our analysis of altered single-cell protein synthesis rates, but not in that of altered protein abundances, have been particularly highlighted as source for potential drug targets in diffuse large B cell lymphoma, namely DNA repair and B cell receptor signaling, e.g., Fc receptor signaling and antigen processing (Derenzini et al., 2015; De Jong et al., 2018) (Figure 4B, Tables S3 and S4). In addition, the endoplasmic reticulum protein transport has been found to be of relevance in B cell malignancies in general (Carew et al., 2006). This is a facet of protein transport that we detected more frequently related to proteins with altered synthesis rates than related to proteins with altered abundances. Subsequently, in this example, focusing on changes in synthesis rates gives a more complete, biologically relevant picture of the cellular machinery that is involved in B cell activation.

We therefore aim to encourage quantitative estimation of rates of gene expression. Experimentally determining synthesis rates of mRNA and protein is still challenging despite strong advances in techniques enabling direct measurements (Larson et al., 2011; Ingolia et al., 2012; Brar and Weissman, 2015; Yan et al., 2016; Calviello and Ohler, 2017). Thus, an estimation using population average mRNA and protein levels, half-lives and the cell division time (Figure 2D and STAR Methods, Equations 15, 16, and 17) may be favorable. For determining synthesis rates in NIH3T3 cells, our approach is thereby overall consistent with our previously published approach (Schwanhäusser et al., 2013) (Figure S4B). Many datasets are already available online reporting one or several of the required quantities on genomic scale for many species (e.g., Friedel et al., 2009; Boisvert et al., 2012; Tani et al., 2012; Rieckmann et al., 2017; Mathieson et al., 2018 and see Liu et al., 2016 for a review on measurement techniques). However, combining multiple measurements can propagate potential measurement errors (Figure S3C).

To summarize, changes in population average mRNA and protein molecule counts per cell allow only limited conclusions on changes in the underlying single-cell gene expression, especially when comparing cell systems which differ strongly in their cell division time. Moreover, the influence of cell division on average population mRNA and protein molecule counts differs for every mRNA and every protein depending on their half-lives. In order to take this gene-specific effect of cell division into account, we therefore recommend to base differential gene expression analysis on parameters of rates of gene expression.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- METHOD DETAILS
  - Derivation of the mRNA Population Average
  - Derivation of the Protein Population Average
  - The Single-Cell Steady State
  - Quantifying mRNA and Protein Synthesis Rates
  - Assumptions Underlying the Derived Formulas
  - The Age Distribution within the Population
  - Populations of Non-identical Cells
  - Measurement Error Effects on Synthesis Rates
  - The Importance of Incorporating Cell Division
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Resting vs. Activated B Cells Data
  - Classical: Differential Protein Expression
  - Alternative: Differential Protein Synthesis
  - GO Enrichment and GO Term Clustering

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2019.07.009.

ACKNOWLEDGMENTS

We thank Matthias Selbach, Max Delbrück Center for Molecular Medicine Berlin, and Gunnar Dittmar, Luxembourg Institute of Health, for valuable discussions. J.W. acknowledges funding by the Helmholtz Association (Personalized Medicine Initiative "iMED") and the German Federal Ministry of Education and Research (BMBF) within the e:Med research and funding (FKZ 01ZX1607F). J.S. was supported by the German Federal Ministry of Education and Research (BMBF) *E:kid*(E:Med) project (FKZ 01ZX1612D),

Please cite this article in press as: Baum et al., Of Gene Expression and Cell Division Time: A Mathematical Framework for Advanced Differential Gene Expression and Data Analysis, Cell Systems (2019), https://doi.org/10.1016/j.cels.2019.07.009

ML-MED project (FKZ 01IS180044C), and by EU-FP7, HEALTH-F4-2013-602156, *HeCaToS* project.

AUTHOR CONTRIBUTIONS

Conceptualization, D.B.; K.B.; J.W.; and J.S.; Methodology, K.B.; J.S.; D.B.; and J.W.; Software, K.B.; Formal Analysis, K.B. and J.S.; Writing – Original Draft, D.B. and K.B.; Writing – Review & Editing, K.B.; D.B.; and J.W.; Visualization, K.B., J.W., and D.B.; Supervision, D.B. and J.W.; Project Administration K.B., D.B., and J.W.; Funding Acquisition: J.W. and J.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 5, 2018
Revised: April 15, 2019
Accepted: July 23, 2019
Published: September 11, 2019

REFERENCES

Aebersold, R., and Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. Nature *537*, 347–355.

Alon, U. (2006). An introduction to systems biology: design principles of biological circuits (Chapman & Hall/CRC).

Becker, K.A., Ghule, P.N., Therrien, J.A., Lian, J.B., Stein, J.L., Van Wijnen, A.J., and Stein, G.S. (2006). Self-renewal of human embryonic stem cells is supported by a shortened G1 cell cycle phase. J. Cell. Physiol. *209*, 883–893.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. J. R. Stat. Soc. B *57*, 289–300.

Boisvert, F.M., Ahmad, Y., Gierliński, M., Charrière, F., Lamont, D., Scott, M., Barton, G., and Lamond, A.I. (2012). A quantitative spatial proteomics analysis of proteome turnover in human cells. Mol. Cell. Proteomics *11*, M111.011429.

Brar, G.A., and Weissman, J.S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. Nat. Rev. Mol. Cell Biol. *16*, 651–664.

Braun, K.A., and Young, E.T. (2014). Coupling mRNA synthesis and decay. Mol. Cell. Biol. *34*, 4078–4087.

Calviello, L., and Ohler, U. (2017). Beyond Read-Counts: ribo-seq Data Analysis to Understand the Functions of the transcriptome. Trends Genet. *33*, 728–744.

Cambridge, S.B., Gnad, F., Nguyen, C., Bermejo, J.L., Krüger, M., and Mann, M. (2011). Systems-wide proteomic analysis in mammalian cells reveals conserved, functional protein turnover. J. Proteome Res. *10*, 5275–5284.

Carew, J.S., Nawrocki, S.T., Krupnik, Y.V., Dunner, K., Jr., McConkey, D.J., Keating, M.J., and Huang, P. (2006). Targeting endoplasmic reticulum protein transport: a novel strategy to kill malignant B cells and overcome fludarabine resistance in CLL. Blood *107*, 222–231.

Carlson, M. (2016). org.Hs.eg.db: Genome wide annotation for Human. https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html.

Cooperman, J., Neely, R., Teachey, D.T., Grupp, S., and Choi, J.K. (2004). Cell division rates of primary human precursor B cells in culture reflect in vivo rates. Stem Cells *22*, 1111–1120.

De Jong, M.R.W., Visser, L., Huls, G., Diepstra, A., Van Vugt, M., Ammatuna, E., Van Rijn, R.S., Vellenga, E., Van Den Berg, A., Fehrmann, R.S.N., et al. (2018). Identification of relevant drugable targets in diffuse large B-cell lymphoma using a genome-wide unbiased CD20 guilt-by association approach. PLoS One *13*, e0193098.

Derenzini, E., Agostinelli, C., Imbrogno, E., Iacobucci, I., Casadei, B., Brighenti, E., Righi, S., Fuligni, F., Ghelli Luserna Di Rorà, A., Ferrari, A., et al. (2015). Constitutive activation of the DNA damage response pathway as a novel therapeutic target in diffuse large B-cell lymphoma. Oncotarget *6*, 6553–6569.

Eden, E., Geva-Zatorsky, N., Issaeva, I., Cohen, A., Dekel, E., Danon, T., Cohen, L., Mayo, A., and Alon, U. (2011). Proteome half-life dynamics in living human cells. Science *331*, 764–768.

Friedel, C.C., Dölken, L., Ruzsics, Z., Koszinowski, U.H., and Zimmer, R. (2009). Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. Nucleic Acids Res. *37*, e115.

Fröhlich, H., Speer, N., Poustka, A., and Beissbarth, T. (2007). GOSim - an R-package for computation of information theoretic GO similarities between terms and gene products. BMC Bioinformatics *8*, 166.

Hargrove, J.L., and Schmidt, F.H. (1989). The role of mRNA and protein stability in gene expression. FASEB J. *3*, 2360–2370.

Ingolia, N.T., Brar, G.A., Rouskin, S., Mcgeachy, A.M., and Weissman, J.S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc. *7*, 1534–1550.

Janzen, W.P. (2014). Screening technologies for small molecule discovery: the state of the art. Chem. Biol. *21*, 1162–1170.

Jones, D.D., Wilmore, J.R., and Allman, D. (2015). Cellular dynamics of memory B cell populations: IgM+ and IgG+ memory B cells persist indefinitely as quiescent cells. J. Immunol. *195*, 4753–4759.

Larson, D.R., Zenklusen, D., Wu, B., Chao, J.A., and Singer, R.H. (2011). Real-time observation of transcription initiation and elongation on an endogenous yeast gene. Science *332*, 475–478.

Legewie, S., Herzel, H., Westerhoff, H.V., and Blüthgen, N. (2008). Recurrent design patterns in the feedback regulation of the mammalian signalling network. Mol. Syst. Biol. *4*, 190.

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. Cell *165*, 535–550.

Llamosi, A., Gonzalez-Vargas, A.M., Versari, C., Cinquemani, E., Ferrari-Trecate, G., Hersen, P., and Batt, G. (2016). What population reveals about individual cell identity: single-cell parameter estimation of models of gene expression in yeast. PLoS Comput. Biol. *12*, e1004706.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. PLoS Comput. Biol. *13*, e1005457.

Macarron, R., Banks, M.N., Bojanic, D., Burns, D.J., Cirovic, D.A., Garyantes, T., Green, D.V., Hertzberg, R.P., Janzen, W.P., Paslay, J.W., et al. (2011). Impact of high-throughput screening in biomedical research. Nat. Rev. Drug Discov. *10*, 188–195.

Mathieson, T., Franken, H., Kosinski, J., Kurzawa, N., Zinn, N., Sweetman, G., Poeckel, D., Ratnu, V.S., Schramm, M., Becher, I., et al. (2018). Systematic analysis of protein turnover in primary cells. Nat. Commun. *9*, 689.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P.D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. *45*, D183–D189.

Miller, C., Schwalb, B., Maier, K., Schulz, D., Dümcke, S., Zacher, B., Mayer, A., Sydow, J., Marcinowski, L., Dölken, L., et al. (2011). Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. Mol. Syst. Biol. *7*, 458.

Milo, R., Jorgensen, P., Moran, U., Weber, G., and Springer, M. (2010). BioNumbers-the database of key numbers in molecular and cell biology. Nucleic Acids Res. *38*, D750–D753.

Padovan-Merhar, O., Nair, G.P., Biaesch, A.G., Mayer, A., Scarfone, S., Foley, S.W., Wu, A.R., Churchman, L.S., Singh, A., and Raj, A. (2015). Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. Mol. Cell *58*, 339–352.

Powell, E.O. (1956). Growth rate and generation time of bacteria, with special reference to continuous culture. J. Gen. Microbiol. *15*, 492–511.

R Core Team. (2016). R: A language and environment for statistical computing.

Rieckmann, J.C., Geiger, R., Hornburg, D., Wolf, T., Kveler, K., Jarrossay, D., Sallusto, F., Shen-Orr, S.S., Lanzavecchia, A., Mann, M., et al. (2017). Social network architecture of human immune cells unveiled by quantitative proteomics. Nat. Immunol. *18*, 583–593.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2013). Corrigendum: global quantification of mammalian gene expression control. Nature *495*, 126–127.

Skinner, S.O., Xu, H., Nagarkar-Jaiswal, S., Freire, P.R., Zwaka, T.P., and Golding, I. (2016). Single-cell analysis of transcription kinetics across the cell cycle. eLife *5*, e12175.

Tani, H., Mizutani, R., Salam, K.A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y., and Akimitsu, N. (2012). Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res. *22*, 947–956.

Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. *13*, 227–232.

Wickham, H. (2016). ggplot2: elegant graphics for data analysis (Springer-Verlag New York).

Wickham, H., Francois, R., Henry, L., and Müller, K. (2018). Dplyr: A Grammar of Data Manipulation.

Yan, X., Hoek, T.A., Vale, R.D., and Tanenbaum, M.E. (2016). Dynamics of translation of single mRNA molecules in vivo. Cell *165*, 976–989.

Yu, G.C., Wang, L.G., Han, Y.Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes Among gene clusters. OMICS A. J. Integr. Biol. *16*, 284–287.

Please cite this article in press as: Baum et al., Of Gene Expression and Cell Division Time: A Mathematical Framework for Advanced Differential Gene Expression and Data Analysis, Cell Systems (2019), https://doi.org/10.1016/j.cels.2019.07.009

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
R version 3.3.2	R Core Team, 2016	RRID:SCR_001905
R package dplyr	Wickham et al., 2018	RRID:SCR_016708
R package ggplot2	Wickham, 2016	RRID:SCR_014601
PANTHER	Mi et al., 2017	RRID:SCR_004869
R package org.Hs.eg.db	Carlson, 2016	https://www.bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html
R package clusterProfiler	Yu et al., 2012	RRID:SCR_016884
R package GOSim	Fröhlich et al., 2007	http://bioconductor.org/packages/release/bioc/html/GOSim.html
Other		
protein expression data in resting and activated human B cells	Rieckmann et al., 2017	Table S3
mRNA expression data in resting and activated human B cells	Rieckmann et al., 2017	Table S8
protein half-lives in resting B cells	Mathieson et al., 2018	Data S2

LEAD CONTACT AND MATERIALS AVAILABILITY

Further requests for information and resources should be directed to and will be fulfilled by the Lead Contact, Jana Wolf (jana.wolf@mdc-berlin.de).

METHOD DETAILS

Unless stated otherwise, abundance refers to molecule count and the natural logarithm is used. $\mathcal{N}(a,b)$ denotes the normal distribution with mean a and standard deviation b. All numerical computations and data analyses were performed in R, version 3.3.2 (R Core Team, 2016) using the package *dplyr* (Wickham et al., 2018). Figures were created using the package *ggplot2* (Wickham, 2016).

Derivation of the mRNA Population Average

mRNA Abundance in Single Cells

The molecule count of a specific mRNA, r, in a single mammalian cell is described by the differential equation:

$$\frac{dr}{da} = v_{sr} - k_{dr} \cdot r, \qquad \text{(Equation 1)}$$

which assumes that r is constantly produced over time a with the rate v_{sr} and degraded proportionally to the number of mRNA molecules, $r(a)$, with the rate constant k_{dr}. The parameter of the transcription rate, v_{sr}, summarizes all processes from transcription initiation to mRNA processing in the cytoplasm. The degradation rate constant, k_{dr}, summarizes all processes of decay of an mRNA molecule. The solution to Equation 1 can be obtained analytically:

$$r(a) = \frac{v_{sr}}{k_{dr}} - \left(\frac{v_{sr}}{k_{dr}} - r(0)\right) \cdot e^{-k_{dr} \cdot a}, \qquad \text{(Equation 2)}$$

where $r(0)$ is the initial mRNA abundance.

mRNA in a Dividing Single Cell

We assume that the mRNA doubles during one cell cycle with length τ in order to compensate for the mRNA loss by cell division, that is $2 \cdot r(0) = r(\tau)$ This condition allows to determine $r(0)$ in Equation 2 and replace it by an expression dependent on τ:

$$r(0) = \frac{v_{sr}}{k_{dr}} \cdot \frac{1 - e^{-k_{dr} \cdot \tau}}{2 - e^{-k_{dr} \cdot \tau}}. \qquad \text{(Equation 3)}$$

The progression time a can thus be interpreted as the age of the cell, i.e. the time since its last cell division. Subsequently, the following expression gives the mRNA abundance of a single cell within one division cycle for the age $a \in [0, \tau]$ of the cell:

$$r(a) = \frac{v_{sr}}{k_{dr}} \cdot \left(1 - \frac{e^{-k_{dr} \cdot a}}{2 - e^{-k_{dr} \cdot \tau}}\right). \quad \text{(Equation 4)}$$

From Dividing Single Cells to a Population

We consider a population of identical cells, i.e. we assume that the kinetic parameters v_{sr}, k_{sp}, k_{dr} and k_{dp} are identical in all cells. However, we assume that the cell cycle occurs asynchronously, therefore, the population is a mixture of cells of different age. Under certain assumptions, in particular (i) the same division time τ for each cell, (ii) cells have grown sufficiently long such that the age distribution has become steady, and (iii) statistical fluctuations are negligible, the age distribution of a steadily growing population can be described by a probability density function (Powell, 1956):

$$\phi(a) = 2 \cdot \frac{\log(2)}{\tau} \cdot 2^{-\frac{a}{\tau}}. \quad \text{(Equation 5)}$$

This distribution defines the percentage of cells ($\phi(a)da$) in each (arbitrarily small) interval of ages, $[a, a + da]$. One cell of age a contributes with $r(a)$ to the total mRNA abundance of the population. The average mRNA abundance over all cells in the population, $\overline{R}$, is determined by summing the mRNA contents over all individual cells, i.e. over all ages, weighted by the probability of the age to occur within the population. This corresponds to calculating the expected value of the mRNA abundance $r(a)$ as continuous function of the random variable age a following the probability density function ϕ (see also law of the unconscious statistician):

$$\overline{R} = E[r(a)] = \int_{-\infty}^{\infty} \phi(a) \cdot r(a) da = \int_{0}^{\tau} \phi(a) \cdot r(a) da. \quad \text{(Equation 6)}$$

The integration borders can thereby be reduced to the interval $[0,\tau]$ because the probability density function ϕ takes non-zero values only therein. Please note that we define $\overline{R}$ as an average over different cellular mRNA contents. Here, these differences are only due to the non-synchronized cell cycle; they do not arise because of general cell-to-cell heterogeneity in abundances.

The expressions for $\phi(a)$ (Equation 5) and $r(a)$ (Equation 4) are used in Equation 6:

$$\overline{R} = \int_{0}^{\tau} 2 \cdot \frac{\log(2)}{\tau} \cdot 2^{-\frac{a}{\tau}} \cdot r(a) da = 2 \cdot \frac{\log(2)}{\tau} \int_{0}^{\tau} e^{-\frac{\log(2) \cdot a}{\tau}} \cdot \frac{v_{sr}}{k_{dr}} \cdot \left(1 - \frac{e^{-k_{dr} \cdot a}}{2 - e^{-k_{dr} \cdot \tau}}\right) da.$$

Using that the antiderivative of the exponential function $e^{-b \cdot x}$ in x is $-1/b \cdot e^{-b \cdot x}$, this integral can be straightforwardly computed giving the population average mRNA molecule count per cell as:

$$\overline{R} = \frac{v_{sr}}{k_{dr} + \log(2)/\tau}. \quad \text{(Equation 7)}$$

Derivation of the Protein Population Average

Protein Abundance in Single Cells

The molecule count of a protein in a single cell, p, depends on the molecule count of its mRNA, r, and is described by the system of differential equations:

$$\frac{dr}{da} = v_{sr} - k_{dr} \cdot r \quad \text{(Equation 8)}$$

$$\frac{dp}{da} = k_{sp} \cdot r - k_{dp} \cdot p, \quad \text{(Equation 9)}$$

with the translation rate constant k_{sp} and the degradation rate constant of the protein k_{dp}. The translation rate constant summarizes all processes of the synthesis of a protein molecule from its mRNA. The degradation rate constant summarizes all processes of decay of a protein molecule. For explanation of the parameters governing the equation for r see section *mRNA Abundance in Single Cells*. The solution for r is independent of p (compare Equation 2). The solution for p depends on r and is given by the analytical expression

$$p(a) = \frac{v_{sr} \cdot k_{sp}}{k_{dr} \cdot k_{dp}} + \left(p(0) - \frac{k_{dr} \cdot (v_{sr} - k_{dp} \cdot r(0))}{k_{dp} \cdot (k_{dr} - k_{dp})}\right) e^{-k_{dp} \cdot a} + \frac{k_{dr} \cdot (v_{sr} - k_{dr} \cdot r(0))}{k_{dr} \cdot (k_{dr} - k_{dp})} e^{-k_{dr} \cdot a} \quad \text{(Equation 10)}$$

for $k_{dr} \neq k_{dp}$ and for the time point a. The initial protein amount is $p(0)$, the initial mRNA abundance is $r(0)$. Note that the solution p is different in case $k_{dr} = k_{dp}$. In the following, we will only show the results for $k_{dr} \neq k_{dp}$, however, exactly the same approach gives the same expression for the average population protein abundance (Equation 13) in case $k_{dr} = k_{dp}$.

Please cite this article in press as: Baum et al., Of Gene Expression and Cell Division Time: A Mathematical Framework for Advanced Differential Gene Expression and Data Analysis, Cell Systems (2019), https://doi.org/10.1016/j.cels.2019.07.009

Protein Abundance in a Dividing Single Cell

Assuming a cell division after time τ, the mRNA and protein abundances should exactly double within this time frame, i.e. $r(\tau) = 2 \cdot r(0)$ and $p(\tau) = 2 \cdot p(0)$. Under these conditions, $r(0)$ and $p(0)$ can be determined as a function of τ (see Equation 3 for the initial mRNA abundance $r(0)$, which is also used to obtain the dependence of $p(0)$ on τ):

$$p(0) = \frac{v_{sr} \cdot k_{sp} \cdot (2 - e^{-k_{dp} \cdot \tau})^{-1}}{(k_{dr} - k_{dp}) \cdot k_{dr} \cdot k_{dp}} \left(k_{dr} - k_{dp} + \left(k_{dp} \cdot \frac{1 - e^{-k_{dr} \cdot \tau}}{2 - e^{-k_{dr} \cdot \tau}} - k_{dr} \right) \cdot e^{-k_{dp} \cdot \tau} + \frac{k_{dp} \cdot e^{-k_{dr} \cdot \tau}}{2 - e^{-k_{dr} \cdot \tau}} \right) \quad \text{(Equation 11)}$$

By inserting Equation 3 and Equation 11 in the expression for the protein abundance, Equation 10, we obtain

$$p(a) = \frac{v_{sr} \cdot k_{sp}}{k_{dr} \cdot k_{dp}} - \frac{v_{sr} \cdot k_{sp} \cdot e^{-k_{dp} \cdot a}}{k_{dp} \cdot (k_{dr} - k_{dp}) \cdot (2 - e^{-k_{dp} \cdot \tau})} + \frac{v_{sr} \cdot k_{sp} \cdot e^{-k_{dr} \cdot a}}{k_{dr} \cdot (k_{dr} - k_{dp}) \cdot (2 - e^{-k_{dr} \cdot \tau})} \quad \text{(Equation 12)}$$

for the protein abundance of a dividing single cell with age, i.e., time after cell division, $a \in [0, \tau]$.

From Dividing Single Cells to a Population

Again, using the density function of the population age distribution $\phi(a)$, and assuming that the kinetic parameters and cell division time are the same for each cell of the population, the average protein molecule count per cell in a population, $\overline{P}$, can be calculated. The integral gives the expected value of the protein abundance function as a function of the age distribution following the probability density function ϕ (compare to the corresponding section for the mRNA abundance):

$$\overline{P} = \int_0^{\tau} \phi(a) \cdot p(a) da.$$

Inserting the expression for $\phi(a)$ from Equation 5 and for $p(a)$ from Equation 12, and using that the antiderivative of $e^{-b \cdot x}$ as function of x is $-1/b \cdot e^{-b \cdot x}$, we can simplify the integral in a lengthy but straightforward calculation to obtain the average protein molecule count per cell, $\overline{P}$, in a population of dividing cells with age distribution given by ϕ:

$$\overline{P} = \frac{v_{sr} \cdot k_{sp}}{(k_{dr} + \log(2)/\tau) \cdot (k_{dp} + \log(2)/\tau)}. \quad \text{(Equation 13)}$$

The Single-Cell Steady State

The employed differential equation system of the single cell gene expression has the following steady state solution, i.e. the solution for $dr/da = dp/da = 0$:

$$r_{ss} = \frac{v_{sr}}{k_{dr}} \quad \text{and} \quad p_{ss} = \frac{v_{sr} \cdot k_{sp}}{k_{dr} \cdot k_{dp}} \quad \text{(Equation 14)}$$

for the mRNA and protein abundance, respectively.

Quantifying mRNA and Protein Synthesis Rates

The derived formulas (Equations 7 and 13) can in particular be used to determine single-cell average transcription rates and translation rate constants. Rearranging Equation 7 (equation in Figure 2D left) yields the transcription rate v_{sr}:

$$v_{sr} = \overline{R} \cdot (\log(2)/\tau + k_{dr}) \quad \text{(Equation 15)}$$

Similarly, rearranging Equation 13 (equation in Figure 2D right) yields the translation rate constant k_{sp}:

$$k_{sp} = \overline{P} \cdot \frac{(\log(2)/\tau + k_{dr})}{v_{sr}} \cdot (\log(2)/\tau + k_{dp}) \quad \text{(Equation 16)}$$

Alternatively, using Equation 15 to replace the transcription rate yields the dependency of the translation rate constant on the mRNA abundance $\overline{R}$ instead of transcription rate and mRNA degradation:

$$k_{sp} = \frac{\overline{P}}{\overline{R}} \cdot (\log(2)/\tau + k_{dp}) \quad \text{(Equation 17)}$$

In all of the above formulas, τ denotes the cell division time, $\overline{R}$ and $\overline{P}$ the population average mRNA and protein abundance, respectively, in the unit molecules/cell. The degradation rate constants k_{dr} and k_{dp} can be replaced by the molecular half-lives of mRNA and protein which are denoted by hl_r and hl_p, respectively. They relate to the degradation rate constants by $k_{dr} = \log(2)/hl_r$ and $k_{dp} = \log(2)/hl_p$. To apply the formulas, these values need to be set for each mRNA-protein pair, for specific cell types and/or conditions.

Assumptions Underlying the Derived Formulas

The assumptions underlying the derivation of the expressions linking gene expression kinetic parameters to average mRNA and protein abundances, $\overline{R}$ and $\overline{P}$, are listed in the following.

To be able to describe gene expression with ordinary differential equations:

- All processes must be spatially continuous events
- The numbers of molecules participating must be sufficiently large
- Cells are considered to produce mRNA and protein continuously from cell birth to cell division, i.e. cell cycle steps are neglected in that core model.
- We neglect any extracellular or intracellular feedback or feed-forward regulation on the processes involved, i.e. we assume that regulations are steady and settled, and, in particular, that the parameters of gene expression are constant for a certain state or condition of the population. This simplification is frequently used, also for modeling gene expression in single cells (Alon, 2006; Legewie et al., 2008; Llamosi et al., 2016).
- We assume a dosage compensation (Padovan-Merhar et al., 2015; Skinner et al., 2016). The resulting constant rate of transcription over the cell cycle, i.e. the production of a constant number of mRNA molecules in a certain time interval, is one way of representing mRNA synthesis which has been used by us and others before (Miller et al., 2011; Schwanhäusser et al., 2011; Skinner et al., 2016).
- By formulating the model in molecule counts per cell, degradation encompasses only molecular degradation. Therefore, the degradation rate constants are assumed to be constant over the cell cycle, e.g. (Eden et al., 2011; Llamosi et al., 2016).
- mRNA and protein abundances have doubled from cell birth to division time τ, i.e. $r(\tau) = 2 \cdot r(0)$ and $p(\tau) = 2 \cdot p(0)$.

For the derivation of the age distribution $\phi(a)$:

- Homogeneous population (i.e. only organisms of one type).
- Sufficiently large population to be able to consider changes as continuous and statistical fluctuations as negligible (this is similar to assumptions made when employing ODE models).
- Cells have grown sufficiently long such that the age distribution is stable but saturation does not affect growth.
- The cell division time τ is identical for all cells in a population (exceptions are considered in Figure S2 and Figure S3).

In order to summarize the mRNA and protein content over all cells to calculate $\overline{R}$ and $\overline{P}$:

- For a considered mRNA-protein pair, all cells have exactly the same kinetic parameter values for production and degradation rate constants (exceptions are considered in Figure S3).

The Age Distribution within the Population

Development of the Age Distribution

We simulated the development of the age distribution in an exponentially growing population of cells (Figure S2C). We started with a population of 10^6 cells that is initially synchronized, i.e. all cells have age zero. We assumed a variation in the cell division time $\tau = 27.5$ h (NIH3T3 cells, (Schwanhäusser et al., 2011)) and sampled it for each cell i from a normal distribution with standard deviation of 15%, $\tau_i \sim 27.5\ \text{h} \cdot \mathcal{N}(1, 0.15)$.

We updated the age of each cell three times per generation (at elapsed times modulo τ of 0 h, 10 h, 20 h). At each updating step, cells that have an age larger than their cell division time give rise to two newborn cells which are assigned new, random cell division times. After each update, 10^6 cells are randomly drawn from the population and form the new population. We followed the population over 25 generations.

The emerging age distributions are compared to a close-to-steady age distribution derived from Powell's age distribution. This age distribution is obtained if the age a_i for each cell i with a cell division time τ_i is sampled from Powell's steady age distribution according to the density function

$$\phi(a_i) = 2 \cdot \frac{\log(2)}{\tau_i} \cdot 2^{-\frac{a}{\tau_i}} \qquad \text{(Equation 18)}$$

(see also Figure S2B). We performed the Kolmogorov-Smirnov-test for sub-populations of 10^4 cells at each updating step, and used FDR correction for multiple testing (Benjamini and Hochberg, 1995). The resulting corrected p-values are used as measure of similarity between the age distributions (Figure S2D).

Variable Cell Division within the Population

For the derivation of our formulas, we assumed identical cells with a fixed cell division time τ. This gives rise to the age distribution derived by (Powell, 1956). Moreover, we investigated the case that the cell division time of the cells within a population, $\widehat{\tau}$, follows a normal distribution with average 27.5 h (as for NIH3T3 cells) and a fixed standard deviation α, $\widehat{\tau} \sim 27.5 \cdot max(0.01, \mathcal{N}(1, \alpha))$ (Figure S2E). We further assumed that the age distribution is quasi-steady, meaning that for cells with a certain cell division time the respective steady age distribution as given by Powell is reached (see Equation 18). The effect of variation in τ and the resulting variation in the age distribution on the connection between kinetic parameters and population average abundances is characterized by the relative deviations $(\overline{R} - \widehat{R})/\overline{R}$ and $(\overline{P} - \widehat{P})/\overline{P}$ of the mRNA and protein population average abundances simulated in the population of cells with variation in τ, $\widehat{R}$ and $\widehat{P}$, from the mRNA and protein abundance calculated without variation in τ according to our proposed formula, $\overline{R}$ and $\overline{P}$.

Please cite this article in press as: Baum et al., Of Gene Expression and Cell Division Time: A Mathematical Framework for Advanced Differential Gene Expression and Data Analysis, Cell Systems (2019), https://doi.org/10.1016/j.cels.2019.07.009

Populations of Non-identical Cells

For the derivation of our formulas (Equations 7 and 13), we assumed a population of identical cells (except for their age) with fixed cell division time, synthesis and degradation rate constants, and doubled abundances from cell birth to division. We challenged these assumptions and investigated whether the relationship between population average abundances and single cell kinetic parameters and cell division time remains similar if assuming different (but fixed, within a cell's life) parameter values around a common population average in different cells within the population (Figures S3A and B).

We assumed that the cell division time τ, transcription rate v_{sr}, translation rate constant k_{sp}, mRNA and protein degradation rate constants k_{dr} and k_{dp}, and the initial mRNA and protein abundances (abundance at cell birth) are subject to additive, Gaussian noise: $\widehat{\beta} = \beta \cdot max(0.001, \mathcal{N}(1, \epsilon))$, with the standard deviation ϵ being 0.3 (0.1 for left panels in Figure S3B, 0.15 for τ).

For a given mRNA-protein pair (Table S1), we sampled populations with a fixed population size (100 to 10^6 single cells, as indicated). In detail, for each cell we

1. draw a random cell division time around τ,
2. draw an age from the resulting age distribution according to Equation 18,
3. draw a random value for each of the four kinetic gene expression parameters around their original value (Table S1),
4. calculate the initial mRNA abundance and initial protein abundance from its sampled cell division time and sampled parameter values,
5. draw a random value around the calculated initial mRNA abundance,
6. draw a random value around the calculated initial protein abundance,
7. calculate the mRNA abundance and protein abundance from the sampled age, kinetic parameter values, initial abundances.

For each thus sampled population of cells, we computed the population average mRNA abundance and the population average protein abundance as arithmetic mean over the values for all cells in the population.

We performed this sampling and computation for 200 populations for each case. The average value of the obtained distributions of populations averages is shifted with respect to the population averages computed for identical cells depending on the employed mRNA-protein pair (and the amount of allowed variation ϵ). We report and interpret this shift for 200 populations of 10^6 cells as a sensitivity measure toward intra-population variability (Table S2). In case of a small shift, our framework derived for populations of identical cells can be considered as a good approximation also for populations of cells with varying parameters.

Measurement Error Effects on Synthesis Rates

Our derived formulas can be used to compute the mRNA transcription rate, v_{sr}, or the protein translation rate constant, k_{sp} (see Eqn Equation 15, Equation 16 and Equation 17, respectively). We now assumed that the employed quantities for computation, k_{dr}, R, τ, P, k_{dp}, can be subject to measurement errors with a standard deviation of 30%, i.e. each quantity $\widehat{\beta}$ follows a log-normal distribution around its original value β, $\widehat{\beta} \sim e^{\mathcal{N}(0,0.3)} \cdot \beta$ (Figure S3C top). These measurement errors will propagate to the synthesis rates estimated from the erroneous measurements, $\widehat{v}_{sr}$ and $\widehat{k}_{sp}$, which will deviate from the synthesis rates derived without error, v_{sr} and k_{sp}. The dispersions of the distributions of these relative deviations, $(v_{sr} - \widehat{v}_{sr})/v_{sr}$ and $(k_{sp} - \widehat{k}_{sp})/k_{sp}$, illustrate the strength of the error propagation and potentiation (Figure S3C bottom).

The Importance of Incorporating Cell Division

We assessed the importance of cell division from three perspectives (Figure 3):

(i) The cell division time affects the relationship between single-cell rates of gene expression and the population average abundances. If we consider a population of cells with certain rates of gene expression, v_{sr}, k_{sp}, k_{dr}, k_{dp}, the observed population averages depend on how fast these cells divide. In particular, the population averages can be very different whether we consider cell division or whether these cells are in steady state ($\tau = \infty$). We quantify their difference by the ratio $\overline{R_{ss}}/\overline{R_\tau}$ or $\overline{P_{ss}}/\overline{P_\tau}$ between the mRNA or protein population averages obtained for a given cell division time τ, $\overline{R_\tau}$ or $\overline{P_\tau}$, and the population averages obtained if neglecting cell division and assuming that the single cells are in steady state, $\overline{R_{ss}}$ or $\overline{P_{ss}}$. Please note that for populations of identical cells in steady state, the single cell steady state coincides with the population average, $\overline{R_{ss}} = r_{ss}$ and $\overline{P_{ss}} = p_{ss}$ (see Equation 14). Thus, using Equations 7 and 13, the ratios are given by

$$\frac{\overline{R}}{\overline{R_\tau}} = \frac{v_{sr}}{k_{dr}} \cdot \frac{k_{dr} + \log(2)/\tau}{v_{sr}} = \frac{k_{dr} + \log(2)/\tau}{k_{dr}} = 1 + \frac{hl_r}{\tau} \qquad \text{(Equation 19)}$$

for the mRNA abundance and the molecular mRNA half-life $hl_r = \log(2)/k_{dr}$, and

$$\frac{\overline{P_{ss}}}{\overline{P_\tau}} = \frac{v_{sr} \cdot k_{sp}}{k_{dr} \cdot k_{dp}} \cdot \frac{(k_{dr} + \log(2)/\tau) \cdot (k_{dp} + \log(2)/\tau)}{v_{sr} \cdot k_{sp}} \qquad \text{(Equation 20)}$$

$$= \frac{(k_{dr} + \log(2)/\tau) \cdot (k_{dp} + \log(2)/\tau)}{k_{dr} \cdot k_{dp}} = 1 + \frac{hl_r}{\tau} \cdot \frac{1/hl_r + 1/hl_p + 1/\tau}{1/hl_p} \quad \text{(Equation 21)}$$

for the protein abundance, the molecular mRNA half-life $hl_r = \log(2)/k_{dr}$ and the molecular protein half-life $hl_p = \log(2)/k_{dp}$.

(ii) The cell division time strongly influences the differential gene expression analysis between conditions. Differences in population averages can result solely from different cell division times in the populations, while the rates of gene expression in the single cells of both populations remain identical. This effect can be quantified by comparing how big differences in population averages can become due to changes in cell division time only. Given the single cell rates of gene expression for two populations i and j are the same, $v^i_{sr} = v^j_{sr} = v_{sr}$, $k^i_{sp} = k^j_{sp} = k_{sp}$, $k^i_{dr} = k^j_{dr} = k_{dr}$, $k^i_{dp} = k^j_{dp} = k_{dp}$, but their cell division times τ_i and τ_j are different, the resulting differences in their mRNA or protein population averages, $\overline{R_i}$ and $\overline{R_j}$, or $\overline{P_i}$ and $\overline{P_j}$, are given by how strongly their below given ratio (fold change) differs from 1:

$$\frac{\overline{R_i}}{\overline{R_j}} = \frac{v^i_{sr}}{k^i_{dr} + \log(2)/\tau_i} \cdot \frac{k^j_{dr} + \log(2)/\tau_j}{v^j_{sr}} = \frac{k_{dr} + \log(2)/\tau_j}{k_{dr} + \log(2)/\tau_i} \quad \text{(Equation 22)}$$

for the mRNA abundances and

$$\frac{\overline{P_i}}{\overline{P_j}} = \frac{v^i_{sr} \cdot k^i_{sp}}{(k^i_{dr} + \log(2)/\tau_i) \cdot \left(k^i_{dp} + \log(2)/\tau_i\right)} \cdot \frac{(k^j_{dr} + \log(2)/\tau_j) \cdot \left(k^j_{dp} + \log(2)/\tau_j\right)}{v^j_{sr} \cdot k^j_{sp}} \quad \text{(Equation 23)}$$

$$= \frac{(k_{dr} + \log(2)/\tau_j) \cdot (k_{dp} + \log(2)/\tau_j)}{(k_{dr} + \log(2)/\tau_i) \cdot (k_{dp} + \log(2)/\tau_i)} \quad \text{(Equation 24)}$$

for the protein abundances.

(iii) Vice versa, even if the same abundances are observed between conditions, the single cell gene expression characteristics governed by the rates of gene expression can be different. This is in particular the case if the different conditions are governed by different cell division times. We quantify this effect of cell division by the ratio (fold change) of synthesis rate constants obtained if using Equations 15 and 17 for the same population average abundances $\overline{R_i} = \overline{R_j}$ and $\overline{P_i} = \overline{P_j}$ but different cell division times (while assuming otherwise identical parameters of gene expression, $k^i_{dr} = k^j_{dr}$, $k^i_{dp} = k^j_{dp}$) between conditions i and j. The ratios are given by

$$\frac{v^i_{sr}}{v^j_{sr}} = \frac{\overline{R_i} \cdot (k^i_{dr} + \log(2)/\tau_i)}{\overline{R_j} \cdot (k^j_{dr} + \log(2)/\tau_j)} = \frac{k_{dr} + \log(2)/\tau_i}{k_{dr} + \log(2)/\tau_j} \quad \text{(Equation 25)}$$

for mRNA synthesis rates and

$$\frac{k^i_{sp}}{k^j_{sp}} = \frac{\overline{P_i}/\overline{R_i} \cdot \left(k^i_{dp} + \log(2)/\tau_i\right)}{\overline{P_j}/\overline{R_j} \cdot \left(k^j_{dp} + \log(2)/\tau_j\right)} = \frac{k_{dp} + \log(2)/\tau_i}{k_{dp} + \log(2)/\tau_j} \quad \text{(Equation 26)}$$

for protein synthesis rate constants. In particular, please note that these ratios are independent of the actual values of the assumed population average abundances.

QUANTIFICATION AND STATISTICAL ANALYSIS

Resting vs. Activated B Cells Data

We employed proteomics and transcriptomics data given for resting and activated human memory B cells in (Rieckmann et al., 2017): After sorting blood cells from four donors, resting memory B cells have been subjected to activation. The proteome of both resting and activated B cells have been measured by mass spectrometry reporting IBAQ values. The transcriptome changes have been estimated by RNAseq of pooled samples of the resting or activated cells. We used protein half-lives for B cells from (Mathieson et al., 2018) assuming similar half-lives for both resting and activated cells. We restricted our analysis to proteins for which we had (i) at least 2 measured protein abundance values for resting as well as activated B cells, (ii) a measurement for the according mRNA for both resting and activated B cells, and (iii) at least 1 protein half-life measurement of good quality (according to (Mathieson et al., 2018)). These restrictions gave rise to 2438 proteins which we compared between conditions.

Classical: Differential Protein Expression

We compared the protein abundances for each single protein between resting and activated B cells by performing two-sided Welch's t-tests employing the IBAQ values as protein abundances (treating zeros as NAs, sample numbers: between 2 and 4 for each condition). We applied Benjamini-Hochberg multiple testing correction. Log fold-changes (FC) were computed by logarithmizing ($\log_{10}$) the ratio of the average abundances of resting vs. activated B cells. Proteins were considered significantly different between resting and activated B cells if the corrected p-value was above 0.05 and the fold-change more than 50% (i.e. $|\log10(FC)| > 0.176$). Please note that due to lack of absolute cellular quantification, the comparison between the IBAQ values can be interpreted as comparison of

Please cite this article in press as: Baum et al., Of Gene Expression and Cell Division Time: A Mathematical Framework for Advanced Differential Gene Expression and Data Analysis, Cell Systems (2019), https://doi.org/10.1016/j.cels.2019.07.009

population average molecule counts per cell between condition 1 and 2 only under the assumption 1= [cell count in sample]$_2$/·[cell count in sample]$_1$.

Alternative: Differential Protein Synthesis

We calculated the rate constant of protein synthesis (i.e. the translation rate constant) according to Equation 17 for each protein for each sample (between 2 and 4 for each condition). For the resting memory B cells we assumed no cell division (cell division time $\tau = \infty$, i.e. $\log(2)/\tau = 0$, (Jones et al., 2015)); for the activated memory B cells a cell division time of $\tau = 16$ h ((Milo et al., 2010), BNID: 109934) was considered. We used the IBAQ values from (Rieckmann et al., 2017) as population average abundance $\overline{P}$, and 2 to the power of the mRNA value as given in (Rieckmann et al., 2017) (which we interpreted as the binary logarithm of the RPKM value) for the average mRNA abundance $\overline{R}$. We used $k_{dp} = \log(2)/hl_p$, with hl_p being the molecular protein half-life for each protein as obtained from (Mathieson et al., 2018). If two half-lives of good quality were measured for a protein they were averaged; weak-quality half-lives were considered with a weight of 1/3 in weighted averaging. Note that the data lack absolute quantification of the protein and mRNA abundances and therefore only a comparison between conditions 1 and 2 is valid, with the assumption 1=[total amount of mRNA nucleotides/cell]$_1$/[total amount of mRNA nucleotides/cell]$_2$·[cell count in sample]$_2$/[cell count in sample]$_1$.

We compared the synthesis rate constants between the two conditions exactly as we compared the population average abundances. In both cases we computed FDR-corrected p-values of two-sided Welch's t-tests and calculated the $\log_{10}$ fold-changes by logarithmizing the average protein synthesis rates of resting vs. activated B cells. Proteins were considered having significantly different protein synthesis rates for a corrected p-value below 0.05 and a fold-change of more than 50% ($|\log10(FC)| > 0.176$).

GO Enrichment and GO Term Clustering

We characterized those proteins that are significantly affected by B cell activation by a GO enrichment analysis from the complete biological process (BP) annotation using PANTHER (Mi et al., 2017). As input, we converted the *ALIAS* identifier to ENTREZ IDs using the R package *clusterProfiler* (Yu et al., 2012) and the gene identifier annotation from the package *org.Hs.eg.db.* (Carlson, 2016) and employed otherwise the default PANTHER settings (PANTHER Overrepresentation Test, released 2017-12-05, PANTHER version 13.1, released 2018-02-03). We only considered GO-terms with 10 and up to 300 annotated genes in order to avoid too general and too specific terms. We considered GO terms significantly enriched for a Benjamini-Hochberg-corrected Fisher's exact test p-value of below 0.01.

We then separately clustered the two sets of significantly enriched GO terms into sub-categories and categories. First, we estimated the similarity between the GO terms using the function *getTermSim* from the R package *GOSim* (Fröhlich et al., 2007) with the distance method *relevance*. Second, we applied a *friends of friends* clustering with a similarity threshold >0.75 using a custom R code. Third, we inspected the larger clusters (size ≥ 3) and manually assigned sub-categories. Remaining terms of the largest cluster and terms not associated to a cluster were assigned manually. Sub-categories related to metabolism, protein, RNA, DNA and immune system were joined in categories, for all other processes sub-category and category coincide.

Cell Systems, Volume *9*

Supplemental Information

Of Gene Expression and Cell Division Time: A Mathematical Framework for Advanced Differential Gene Expression and Data Analysis

Katharina Baum, Johannes Schuchhardt, Jana Wolf, and Dorothea Busse

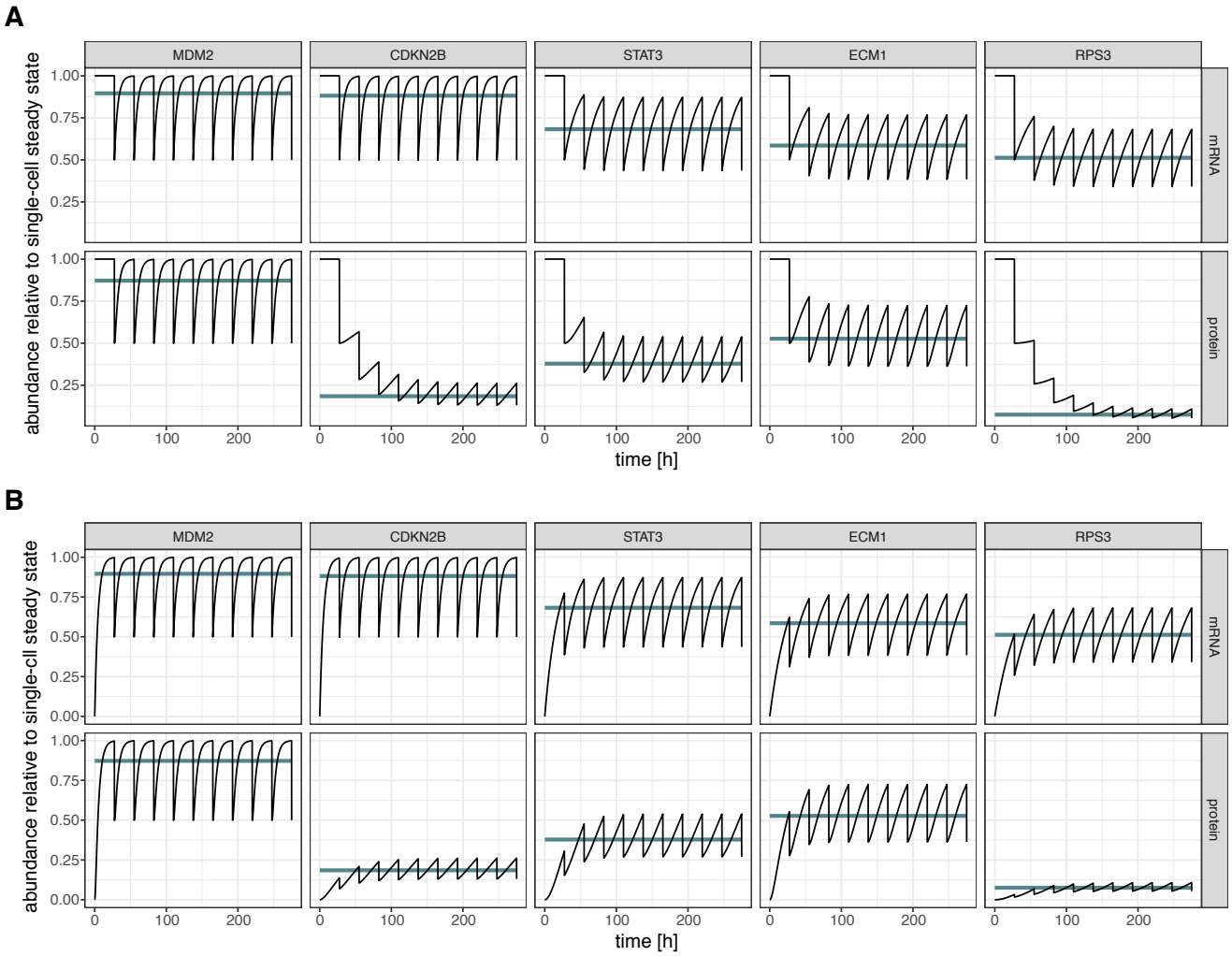


Figure S1. Single cell kinetics of mRNA and protein abundance. Related to Figure 2, Figure 3.
Depicted are relative mRNA and protein abundances of MDM2, CDKN2B, STAT3, ECM1 and RPS3 (black lines) over time for a cell which divides into two cells at $\tau = 27.5$ h; only one descendant cell is tracked (values of kinetic parameters given in Table S1). The blue lines give the population average abundances (relative to single cell steady state levels). Starting abundances are the steady state levels (**A**), or zero abundance (**B**). Two observations are made: First, the different mRNA and protein half-lives influence the relationship between the population average mRNA and protein abundances R and P (blue lines) and the corresponding single cell steady states (values are normalized to steady state, therefore the steady state values equal one). The longer the half-lives the more distant the population averages are from the steady states (compare also Figures 2B and 3). Second, the mRNA abundances and protein abundances within the single cells converge very fast: After 1-7 divisions, the abundance at cell birth of a daughter cell is similar to the abundance at cell birth of its mother cell, for both mRNA and protein ($r(\tau) = 2 \cdot r(0)$ and $p(\tau) = 2 \cdot p(0)$). mRNAs and proteins with long half-lives tend to take longer until the transient phase for reaching this state is completed.

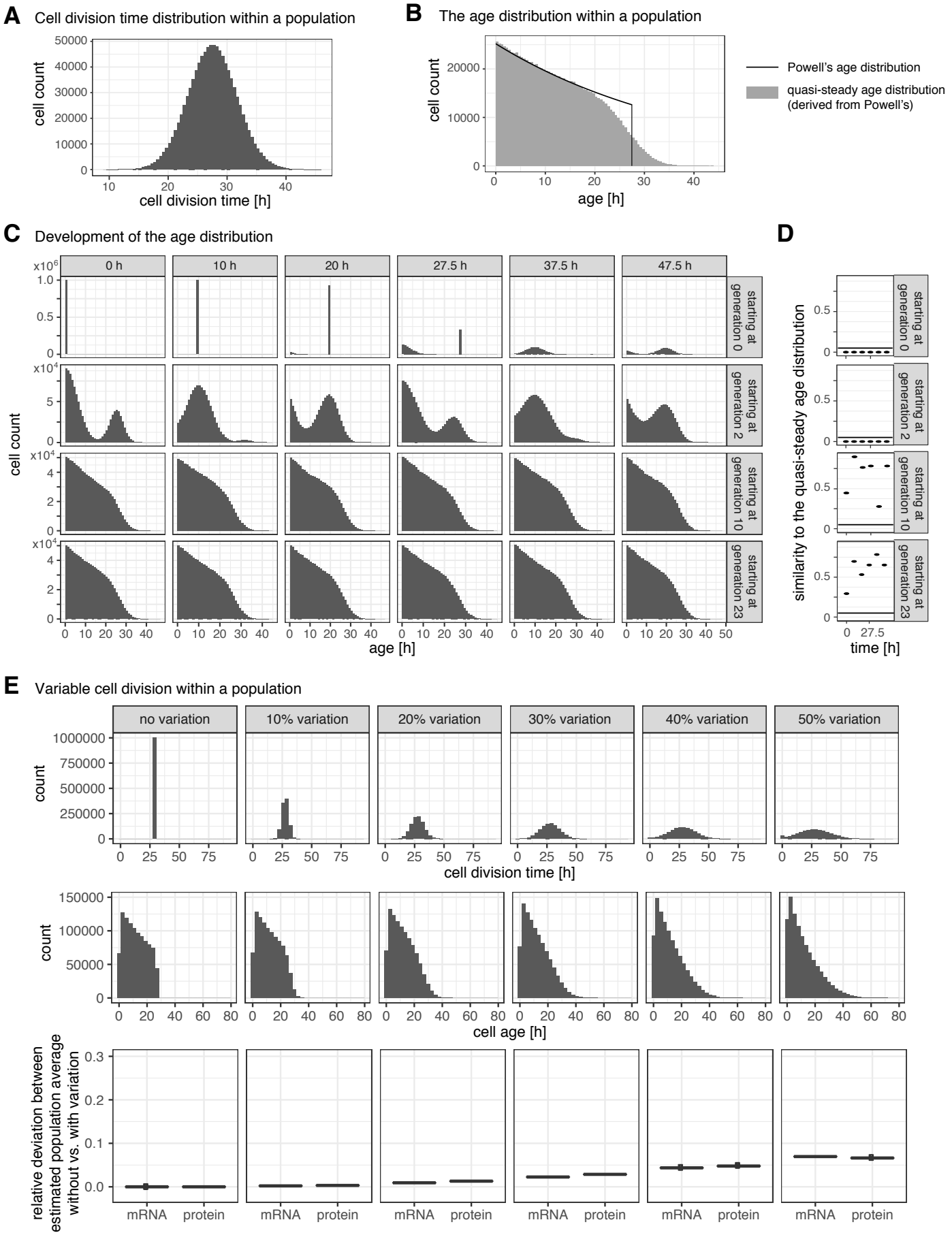


Figure S2. The age distribution within a population. Related to Figure 2.

A: Normal distribution of cell division times τ with a mean of 27.5 h and a standard deviation of 15%. Histogram for 10^6 cells.

B: Histogram of a quasi-steady age distribution for a population with cell division times as in A as derived by (Powell, 1956). Solid line: age distribution for cell populations with a cell division time of exactly 27.5 h.

C: Simulation of the development of an age distribution of 10^6 initially synchronized cells over time. The cell division time was updated three times per generation (0h, 10h, 20h) for 25 generations, cell division times of new-born cells were randomly assigned from a distribution as described in A. Distributions are shown only for generations 0-1 (first row), 2-3 (second row), 10-11 (third row), and 23-24 (fourth row). Finally, a stable age distribution evolves which is similar to that from B. Please note that the binwidths in the histograms in B and C are different and therefore the values on the y-axis differ.
D: Similarity of the simulated age distributions from C to the quasi-steady age distribution in B (as measured by the Benjamini-Hochberg corrected p-values of the Kolmogorov-Smirnov test for 10^4 cells at each update of the population). In each row, the similarities obtained for the six corresponding histograms from B are given as circles. Similarities above the black horizontal line mean that the distributions are statistically identical (corrected p-value < 0.05). Over time, the age distribution develops towards the quasi-steady distribution corresponding to Powell's.
E: Histograms of the cell division times τ within a population of 10^6 cells for different degrees of variation around 27.5 h (top) and the corresponding quasi-steady age distributions within the populations (middle). Bottom: Boxplots of relative deviation of the simulated STAT3 mRNA and protein population average abundances with variation in cell division times (and otherwise identical kinetic parameters) from the respective abundances without variation, for 100 populations each. Only slight differences up to 7% are observed even for large variation in the cell division times and consequently age distributions.

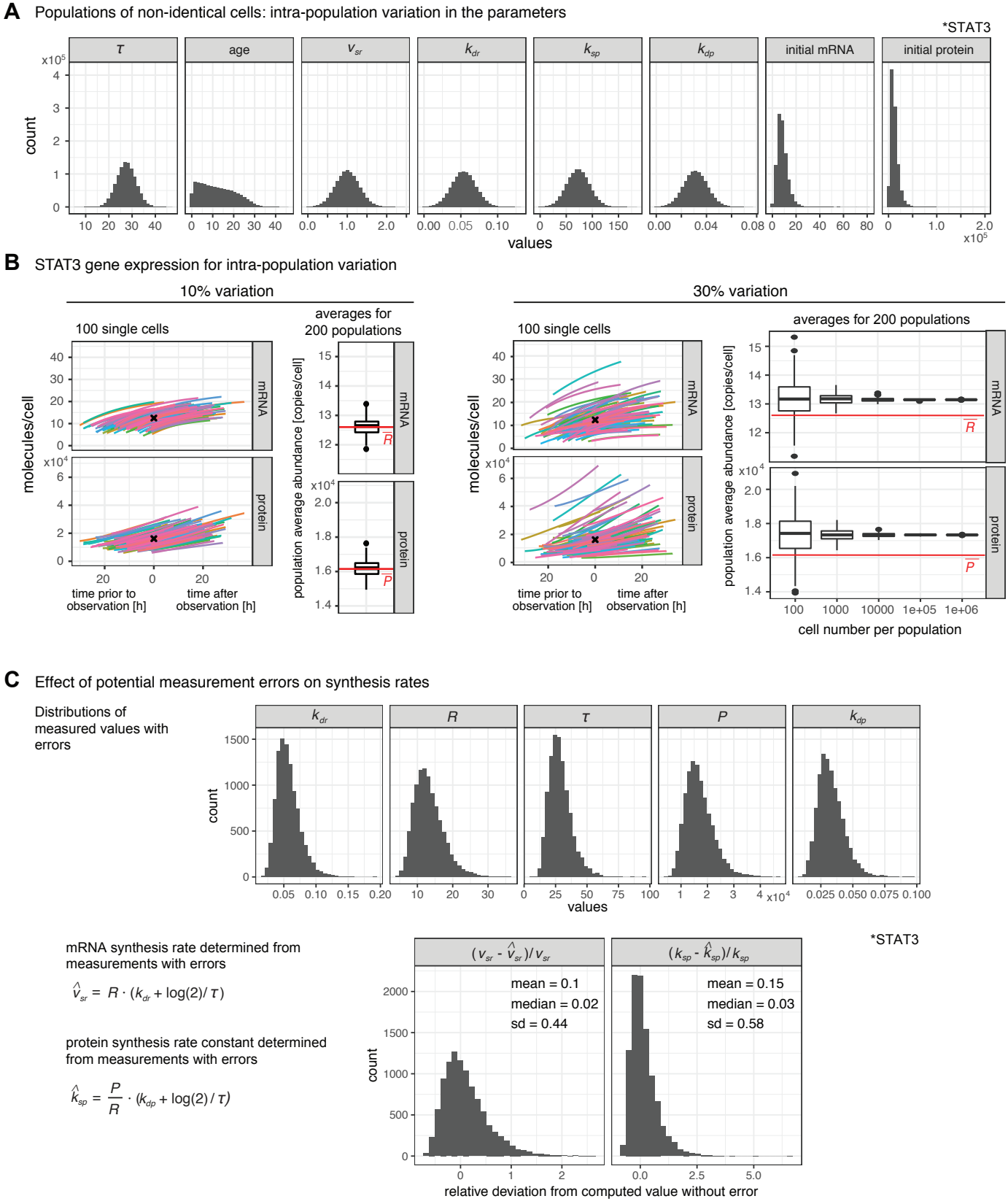


Figure S3. Populations of non-identical cells, and effect of potential measurement errors on synthesis rates. Related to Figure 2.

A: Distributions assumed for the cell division time, cell age, kinetic parameters and initial conditions for representing variation between single cells within a population (histograms of 10^6 sampled values for the example of STAT3, parameter values given in Table S1, τ = 27.5 h).

B: Simulations of STAT3 gene expression dynamics (mRNA top and protein bottom) in 100 single cells are shown allowing for 10% (left) and 30% (right) variation in v_{sr}, k_{dr}, k_{sp}, k_{dp} and the initial abundances at cell birth, and 15% for the cell division time. The population averages (arithmetic mean) of 200 such simulations of a population of ≥ 100 cells (100, 1000, 10^4, 10^5 or 10^6 cells, right) are presented in boxplots. The red lines show the calculated average population mRNA and protein abundances using the equations in Figure 2D.

With increasing variability, the small positive shift between the simulated average mRNA and protein abundances and the analytically derived abundances assuming identical cells increases slightly. With increasing cell number per population, the variation between populations is reduced and the shift turns more stable.

C: Effect of possible measurement errors on the calculated mRNA and protein synthesis rate constants. For the measured quantities: degradation rate constant k_{dr}, population average mRNA abundance R, cell division time τ, population average protein abundance P and protein degradation rate constant k_{dp} we assumed a log-normal distribution with standard deviation of 30% (top, shown for quantities of STAT3 for 10^4 sampled values, see Table S1). The relative deviations of the calculated synthesis rate constants with and without measurement error are characterized by the resulting distributions. These have a larger width, standard deviations are 44% and 58% for mRNA (bottom left) and protein (bottom right), respectively. Similar dispersions between 41-46% for the transcription rate and 57-64% for the translation rate constant are obtained for other mRNA-protein pairs (Table S1) for cell division times sampled around 27.5 h.

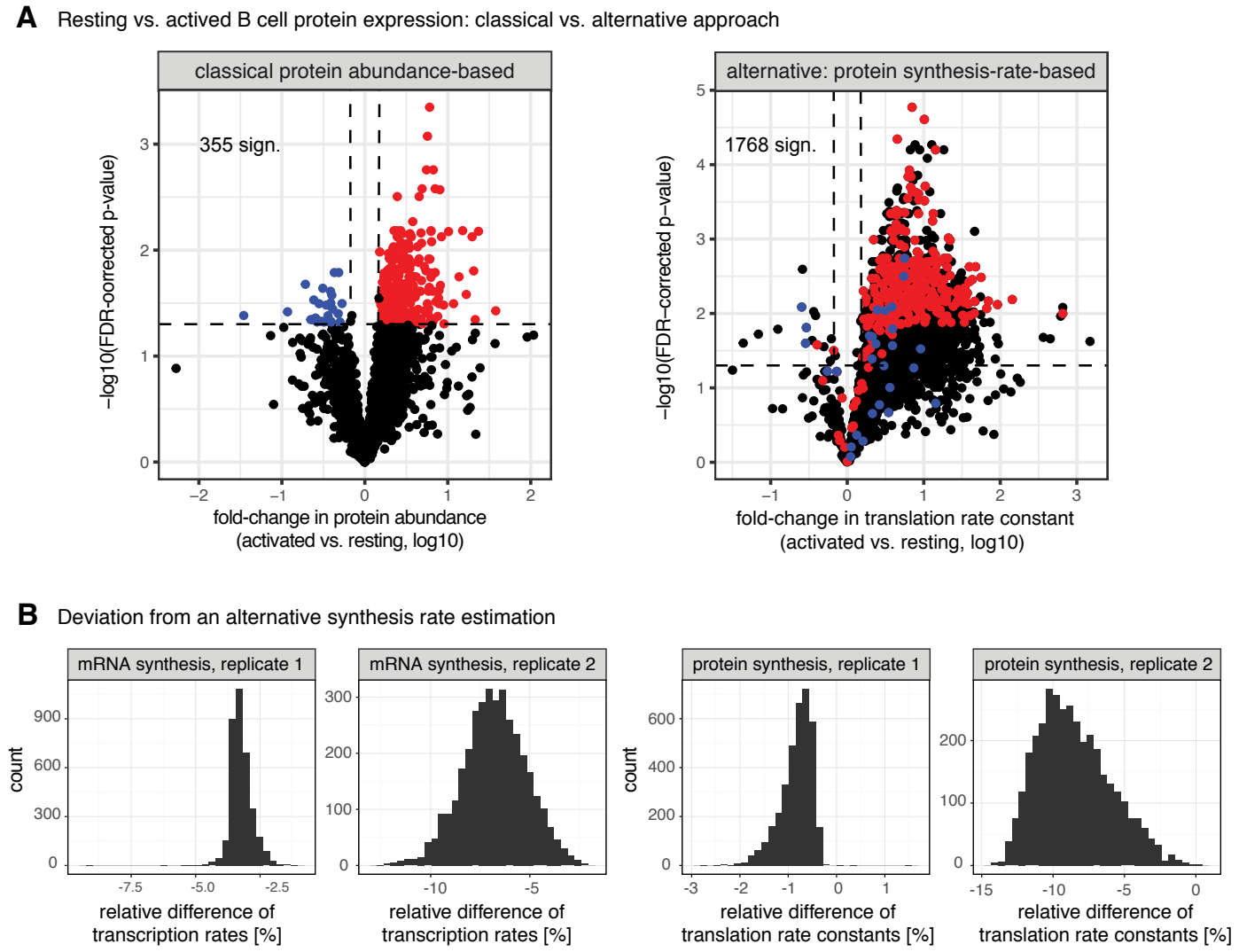


Figure S4. Application of the derived formulas. Related to Figure 4.

A: Differential protein expression in resting vs. activated B cells. Benjamini-Hochberg-corrected Welch's test p-values vs. fold changes of protein abundances as measured by IBAQ (Rieckmann et al., 2017) (classical approach, left) or of protein synthesis rate constants as computed from Eq. Figure 2D (STAR methods Eq. 17, alternative approach, right). The 327 proteins with significantly increased abundance using the classical approach are marked in red, the 28 proteins with significantly decreased abundance using the classical approach in blue (both left and right). 1768 proteins were detected as significantly different, 1442 of which up-regulated, using the alternative approach.

B: Deviation from an alternative synthesis rate estimation. Transcription and translation rates calculated according to the transformed equations in Figure 2D (STAR methods Eqs. 15 and 17) are compared to those calculated in (Schwanhäusser et al., 2013) (the latter indexed by 'Schw'). Shown are the differences of the values calculated for the 3569 mRNA-protein pairs with complete data for the two replicate data sets reported in (Schwanhäusser et al., 2013). Relative differences of the transcription rates, $(v_{sr}^{Schw} - v_{sr})/v_{sr}$, (left panels) and of the translation rate constants, $(k_{sp}^{Schw} - k_{sp})/k_{sp}$, (right panels) are given in percent for each replicate. In contrast to the consideration of a heterogeneous age-distribution in a population of growing cells, our earlier approach (Schwanhäusser et al., 2011) considers the time-average over a cell cycle. Overall, we found only small relative differences between the two approaches due to the near homogeneous age distribution of NIH3T3 cells.

Table S1: Parameter values for gene expression of STAT3, MDM2, CDKN2B, ECM1, RPS3. Related to Figure 2, Figure 3, Figure S1, Figure S2, Figure S3 and Table S2.
Kinetic parameter values derived from (Schwanhäusser et al., 2013) in NIH3T3 cells.

species	v_{sr} [no./h]	k_{sp} [1/h]	k_{dr} [1/h]	k_{dp} [1/h]	species long name
STAT3	1	72.5	log(2)/12.8	log(2)/22.1	Signal transducer and activator of transcription 3
MDM2	610.5	11.5	log(2)/3.2	log(2)/0.74	E3 ubiquitin-protein ligase
CDKN2B	2.98	386.62	log(2)/3.69	log(2)/103.51	Cyclin-dependent kinase inhibitor 2B
ECM1	3.72	15.92	log(2)/19.48	log(2)/3.06	Extracellular matrix protein 1
RPS3	15.25	913.95	log(2)/26.14	log(2)/159.34	40S ribosomal protein S3

Table S2: Sensitivity to variability between cells of the populations. Related to Figure 2, Figure S3.
We considered populations of non-identical cells with respect to the kinetic parameters of gene expression and cell division, and non-exact doubling of the abundances from cell birth to division. The sensitivity of the derived formulas (Eqs. Figure 2D) towards this intra-population variation is quantified by the relative difference, $(\widehat{R}\text{-}\overline{R})/\overline{R}$ or $(\widehat{P}\text{-}\overline{P})/\overline{P}$, between the population averages obtained when sampling 200 populations of 10^6 variable cells, $\widehat{R}$ or $\widehat{P}$, and the population average obtained for a population of identical cells, $\overline{R}$ or $\overline{P}$ (Eqs. Figure 2D). This relative shift is reported for five mRNA-protein pairs (Table S1) and a standard deviation of 30% in the kinetic parameters of gene expression and the initial abundances, and a standard deviation of 15% for different cell division times τ of 16 h, 27.5 h or 65.5 h. The sensitivity towards intra-population variability for the presented combinations of half-lives and cell division times is on average 5.8% for mRNA and 12.8% for protein. Even for the special case of the very unstable MDM2 mRNA and protein, the shift is only at most 32%.

species	mRNA half-life	protein half-life	τ	sensitivity mRNA	sensitivity protein
MDM2	short	short	16 h	7.6%	21.3%
	short	short	27.5 h	9.7%	25.3%
	short	short	65.5 h	12.4%	31.9%
CDKN2B	short	long	16 h	1.4%	9.6%
	short	long	27.5 h	2.9%	13.1%
	short	long	65.5 h	6.1%	19.5%
STAT3	intermediate	intermediate	16 h	2.5%	4.3%
	intermediate	intermediate	27.5 h	4.3%	7.3%
	intermediate	intermediate	65.5 h	7.7%	13.9%
ECM1	long	short	16 h	7.1%	7.5%
	long	short	27.5 h	9.2%	9.5%
	long	short	65.5 h	12.0%	13.2%
RPS3	long	long	16 h	0.8%	2.1%
	long	long	27.5 h	2.0%	2.9%
	long	long	65.5 h	4.9%	5.8%