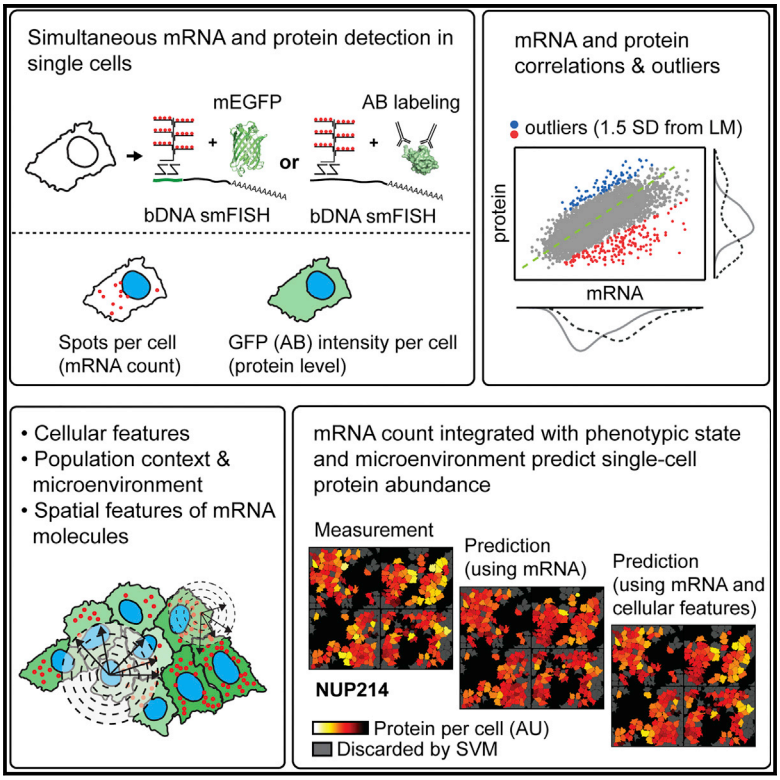
Article

Cell Systems

Multivariate Control of Transcript to Protein Variability in Single Mammalian Cells

Graphical Abstract



Highlights

- Image-based quantification of mRNA and protein abundance in thousands of single cells
- Variability in mRNA-protein ratios is determined by cellular state and microenvironment
- Cellular state and microenvironment influence *JUN* gene expression at multiple levels
- Variability in cytoplasmic patterning of *JUN* mRNA depends on the microenvironment

Authors

Doris Popovic, Birgit Koch, Moritz Kueblbeck, Jan Ellenberg, Lucas Pelkmans

Correspondence

lucas.pelkmans@imls.uzh.ch

In Brief

Popovic et al. show that in adherent mammalian cells, variation in the ratio between mRNA and protein levels of the same gene stems from differences in the phenotypic state and microenvironment of individual cells and that the level of protein expressed by a gene can be predicted by integrating these properties with its mRNA abundance in individual cells.

Popovic et al., 2018, Cell Systems 7, 1–14
October 24, 2018
https://doi.org/10.1016/j.cels.2018.09.001

CellPress

Please cite this article in press as: Popovic et al., Multivariate Control of Transcript to Protein Variability in Single Mammalian Cells, Cell Systems (2018), https://doi.org/10.1016/j.cels.2018.09.001

Cell Systems
Article

Multivariate Control of Transcript to Protein Variability in Single Mammalian Cells

Doris Popovic,[1] Birgit Koch,[2] Moritz Kueblbeck,[2] Jan Ellenberg,[2] and Lucas Pelkmans[1,3,*]
[1]Department of Molecular Life Sciences, University of Zurich, Zürich, Switzerland
[2]European Molecular Biology Laboratory, Heidelberg, Germany
[3]Lead Contact
*Correspondence: lucas.pelkmans@imls.uzh.ch
https://doi.org/10.1016/j.cels.2018.09.001

SUMMARY

A long-standing question in quantitative biology is the relationship between mRNA and protein levels of the same gene. Here, we measured mRNA and protein abundance, the phenotypic state, and the population context in thousands of single human cells for 23 genes by combining a unique collection of cell lines with fluorescently tagged endogenous genomic *loci* and quantitative immunofluorescence with branched DNA single-molecule fluorescence in situ hybridization and computer vision. mRNA and protein abundance displayed a mean single-cell correlation of 0.732 at steady state. Single-cell outliers of linear correlations are in a specific phenotypic state or population context. This is particularly relevant for interpreting mRNA-protein relationships during acute gene induction and turnover, revealing a specific adaptation of gene expression at multiple steps in single cells. Together, we show that single-cell protein abundance can be predicted by multivariate information that integrates mRNA level with the phenotypic state and microenvironment of a particular cell.

INTRODUCTION

The linearity and accuracy of information transfer from a gene to its gene products that define the cellular phenotype have been investigated for many years (Liu et al., 2016), in particular, the correlation between mRNA and protein abundance. An important distinction among these studies is between those that analyze correlations between protein and mRNA levels across multiple genes in one sample and those that analyze such correlations for individual genes across multiple samples. Numerous studies using bulk approaches have been performed to analyze the extent of correlation between these two molecular species, identifying gene-specific and contextual effects, as well as non-linear relationships (Csárdi et al., 2015; Edfors et al., 2016; Fortelny et al., 2017; Golding et al., 2005; Greenbaum et al., 2003; Ideker et al., 2001; Jovanovic et al., 2015; Li et al., 2014; Liu et al., 2016; Peshkin et al., 2015; Schwanhäusser et al., 2011; Washburn et al., 2003; Wilhelm et al., 2014).

A further distinction should be made for studies that analyze correlations between mRNA and protein across single cells that are genetically identical and exposed to identical conditions. The first study quantifying both molecular species and their correlation in bacteria using endogenous gene tagging (Taniguchi et al., 2010) reported this correlation to be largely non-existent. In mammalian cells, depending on the cell type and genes studied, reported single-cell correlations vary (Albayrak et al., 2016; Genshaft et al., 2016; Peterson et al., 2017; Stoeckius et al., 2017). Accurate quantification of mRNA and protein levels in single mammalian cells has, however, been technically challenging due to the lack of systematic endogenous gene tagging and the problem of sensitivity and sampling bias. Moreover, adherent mammalian cells display emergent properties at the population level due to local effects, resulting in a wide variety of cell sizes, shapes, and positions in the cell cycle (Gut et al., 2015; Snijder et al., 2009). These determine a large amount of cell-to-cell variability in cytoplasmic mRNA abundance (Battich et al., 2015) and may well influence mRNA-to-protein relationships in single cells.

Here, we use a recently made available collection of human cell lines (HeLa cells and induced pluripotent stem cells) in which various proteins were tagged with monomeric enhanced green fluorescent protein (mEGFP) at their endogenous genomic loci using CRISPR-Cas9 or zinc-finger technology (Cai et al., 2018; Koch et al., 2018; Politi et al., 2018; Reichmann et al., 2018; Roberts et al., 2017; Walther et al., 2018). On these cells, we applied a sensitive high-throughput (HT) automated method of single-molecule fluorescence in situ hybridization (sm-FISH) (Battich et al., 2013), in combination with large-scale high-resolution imaging and quantitative image processing. To complement the use of fluorescent protein tagging, we also combined HT automated sm-FISH with indirect immunofluorescence imaging to quantify both mRNA and protein levels of non-tagged genes. This allowed us to obtain measurements of protein and mRNA abundance of endogenously expressed genes from thousands of single cells, as well as a large set of features quantifying their phenotypic state and the population context of each single cell. These measurements reveal that while human adherent cells display high cell-to-cell variability in mRNA and protein abundance, correlations between these molecular species across single cells are evident for each of the 23 genes analyzed. Moreover, when the phenotypic state and microenvironment of individual cells is considered, outliers in simple linear correlation analyses can, in addition, be explained. Finally, we show that during gene induction of *JUN*, an immediate early response gene, the lag times between, and the rates of mRNA and protein

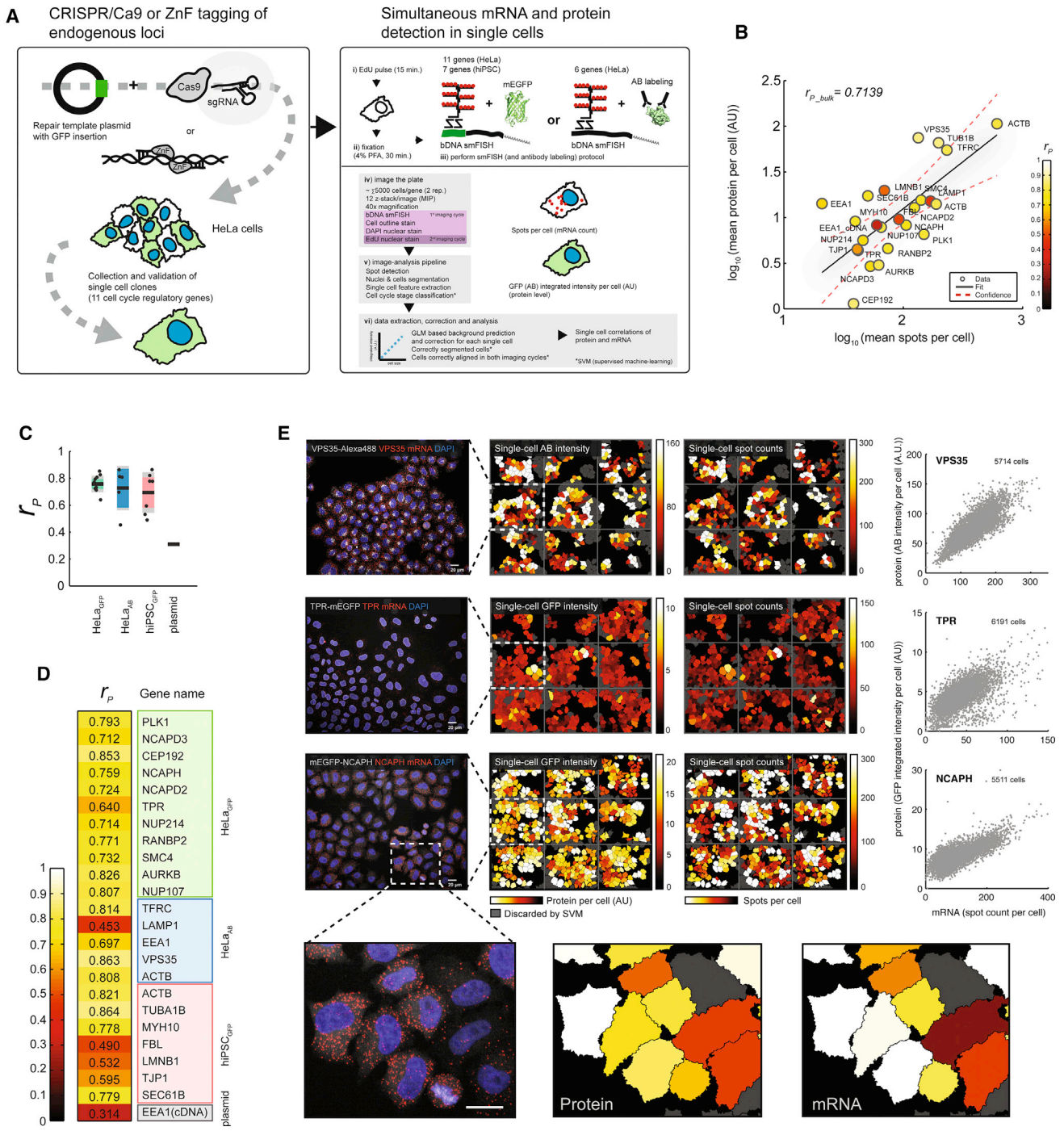


Figure 1. Image-Based Quantification of mRNA and Protein Levels and Cell-to-Cell Variability in Their Abundance

(A) Scheme of the workflow, left to right: generation of endogenously tagged cells, detection and quantification of both protein and mRNA, extraction of cellular features, and data clean-up (badly segmented cells and border cells).

(B) Scatterplot of bulk (cell-averaged) mRNA and bulk (cell-averaged) protein abundances of all inspected genes, with indicated correlation between these values across all 23 genes. Each gene is, in addition, colored according to the Pearson's correlation coefficient (mean of two technical replicates as listed in (D) between its mRNA and protein abundance across single cells.

(C) Plots of single-cell correlation coefficients between mRNA and protein abundances for individual genes, separated by cell line (HeLa and hiPS cells) and protein detection method (GFP or antibody, AB). Plasmid indicates the single-cell correlation coefficient between mRNA and protein abundance of EEA1 expressed as cDNA from a plasmid.

(D) Pearson's correlation coefficients (mean r_P of two technical replicates) between mRNA and protein abundance for individual genes calculated across all single cells; same values are used by the coloring of genes in (B).

(legend continued on next page)

synthesis and degradation are different for single cells in different phenotypic states and microenvironments. We also provide evidence that some of these differences could be the result of regulated intracellular positioning of cytoplasmic mRNA transcripts at the single-cell level.

RESULTS

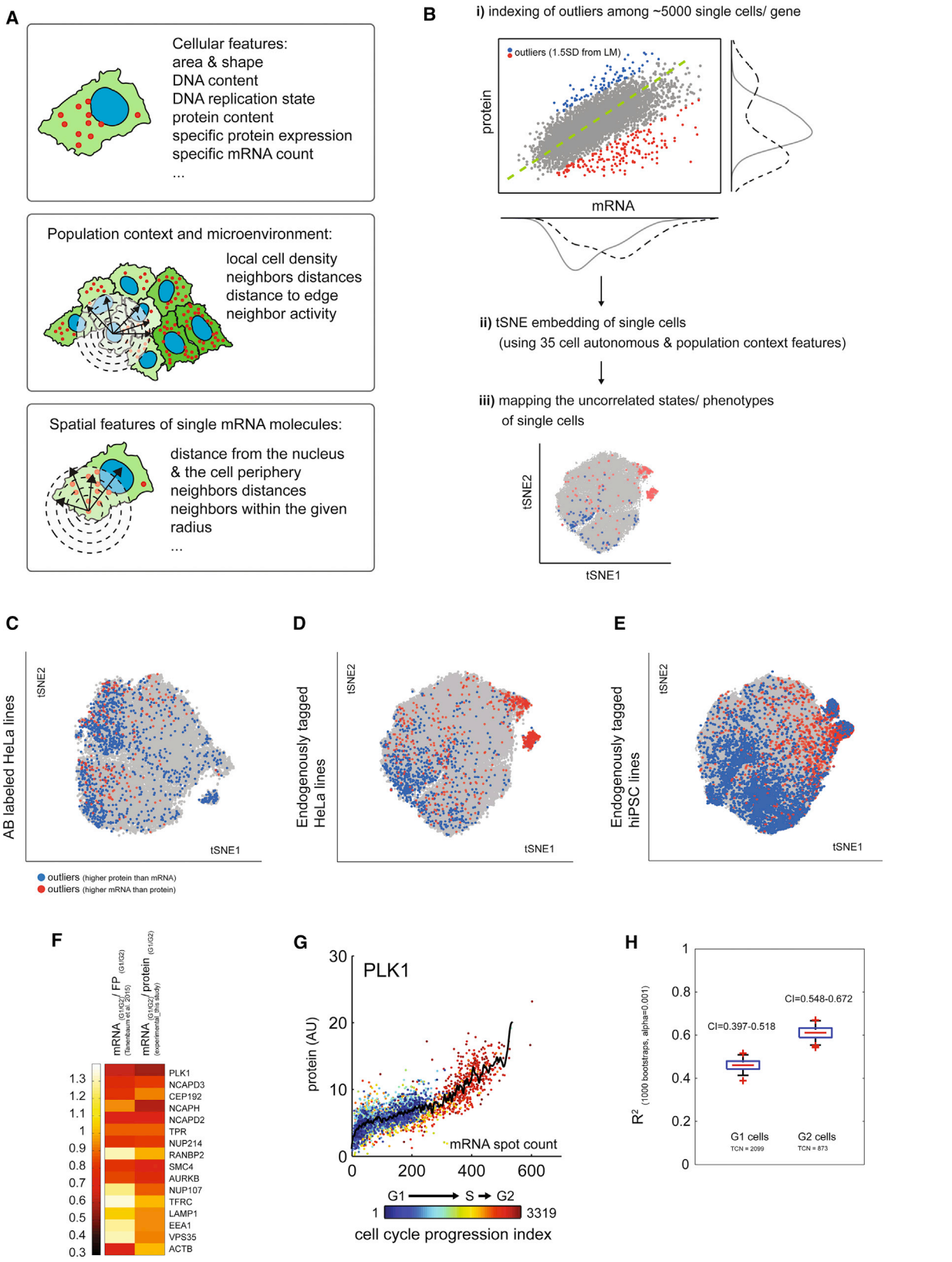
Image-Based Measurements of Endogenous mRNA and Protein Levels in Human Cells

Endogenous gene tagging with a fluorescent protein allows highly sensitive quantification of endogenous protein levels in single cells (Politi et al., 2018; Wachsmuth et al., 2015). Furthermore, the fluorescence intensity is known to be proportional to the number of protein molecules and can therefore be used to determine the relative amounts of proteins within cells (Lo et al., 2015). We therefore decided to use two recently made available sets of human cell lines, the first comprising 11 lines derived from HeLa cells (Koch et al., 2018; Wachsmuth et al., 2015) and the second comprising 7 lines derived from human-induced pluripotent stem (hiPS) cells (Roberts et al., 2017). In each cell line, one gene was tagged with a mEGFP at the endogenous locus using CRISPR-Cas9 or zinc-finger technology. To quantify the mRNA abundance of the tagged genes in single cells, we used sm-FISH with branched DNA (bDNA)-based oligonucleotide probes (bDNA sm-FISH) against the mEGFP sequence in the transcript. bDNA sm-FISH enables absolute mRNA quantifications with detection efficiencies above 85% in the cytoplasm, or nucleus, of single cells. At the cell-averaged level, quantification of mRNA abundance with bDNA sm-FISH agrees well with RNA-seq across 857 human genes ($r = 0.8–0.9$) (Battich et al., 2013). Moreover, given its false-positive count of less than 1 transcript per cell and 5-fold higher signal-to-noise ratio compared to other sm-FISH methods, bDNA sm-FISH is well suited also for genes with low expression levels and works in a HT setup (Battich et al., 2013, 2015). Thus, we combined the imaging of fluorescently tagged proteins with bDNA sm-FISH in automated large-scale microscopy (Figure S1A). Additionally, we used bDNA sm-FISH with indirect immunofluorescence to quantify protein levels of 5 untagged genes (Figure 1A). Using a previously established platform for image-based transcriptomics (Battich et al., 2013; Stoeger et al., 2015), we obtained reproducible transcript counts in thousands of single cells, together with their relative background-corrected protein quantities derived from the integrated fluorescence intensities of the mEGFP or fluorescent antibody signal (Figures 1A and S1B–S1G; Tables S1–S3). We also inferred the cell-cycle position of cells by pulsing them with EdU for 15 min prior to the fixation (Gut et al., 2015).

In total, we performed these measurements for 23 genes that encode components of the cytoskeleton, endocytic, and exocytic organelles, the nuclear pore complex, cell adhesion structures, cell cycle regulators, and kinases. At the mean (bulk) level, we observed widely varying amounts of cytoplasmic mRNA spot counts and protein intensities, with the lowest expressed gene being EEA1 at 20 transcripts per cell (which is within 53% of the lowest expressed genes according to RNA-seq) and the highest expressed gene being ACTB at >610 transcripts per cell (which is within 2% of highest expressed genes according to RNA-seq) (Battich et al., 2013) (Figure 1B). The correlation between bulk (cell-averaged) mRNA and protein levels across the 23 genes was $r = 0.714$ (Figure 1B). For a bulk analysis, this number of genes is obviously small compared to other studies, which studied bulk mRNA-protein relationships across several thousands of genes (Edfors et al., 2016; Jovanovic et al., 2015; Lawless et al., 2016; Schwanhäusser et al., 2011). Nevertheless, the explained variance (R^2) across genes in our bulk analysis of HeLa and hiPS cells combined was 0.51, which compares well to previous bulk studies reporting an (uncorrected) R^2 of 0.47–0.54 in mouse bone-marrow-derived dendritic cells (Jovanovic et al., 2015), 0.37–0.41 in MCF7 and NIH3T3 cells (Schwanhäusser et al., 2011), 0.15–0.62 in a range of different cancer-derived cell lines (Edfors et al., 2016), and 0.58 in *S. cerevisiae* (Lawless et al., 2016).

At the single-cell level, we observed that correlations between the cytoplasmic mRNA spot count and protein intensity across thousands of single cells varied for individual genes, ranging from $r = 0.453$ to $r = 0.864$, with a mean of $r = 0.732$ (Figures 1C and 1D). These correlations are generally higher than previous findings using single-cell measurements, which reported r values lower than 0.1 (Albayrak et al., 2016; Darmanis et al., 2016; Genshaft et al., 2016), but comparable to a recent study which reported r values up to 0.86 based on single-cell sequencing of transcripts and DNA-barcoded antibodies (Peterson et al., 2017). This can be appreciated in cell population overviews in which single-cell expression levels of both molecular species are visualized (Figure 1E). The extent of single-cell correlation did not depend on the mean mRNA or protein expression level of the gene, as was evident for genes with different expression levels (Figures S1H–S1J). For instance, we found similar single-cell correlations for *EEA1* ($r = 0.697$) and *NCAPD2* ($r = 0.724$), even though *EEA1* is expressed at a mean mRNA spot count of 20 molecules per cell (Figure 1B), while *NCAPD2* has a mean mRNA spot count of 124 molecules per cell (Figure 1B). The scaling of protein with mRNA abundance was also gene specific, as evident from the different slopes of their linear regression curves of single-cell mRNA versus protein abundances (Figure S1K). Moreover, genes with unimodal mRNA and protein distributions (e.g., endocytosis regulators, nuclear pore proteins, and cytoskeletal components) displayed similar correlations ($r = 0.453–0.864$) as genes with bi- or multimodal distributions (e.g., *CEP192, NCAPD2, NCAPD3, NCAPH, PLK1,* and *SMC4*) ($r = 0.712–0.853$) (Figures S1K and S1L). We also observed that higher cell-to-cell variability in mRNA abundance was accompanied by a higher cell-to-cell variability in protein content ($p < 10^{-4}$) (Figures S1L–S1O). Finally, when we transiently expressed an exogenous cDNA version of *EEA1* without introns and lacking the endogenous 3′- and 5′-untranslated ends on

(E) Example images of single cells from image-based quantifications (protein stain or mEGFP signal in gray, bDNA sm-FISH in red). Middle-right: visualization of the quantified protein (integrated intensity [AU]) and transcript abundance (spots) per cell by pseudo-coloring single-cell segmentations. Cells discarded by supervised machine learning (SVM) are gray (badly segmented and border cells). Right: scatterplots of all measured single cells for the corresponding example images.



(legend on next page)

the transcript (Lawe et al., 2000), the single-cell correlation between cytoplasmic mRNA and protein abundance (r = 0.31; 99.99% CI = 0.26–0.37) was more than 2-fold lower than for the endogenous gene (r = 0.70; 99.99% CI = 0.67–0.73) (see Supplemental Information and STAR Methods), suggesting the involvement of endogenous transcriptional and posttranscriptional mechanisms in establishing single-cell relationships between both molecular species. These measurements reveal that for the 23 genes interrogated, cytoplasmic levels of mRNA and their corresponding protein abundance are, at steady state, generally well correlated at the single-cell level in mammalian cells and that the scaling and extent of correlation between these molecular species are gene specific.

Correlation Outliers Reveal Cellular State-Specific mRNA-to-Protein Ratios in Individual Cells

We next asked whether the single cells that are outliers in a linear correlation analysis between steady-state mRNA and protein abundance across single cells from a whole cell population display specific phenotypic states or microenvironments, which may suggest the presence of a non-linear relationship. To explore this, we used our image-based transcriptomics platform (Battich et al., 2013; Stoeger et al., 2015; Battich et al., 2015) to extract ~180 features that quantify properties of the cell and nucleus shape and area, of protein and DNA content and texture, and of the extent of local cell density, number of neighbors, and relative location to other cells and to empty space in the cell population. We also extracted features describing DNA replication activity (using the EdU stain) and 34 features quantifying the spatial positioning of each mRNA molecule within each single cell (Figure 2A).

We selected single cells that were outliers in mRNA-to-protein ratios as determined by being 1.5 standard deviations away from linear regression fits between single-cell mRNA and protein abundance of each gene (Figure 2B) and visualized whether these cells are, in particular, phenotypic or microenvironmental states by highlighting them in two-dimensional (2D) maps (Figures 2C–2E). These maps are created with all cells from the various analyzed genes, pooled per cell line and protein detection method, by means of Barnes Hut t-distributed stochastic neighbor embedding (tSNE) (van der Maaten, 2008), using a multivariate set of principal component-reduced features of the phenotypic state of single cells (e.g., DNA content, replication, and cell and nucleus size) and their microenvironment (e.g., local cell density). tSNE is a machine learning algorithm used for non-linear dimensionality reduction of high-dimensional data and its visualization in a low-dimensional space. More specifically, this transforms the data such that similar objects (single cells with similar feature values) will be placed close to each other, while dissimilar objects will be placed far from each other. For single cells, this can be evaluated by projecting the original values of single-cell features onto each data point (which are single cells) within the 2D tSNE plot, showing that cells of similar properties are close to each other and occupy a specific region in the plot (Figure S2A). The Barnes Hut variant of tSNE significantly reduces computational cost, hence being more applicable to larger datasets that contain more than several thousand single cells (van der Maaten, 2014).

When we highlighted the outliers from all genes combined (Figures 2C–2E), or separately per gene (Figures S2A–S2D), we observed that they non-randomly distributed in the tSNE plots of both HeLa and hiPS cells. This suggests the existence of phenotypic state- and population context-specific mRNA-to-protein relationships in single cells. For instance, mitotic cells had more mRNA than protein for the genes *FBL* (Fibrillarin) and *LMNB1* (Lamin B1) and nuclear pore complex components (*NUP214, TPR*, *RANBP2*, and *NUP107*) when compared to interphase cells. This finding supports previous reports from bulk measurements showing that their transcription continues during mitosis while protein translation is halted (Palozola et al., 2017; Stumpf et al., 2013; Tanenbaum et al., 2015) and extends it to the single-cell level. We observed the opposite trend for genes encoding endocytic and cytoskeletal proteins (*TFRC*, *ACTB*, and *TUBA1B*), as these proteins have generally long half-lives (Boisvert et al., 2012; Tani et al., 2012), resulting in more protein than mRNA in mitotic cells (Figures S2A–S2D), when compared to interphase cells. We also found that outliers with higher protein levels than mRNA, enriched in regions of the maps representing cellular states of a duplicated genome (late S and G2 phases). When we quantified the differences in mRNA-to-protein ratios between single cells in the G1 and G2 phases of the cell cycle across all genes and compared the obtained values to measurements on ribosome occupancy along the cell cycle obtained with an entirely orthogonal, bulk approach performed in a different cell line (Tanenbaum et al., 2015), we observed a good agreement of results (Figures 2F, S2E, and S2F). Such

Figure 2. Explaining Variability in Protein Abundance and Cytoplasmic mRNA Abundance in Single Cells for Various Genes

(A) Schematic representation of features describing the cellular state, population context, and microenvironment, as well as spatial patterns of single mRNA molecules within individual cells.

(B) Scheme of indexing single cells that are outliers (1.5 standard deviations [SD]) from a linear regression fit between single-cell mRNA and protein abundance and their visualization in a two-dimensional (2D) tSNE plot. All single cells from both replicates are embedded in 2D BH-tSNE, using *Z* scored features describing the phenotypic state and microenvironment of each cell. Features used for BH-tSNE are listed in Figure S2.

(C–E) BH-tSNE embedding of all single cells of all genes measured, separated by whether the protein was detected with antibodies (C) or by fluorescent protein-tagging at endogenous loci (D and E). Blue: outliers having higher protein levels; red: outliers having higher mRNA levels than expected from the linear fit (residuals 1 SD away from the linear fit).

(F) Ratios of mRNA abundance between G1 and G2 cells divided by the ratios of the density of ribosome footprints per gene (FP) between G2 and G1 cells, obtained from the dataset published in Tanenbaum et al. (2015) or of mRNA abundance between G1 and G2 cells divided by the ratios of protein abundance between G2 and G1 cells as measured in this study.

(G) Scatterplot of single-cell mRNA and protein abundances of *PLK1*, in which each cell is colored according to its position in the cell cycle (cell cycle progression index). Line represents smoothing spline (parameter = 0.005).

(H) Plots of explained variance in single-cell protein abundance in PLK1 using its mRNA abundance as an independent variable, based on 100 bootstraps of linear model fits performed separately to all G1 and all G2 cells. Lines indicate 99.9% confidence intervals (CIs).

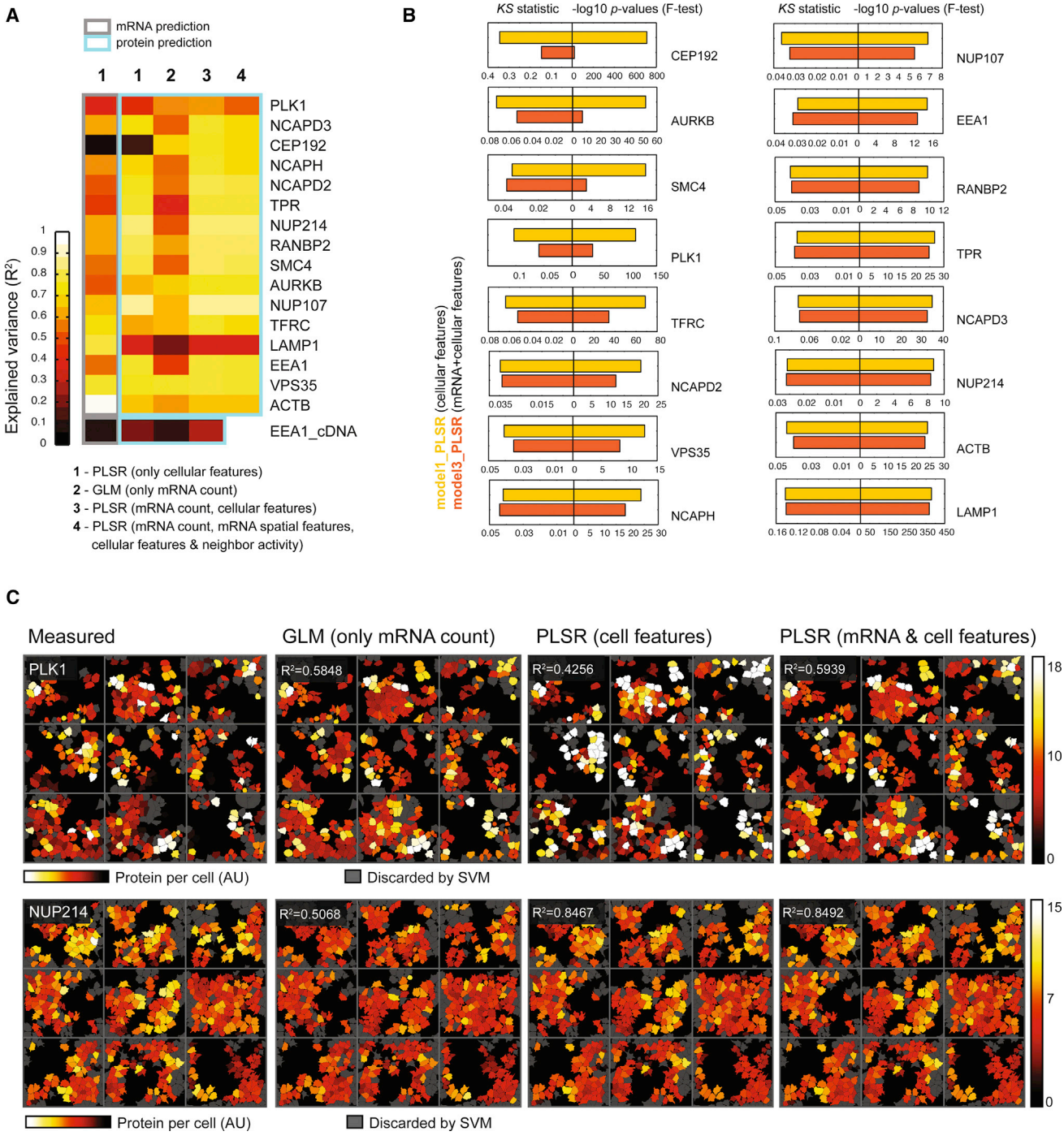


Figure 3. Multivariate Models of mRNA Abundance and Cellular State Accurately Predict Single-Cell Protein Abundance

(A) Explained variance (R^2) of single-cell mRNA abundance (gray-boxed left column) and protein abundance (light-blue boxed columns) for the genes quantified in HeLa cells using various types of (multivariate) linear models (GLM or PLSR) based on combinations of features. (1) PLSR model based on cellular features. (2) GLM model using mRNA abundance. (3) PLSR model combining cellular features (reduced to principal components) and mRNA abundance. (4) PLSR model combining cellular features, mRNA abundance, spatial features of mRNA subcellular patterning, and neighbor activity. Models were learned on one population of cells and tested on a separate population of cells from a replicate experiment.

(B) Plots of Kolmogorov-Smirnov (KS) statistic values and p values ($-\mathrm{Log}_{10}$) of F-tests comparing the performance of model 2 (PLSR models based on cellular features) and model 3 (PLSR models combining cellular features and mRNA abundance) to predict single-cell protein abundances measured in a separate

(legend continued on next page)

cell cycle-dependent differences in mRNA-to-protein ratios can also be expected for genes whose expression is rapidly induced at a specific point in the cell cycle. When we looked at the single-cell protein and mRNA abundances of *PLK1*, a central regulator of mitosis, which is induced during the late S phase (Carmena and Earnshaw, 2003; Zitouni et al., 2014), and color-coded cells according to their position in the cell cycle (Gut et al., 2015), we found that the scaling of protein with mRNA was stronger in G2 than in G1 cells (Figures 2G and 2H) (R^2_{G1} = 0.462, 99.99% CI = 0.397–0.518; R^2_{G2} = 0.608, 99.99% CI = 0.548–0.672) (see Supplemental Information and STAR Methods). This explains why the majority of single-cell outliers with more protein than mRNA relative to a linear fit calculated from the whole cell population were G2 cells. Since the distribution of cell cycle stages among outlier cells is, for each analyzed gene, highly different compared to the whole cell population (Figure S2B and Table S4 for Fisher's exact tests), this indicates that a major source of cell-to-cell variability in protein-to-mRNA relationships comes from cell cycle-controlled differences in the transcription of these genes and the translational efficiency of their mRNAs.

Multivariate Models of mRNA Abundance and Cellular State Accurately Predict Single-Cell Protein Abundance

Because single-cell outliers in mRNA-to-protein correlations distributed non-randomly in population context and phenotypic state space, we reasoned that models predicting single-cell protein abundance from mRNA measurements should perform better if they include the phenotypic and microenvironmental state of single cells. To test this hypothesis, we learned data-driven multilinear models for each gene using partial least squares regression (PLSR) on a principal component-reduced multidimensional space consisting of features describing the cellular and microenvironmental state of cells. We then tested these models by applying them to a separate population of cells from a replicate experiment. While cytoplasmic mRNA abundance in single cells is generally well predicted, as reported previously using multilinear models (Battich et al., 2015), single-cell protein abundance is even better predicted by such models (model 1) (Figures 3A, S3A, and S3B; Table S4). With a few exceptions, these models predict protein abundance in single cells better than models that use only mRNA abundance (model 1 versus 2 on protein abundance prediction). Adding mRNA abundance as a predictor to models using cellular features improves their ability to explain the measured protein abundance in single cells (model 1 versus model 3 on protein abundance prediction) as evident from F-test p values (Figure 3B), and for some genes also from the explained variance and KS statistics (Figures 3A and 3B), but the effect is often small, indicating that cellular features already contain most of the information. Notably, this did not work for a transiently expressed gene using a cDNA construct and a viral promoter (Figures 3B, S3E, and S3F; Table S4). Thus, non-endogenous gene expression does not only alter single-cell correlations between mRNA and protein abundance but also results in an inability to explain the varying mRNA-to-protein ratios by heterogeneity in the microenvironment or phenotypic state of single cells. This suggests that outliers to single-cell correlations between mRNA and protein abundance in genetically identical mammalian cells are determined by the phenotypic state or population context of single cells, which depend on endogenous mechanisms of transcriptional and posttranscriptional gene regulation.

Single-Cell mRNA-to-Protein Ratios in a Highly Dynamic Gene Induction System

While some of the genes analyzed are low abundant and show a large amount of cell-to-cell variability in their expression, they did not include typical non-steady-state conditions that apply to immediate early response genes whose expression is rapidly induced upon a stimulus and whose mRNA and protein turnover rates are particularly high (Bahrami and Drabløs, 2016; Cook et al., 1999). To investigate the behavior of non-steady-state single-cell mRNA and protein abundance over time, we performed gene induction experiments on *JUN*, an important transcription factor and immediate early response gene whose expression is induced by epidermal growth factor (EGF) (Figure 4A). Cells grown for 72 hr were serum-starved for 16 hr and subsequently stimulated with EGF to induce *JUN* gene transcription. At 7, 15, 20, 30, 40, 60, 90, and 210 min after induction, we fixed cells and performed bDNA sm-FISH and immunofluorescence staining to quantify both single-cell cytoplasmic and nuclear mRNA, as well as protein abundance of endogenous *JUN* (Figure S4A; Tables S5 and S6). We also measured nuclear mRNA abundance to quantify the lag between nuclear and cytoplasmic mRNA appearance, as well as the appearance of protein, as a function of time. This becomes relevant for the correct interpretation of single-cell correlations between mRNA and protein abundance during acute gene induction experiments, which are initially far from steady state.

Although the single-cell distributions of nuclear mRNA, cytoplasmic mRNA, and protein abundance displayed a large amount of variability (Figure S4A) and were at certain time points bimodal (Figure S4B), the time course revealed clearly distinguishable consecutive waves (Figure 4A) of mRNA and protein production. Upon gene induction, nuclear mRNA was the first to show a peak in abundance. It was followed, with a time delay of 20 min on average, by a wave of cytoplasmic mRNA abundance that was followed, with a second time delay of 50 min on average, by a wave of protein abundance. When we combined all cytoplasmic mRNA and protein measurements from all single cells and all time points and plotted these (Figure 4B), we observed an almost complete coverage of the "phase space" by single cells, showing the importance of knowing to which time point the cells belong, considering the wide variety of states such a dynamic gene induction system can have. Notably, while linear correlations between cytoplasmic mRNA and protein abundance in single cells were low at early

population of HeLa cells. Full overview of the number of principal components used for models and the various statistical tests calculated are provided in Table S4.

(C) Example of pseudo-colored single-cell segmentation images (genes: *PLK1, NUP214*) for measured values of protein abundance in each single cell, compared to those predicted with linear models using only mRNA abundance (GLM), predicted with PLSR models using only cellular features, and predicted with PLSR models combining cellular features with mRNA abundance.

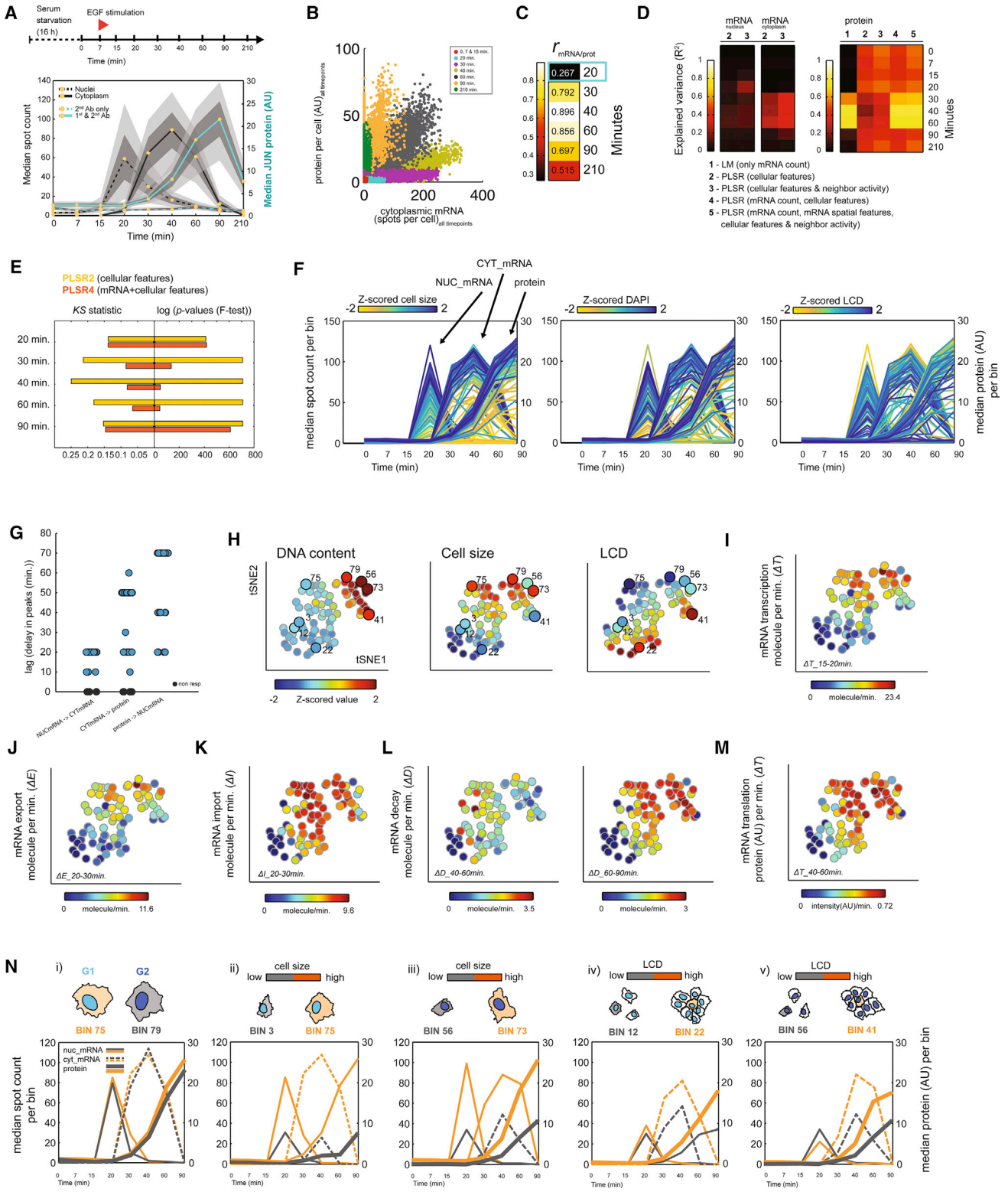


Figure 4. mRNA-Protein Relationships of *JUN* during EGF Gene Induction

(A) Expression profiles for mRNA and protein abundance of *JUN* in HeLa cells after serum starvation and addition of EGF. Colors represent 0.375 and 0.625 quantiles (dark gray) and 0.25 and 0.75 quantiles (light gray) around the median (0.5 quantile) of single-cell quantifications.

(B) Scatterplot of mRNA and protein abundance for all single cells across all time points of EGF stimulation. Time points are colored differently.

(C) Pearson's correlation coefficients of mRNA and protein abundance across single cells for the various depicted time points.

(legend continued on next page)

(0–20 min) and late (90–210 min) time points, they were high at 40 and 60 min (Figures 4C and S4C), which are the time points when cytoplasmic mRNA abundance reaches its maximum or begins to decline, and protein abundance starts to increase or approaches its maximum (Figure 4A). We moreover find that nuclear mRNA abundance is less well predicted by features of the cellular state and population context than cytoplasmic mRNA abundance (Figure 4D). This supports the notion that a delay between mRNA transcription and nuclear export can buffer unpredictable variability (Bahar Halpern et al., 2015; Battich et al., 2015; Singh and Bokes, 2012; Sturrock et al., 2017; Xiong et al., 2009) and that cytoplasmic mRNA abundance of *JUN* is a good predictor of its protein abundance in single mammalian cells at 30–60 min after gene induction (Figures 4D and 4E).

We next asked whether variation in synthesis, nuclear export, and degradation of mRNA, as well as synthesis and degradation of protein, are dependent on the phenotypic state or population context of a cell. We first inspected whether cellular features change during the time of EGF-mediated gene induction, which was not the case except at 210 min after gene induction. Subsequently, we clustered single cells from all time points in which no phenotypic change upon EGF exposure was observed (up to 90 min), based on similarities in their multivariate set of phenotypic states (DNA content, nucleus size, and cell size) and population context features (local cell density) (Figures S5A–S5E), using self-organizing maps (SOMs), which apply neural network-based unsupervised machine learning for single-cell clustering and dimensionality reduction, recently developed for complex multidimensional data originating from flow cytometry (Van Gassen et al., 2015). A Kolmogorov-Smirnov-based nonparametric test revealed that the distributions of the median feature values across the SOM clusters approximate the feature value distributions across all single cells, which was not the case for medians of randomly grouped single cells (Figures S5C–S5E). Since the clusters contained cells from all time points, this allowed us to compare the dynamics of nuclear and cytoplasmic mRNA counts and protein abundance between clusters of cells with different phenotypic and microenvironmental properties (Figure 4F). To analyze these dynamics in more detail, we performed cross-correlation analysis between nuclear mRNA, cytoplasmic mRNA, and protein abundance over time for each of the groups separately (Figure S5F), and we calculated the delays associated with the observed dynamics (Figure S5H). While cross correlations were very high within each of the subgroups of cells (Figure S5F), the lag times at which maximum correlation between the abundance of all three molecular species was found, varied, as did the time differences (termed deltas) between their first appearances, the peaks, and the disappearances (Figures 4G, S5G, and S5H). To visualize whether this variation is accompanied by variation in specific cellular properties, we first embedded all the SOM clusters within a 2D tSNE space, based on their median values for DNA content, nuclear size, cell size, and local cell density (Figure 4H). When we projected the values of those 4 features onto the tSNE plot (Figures 4H–4M), we observed that cells that are large, in the G2 phase of the cell cycle, and are growing in sparsely populated regions have much higher mRNA transcription and protein translation rates, resulting in higher levels of nuclear mRNA, cytoplasmic mRNA, and protein than cells that are small and grow in densely populated regions (Figures 4N and S5I).

When we compared the median increase of mRNA and protein in the different clusters over the gene induction time course, we observed that adaptations in gene expression output can occur at the transcriptional and posttranscriptional level (Figures 4Ni–4Nv). First, we compared single-cell clusters that were similar in all quantified cellular features except in DNA content (bin 75 versus 79) and did not observe any adaptation (Figure 4Ni). Thus, G1 cells that are as large as G2 cells and experience the same local cell density transcribe similar amounts of *JUN* mRNA and synthesize similar amounts of protein. Next, we compared cells that experience a similar amount of local cell

(D) Left: Explained variance (R^2) in single-cell nuclear and cytoplasmic mRNA abundance of *JUN* using multivariate linear models (PLSR) based on cellular features (model 2) and cellular features plus neighbor activity features (model 3) reduced to principal components. Models are tested on cells from a replicate experiment that were stimulated with EGF for a same period of time. Right: Explained variance (R^2) in single-cell protein abundance of JUN based on linear models using only cytoplasmic mRNA abundance (model 1), based on PLSR models using cellular features (model 2), using cellular features and neighbor activity (model 3), using cellular features and mRNA abundance (model 4), and using cellular features, neighbor activity, mRNA abundance, and spatial subcellular patterning features of mRNA (model 5) as predictors.

(E) Plots of Kolmogorov-Smirnov (KS) statistic values and p values ($-\mathrm{Log}_{10}$) of F-tests comparing the performance of PLSR model 2 (based on cellular features) and model 3 (based on cellular features and mRNA abundance) to predict JUN protein abundances in single cells from a replicate experiment that were stimulated with EGF for a same period of time.

(F) Median expression profiles of inferred *JUN* stimulation trajectories obtained after clustering single cells from the various time points based on their similarity in their cellular features and local cell density (population context) using self-organizing maps (SOMs) (see Figure S5). Color map represents mean-normalized values of DNA content, cell size, and local cell density (LCD), calculated as the median of all the cells that belong to a particular cluster.

(G) Time delays between peaks in nuclear and cytoplasmic mRNA abundance, between peaks in protein and cytoplasmic mRNA abundance, and between peaks in protein and nuclear mRNA abundance in all SOM clusters (see Figure S5). Black color indicates SOM clusters of cells that were non-responders.

(H) BH-tSNE embedding of SOM clusters based on *Z* scored median values of cellular and population context (LCD) features. Color map represents *Z* scored median values for DNA content, cell size, and LCD, calculated for each of the SOM clusters.

(I–M) Projection on tSNE-embedded SOM clusters of transcription rates (from deltas between nuclear mRNA counts at 15 and 20 min after EGF stimulation) (I), mRNA nuclear export rates (from deltas between nuclear mRNA counts at 20 and 30 min after EGF stimulation) (J), mRNA cytoplasmic import rates (from deltas between cytoplasmic mRNA counts at 20 and 30 min after EGF stimulation) (K), cytoplasmic mRNA decay rates (from deltas between cytoplasmic mRNA counts at 40 and 60 min as well as at 60–90 min after EGF stimulation) (L), and of protein translation rates (from deltas between protein abundance at 40 and 60 min after EGF stimulation) (M). Deltas (spot count for mRNA and AU for protein intensities) were divided by the time (in min) between the respective time points to obtain rates.

(N) Comparison of changes of median values of nuclear and cytoplasmic mRNA abundance and protein abundance between pairs of SOM clusters highlighted in (H) that contain cells from a different cycle stage (G1 and G2) but have the same size and LCD (i), cells that are all in G1 and have a similar LCD but different cell size (ii), cells that are all in G2 and have a similar LCD but different cell size (iii), cells that are all in G1 and have the same size but a different LCD (iv), or cells that are all in G2 and have the same size but a different LCD (v). Bin number indicated corresponds to the cluster number in (H).

density but are different in cell size (bins 3 and 56 versus 75 and 73). We observed a strong difference in nuclear and cytoplasmic mRNA abundance and in protein levels (Figures 4Nii and 4Niii). This indicates that adaptation of *JUN* expression to cell size occurs primarily at the transcriptional level, which may be part of a general mechanism to assure constant mRNA concentration as cells increase in volume (Battich et al., 2015; Padovan-Merhar et al., 2015). However, when we compared single-cell clusters with similar cell size but experiencing different local cell densities (bins 12 versus 22 and 56 versus 41), we observed a different and unexpected type of adaptation. While the transcriptional output was similar in these cells, as expected given their similar sizes, their cytoplasmic mRNA abundance was different, and so was their protein translational output. This indicates that cells experiencing a high local cell density degrade their cytoplasmic mRNA slower and produce more protein than cells experiencing low local cell density (Figures 4Niv and 4Nv).

Together, these results illustrate that *JUN* gene expression dynamics are highly variable between individual cells, relating to specific properties of the phenotypic state and microenvironment. While we demonstrated some of these effects specifically by comparing single-cell clusters that are similar in all except one of the quantified differences, they usually occur simultaneously, which can lead to non-intuitive outcomes. For instance, cells that experience high local cell density are normally also smaller than cells experiencing low local cell density. While the smaller size may lead to a lower transcriptional output in the nucleus, the higher cell density may at the same time reduce mRNA degradation in the cytoplasm and therefore prevent a concomitant reduction in translational output (e.g., compare bin 73 in Figure 4Niii with bin 41 in Figure 4Nv). This results in cells with highly varying transcriptional outputs but similar protein output, suggesting an intricate regulation of mRNA-to-protein relationships in single cells.

Cell-to-Cell Variability in Subcellular Localization of mRNAs

To investigate these non-intuitive adaptations of *JUN* gene expression further, we examined the intracellular localization of *JUN* transcripts. We observed that *JUN* transcripts display different cytoplasmic patterns over time. 30 min after gene induction, *JUN* mRNAs appear more clustered, whereas at 60 min after induction, they are found to be more dispersed in the cytoplasm (Figure 5A). In contrast, the cytoplasmic transcripts of *HPRT1*, a constitutively expressed housekeeping gene, remained unaltered during the course of the gene induction experiment (Figures S5J and S5K). To systematically quantify the spatial reorganization of *JUN* transcripts in an unbiased manner, we performed unsupervised clustering of single cells based on spatial features extracted from the mRNA spots as previously published (Battich et al., 2013). This analysis revealed three principal classes of spatial patterning of *JUN* transcripts in the cytoplasm of cells, namely a perinuclear non-clustered (class 1), a peripherally clustered (class 2), and a peripherally non-clustered (class 3) pattern (Figure 5B). Generally, class 1 and class 3 patterns are less frequently observed than class 2 patterns. Class 1 patterns occur primarily at earlier time points after gene induction (30 min), while class 3 patterns occur mostly at later time points (60 min) (Figure 5C).

When we quantified the propensity for spatial patterns for different single-cell clusters depending on the phenotypic state and microenvironmental features of cells, we noticed that class 1 patterns are seen specifically in large cells experiencing low local cell density, while class 3 patterns are specifically seen in small cells experiencing high local cell density (Figures 5D–5G) These differences in patterning corresponded to the cells' ability to maintain a similar cytoplasmic mRNA abundance despite highly different transcriptional outputs (Figures 5H and 5I). In large cells experiencing low cell density (bin 73), which have a large transcriptional output, the *JUN* transcripts quickly adopt a class 2 pattern, which persists until cytoplasmic mRNA is degraded (Figure 5I). However, in cells that are small and experience high local cell density (bin 41), the class 2 pattern is less frequently observed, and *JUN* transcripts adopt a class 3 pattern, in which propensity increases over the duration of gene induction (Figure 5I). This may reflect the existence of a specific mechanism in small cells in densely populated areas to protect *JUN* transcripts from degradation, allowing these cells to have similar cytoplasmic mRNA content despite their lower transcriptional output, and concomitantly to have a similar protein translational output. In summary, these results show that also cytoplasmic patterning of mRNA varies in a non-random fashion between genetically identical cells, which for *JUN* may reflect a posttranscriptional mechanism of gene expression adaptation to the cellular microenvironment.

DISCUSSION

In this work, we used a set of recently made available human cell lines in which specific genes are endogenously tagged with a fluorescent protein epitope (Cai et al., 2018; Koch et al., 2018; Politi et al., 2018; Roberts et al., 2017). We combined imaging of protein with bDNA-based single-molecule FISH to obtain sensitive and reproducible readouts of both mRNA and protein abundance of the same gene in thousands of single cells. This was complemented with quantitative immunofluorescence for untagged genes. These quantifications show that single-cell correlations between the steady-state abundance of mRNA and protein of the same gene are generally high in adherent mammalian cells. How can this be different from previous findings in bacteria (Taniguchi et al., 2010)? Besides possible technical explanations, we studied adherent mammalian cells that do not grow in isolation but are part of a growing cell population. Through collective or population context effects, these cell populations display emergent properties that strongly influence gene expression, position in the cell cycle, and the phenotypic state of single cells. This represents a major source of cell-to-cell variability that affects both mRNA and protein abundance, which may be different in liquid cultures of bacteria, where individual cells are forced to stay in isolation. In addition, mRNA and protein turnovers are slower in mammalian cells compared to bacteria, which can have a strong impact on single-cell correlations (Shamir et al., 2016). Moreover, mammalian cells have a nucleus and can display considerable nuclear retention of transcripts, which can passively filter noise in transcription (Bahar Halpern et al., 2015; Battich et al., 2015; Singh and Bokes, 2012; Sturrock et al., 2017; Xiong et al., 2009) and have active mechanisms of noise filtering through feedback loops (Alon, 2007). As a result,

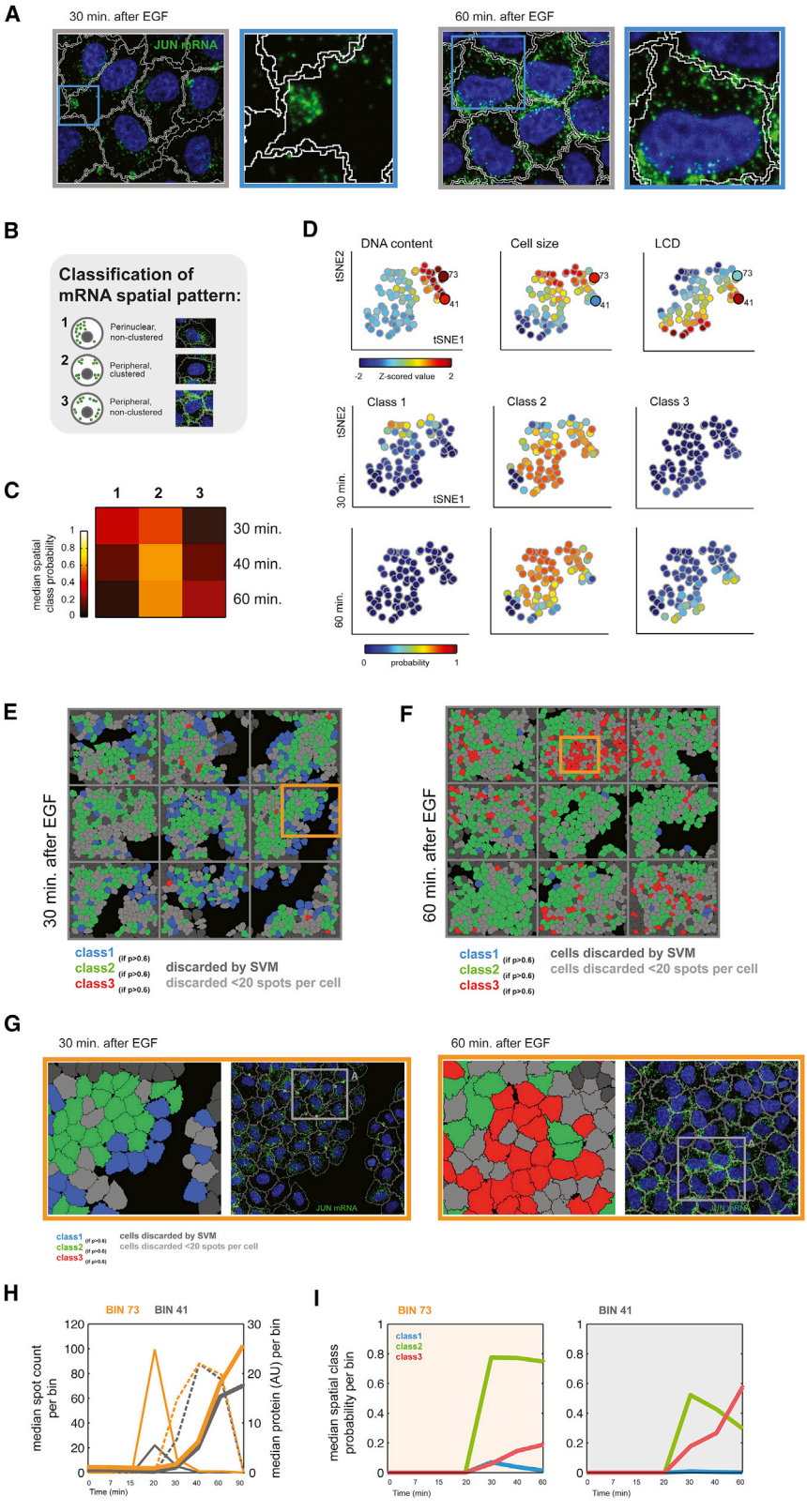


Figure 5. Cell-to-Cell Variability in Subcellular Localization of mRNA

(A) Images of observed spatial patterns of *JUN* cytoplasmic mRNA. Green signal belongs to the bDNA smFISH-labeled single mRNA molecules of *JUN*; blue signal represents DNA stain.

(B) Three distinct classes of intracellular spatial patterns observed for *JUN* mRNA, based on single-cell clustering of all single cells from all time points of the EGF induction experiment using single-cell features of spatial patterning of mRNA spots.

(C) Median values of single-cell classification probabilities to the three classes for all cells of a particular time point after EGF induction.

(D) BH-tSNE embedded SOM clusters, colored for median values of DNA content, cell size, and LCD (top, as in Figure 4) and colored for median probability of all cells in each cluster to belong to the three spatial pattern classes during the time of appearance of mRNA in the cytoplasm (30 min) and during its decay (60 min).

(E and F) Pseudo-colored single-cell segmentation images of *JUN* mRNA according to the highest probability of each single cell for a particular spatial pattern class at 30 min (E) and 60 min (F) after addition of EGF. Colors represent discretized probability of 1 to belong to a particular spatial class, based on a threshold of 0.6. Dark gray cells were discarded by SVM (image border or badly segmented cells), and light gray cells were discarded due to having less than 20 spots per cell, which prevents the calculation of spatial patterns.

(G) Zoomed insets from (E) and (F), with their corresponding original bDNA smFISH image. Green: bDNA smFISH labeling of *JUN* mRNA; blue: DAPI.

(H) Changes in median nuclear and cytoplasmic mRNA and protein abundance of cells that belong to the indicated bin (SOM cluster), which both contain G2 cells but of different size and local cell density (LCD). Bin 73 contains large cells at low LCD and bin 41 small cells at high local cell density. Bin numbers are also indicated in the tSNE plot of SOM clusters in (D).

(I) Median probability of cells belonging to the three mRNA subcellular spatial pattern classes for the two bins from (G) during EGF gene induction.

transcript abundance in the cytoplasm of a single cell is more predictable by its phenotypic state and population context than in the nucleus, which therefore serves as a better predictor of protein abundance, resulting in higher single-cell correlations. Finally, we observed that single-cell correlations between endogenous levels of mRNA and protein are at least 2-fold higher than single-cell correlations obtained from a transiently expressed cDNA with a non-endogenous promoter. This underscores the important role of endogenous transcriptional and posttranscriptional gene regulation in controlling mRNA and protein variability in single cells. In addition, the approach applied here cannot avoid technical noise in single-cell measurements, implying that real correlations may even be higher.

Nevertheless, our measurements do indicate the existence of a considerable fraction of single cells that do not fit a simple linear relationship between mRNA and protein abundance. A multivariate analysis of these cells revealed that these outliers display specific phenotypic properties or experience a specific population context. Such outliers are therefore likely not the result of uncontrolled variation in gene expression but of regulatory mechanisms that differentially adapt the transcriptional or translational output of a single cell to its size, position in the cell cycle, or its microenvironment, as reported previously for bulk and single-cell studies (Jovanovic et al., 2015; Padovan-Merhar et al., 2015; Tanenbaum et al., 2015; Zopf et al., 2013). This holds true for constitutively expressed genes, whose protein-to-mRNA ratio is lower in G1 and higher in G2 cells (Tanenbaum et al., 2015), for genes acutely expressed at specific points of the cell cycle where they show sudden alterations of this ratio, and especially when genes are rapidly induced and show a fast turnover during gene induction experiments (Jovanovic et al., 2015). A more detailed analysis of the latter scenario for *JUN*, an early response transcription factor that induces cell proliferation, provided evidence that such single-cell adaptations can specifically occur at multiple steps of gene expression, be it the duration and total output of gene transcription, nuclear retention and export of mRNA, cytoplasmic mRNA stability, or duration and total output of protein translation. This not only underscores the complexity of transcriptional and posttranscriptional gene regulation mechanisms acting in single cells (Vera et al., 2016) but shows that this complexity is adapted to the multiple and sometimes opposing physiological needs of a single cell, which can result in non-intuitive outcomes. For *JUN*, we moreover reveal that its cytoplasmic transcripts adopt a clustered pattern at early time points of gene induction, which progress into a non-clustered localization pattern at later time points of gene induction, specifically in cells that are small and experience high local cell density. This alteration in subcellular patterning may allow small densely populated cells to have a similar translational output as large cells in sparsely populated regions despite having a much lower transcriptional output. While the molecular mechanisms and causality underlying this phenomenon remain to be investigated, it fits the general notion that the stability of transcripts can be regulated by altering their subcellular localization (Buxbaum et al., 2015; Martin and Ephrussi, 2009; Moor and Itzkovitz, 2017). We therefore conclude that to derive meaningful interpretations from quantitative analyses of mRNA-protein relationships in single cells, these measurements must be performed on a gene-by-gene basis and must take into account the multivariate information of the phenotypic state of a single cell and its microenvironment, as well as the subcellular patterning of transcripts.

STAR★METHODS

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

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell Lines Source and Cultivation
- METHOD DETAILS
  - Experimental Design Details
  - Labelling of Cells with EdU
  - Image-Based Transcriptomics and Antibody Labelling
  - Measurement and Quantification of Nuclear Transcripts
  - EGF Stimulation Experiment
  - Transient Transfection of cDNA
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Feature Extraction
  - Background Subtraction for mEGFP Tagged HeLa Cells
  - Selection of Genes for the Transcript and Protein Quantification
  - Generation of Components for Transcript and Protein Abundance Prediction
  - Cell Cycle Significance Calculation
  - Embedding of Single Cells in Two Dimensional bh-tSNE Feature Space
  - Comparison of Distributions Using Kolmogorov-Smirnov Statistics and F-Test
  - Classification of Cells Based on the Transcript Spatial Pattern
  - Binning of Cells Based on Cellular Features
  - Analysis of Signal Cross-Correlation

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and six tables and can be found with this article online at https://doi.org/10.1016/j.cels.2018.09.001.

ACKNOWLEDGMENTS

We thank the members of Pelkmans lab for discussions and René Holtackers for technical support. D.P. is supported by an EMBO (ALTF 1484-2015) and HFSP (LT-000531/2016) Long-Term Fellowship. L.P. is supported by the Swiss National Science Foundation and the University of Zurich. Research in the J.E. lab is funded by the European Molecular Biology laboratory (B.K., M.K., J.E.) and by grants from EU_FP7_MitoSys (Grant Agreement 241548), EU-H2020-iNEXT (Grant Agreement 653706) and by the National Institutes of Health 4D Nucleome (4DN) Common Fund (U01 EB021223/U01 DA047728).

AUTHOR CONTRIBUTIONS

L.P. and D.P. conceived the project; B.K., M.K., and J.E. created and provided endogenously tagged HeLa cell lines; D.P. performed all experiments and analyzed all data; L.P. and D.P. wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 25, 2018
Revised: June 28, 2018
Accepted: September 5, 2018
Published: October 17, 2018

REFERENCES

Albayrak, C., Jordi, C.A., Zechner, C., Lin, J., Bichsel, C.A., Khammash, M., and Tay, S. (2016). Digital quantification of proteins and mRNA in single mammalian cells. Mol. Cell *61*, 914–924.

Alon, U. (2007). Network motifs: theory and experimental approaches. Nat. Rev. Genet. *8*, 450–461.

Bahar Halpern, K., Tanami, S., Landen, S., Chapal, M., Szlak, L., Hutzler, A., Nizhberg, A., and Itzkovitz, S. (2015). Bursty gene expression in the intact mammalian liver. Mol. Cell *58*, 147–156.

Bahrami, S., and Drabløs, F. (2016). Gene regulation in the immediate-early response process. Adv. Biol. Regul. *62*, 37–49.

Battich, N., Stoeger, T., and Pelkmans, L. (2013). Image-based transcriptomics in thousands of single human cells at single-molecule resolution. Nat. Methods *10*, 1127–1133.

Battich, N., Stoeger, T., and Pelkmans, L. (2015). Control of transcript variability in single mammalian cells. Cell *163*, 1596–1610.

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.

Buxbaum, A.R., Haimovich, G., and Singer, R.H. (2015). In the right place at the right time: visualizing and understanding mRNA localization. Nat. Rev. Mol. Cell Biol. *16*, 95–109.

Cai, Y., Hossain, M.J., Hériché, J.-K., Politi, A.Z., Walther, N., Koch, B., Wachsmuth, M., Nijmeijer, B., Kueblbeck, M., Martinic-Kavur, M., et al. (2018). Experimental and computational framework for a dynamic protein atlas of human cell division. Nature *561*, 411–415.

Carmena, M., and Earnshaw, W.C. (2003). The cellular geography of aurora kinases. Nat. Rev. Mol. Cell Biol. *4*, 842–854.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. *7*, R100.

Cook, S.J., Aziz, N., and McMahon, M. (1999). The repertoire of fos and jun proteins expressed during the G1 phase of the cell cycle is determined by the duration of mitogen-activated protein kinase activation. Mol. Cell. Biol. *19*, 330–341.

Csárdi, G., Franks, A., Choi, D.S., Airoldi, E.M., and Drummond, D.A. (2015). Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. PLoS Genet. *11*, e1005206.

Darmanis, S., Gallant, C.J., Marinescu, V.D., Niklasson, M., Segerman, A., Flamourakis, G., Fredriksson, S., Assarsson, E., Lundberg, M., Nelander, S., et al. (2016). Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells. Cell Rep. *14*, 380–389.

Edfors, F., Danielsson, F., Hallström, B.M., Käll, L., Lundberg, E., Pontén, F., Forsström, B., and Uhlén, M. (2016). Gene-specific correlation of RNA and protein levels in human cells and tissues. Mol. Syst. Biol. *12*, 883.

Fortelny, N., Overall, C.M., Pavlidis, P., and Freue, G.V.C. (2017). Can we predict protein from mRNA levels? Nature *547*, E19–E20.

Genshaft, A.S., Li, S., Gallant, C.J., Darmanis, S., Prakadan, S.M., Ziegler, C.G.K., Lundberg, M., Fredriksson, S., Hong, J., Regev, A., et al. (2016). Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. Genome Biol. *17*, 188.

Golding, I., Paulsson, J., Zawilski, S.M., and Cox, E.C. (2005). Real-time kinetics of gene activity in individual bacteria. Cell *123*, 1025–1036.

Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. *4*, 117.

Gut, G., Tadmor, M.D., Pe'er, D., Pelkmans, L., and Liberali, P. (2015). Trajectories of cell-cycle progression from fixed cell populations. Nat. Methods *12*, 951–954.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science *292*, 929–934.

Jovanovic, M., Rooney, M.S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E.H., Fields, A.P., Schwartz, S., Raychowdhury, R., et al. (2015). Immunogenetics. dynamic profiling of the protein life cycle in response to pathogens. Science *347*, 1259038.

Kamentsky, L., Jones, T.R., Fraser, A., Bray, M.A., Logan, D.J., Madden, K.L., Ljosa, V., Rueden, C., Eliceiri, K.W., and Carpenter, A.E. (2011). Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software. Bioinformatics *27*, 1179–1180.

Koch, B., Nijmeijer, B., Kueblbeck, M., Cai, Y., Walther, N., and Ellenberg, J. (2018). Generation and validation of homozygous fluorescent knock-in cells using CRISPR-Cas9 genome editing. Nat. Protoc. *13*, 1465–1487.

Lawe, D.C., Patki, V., Heller-Harrison, R., Lambright, D., and Corvera, S. (2000). The FYVE domain of early endosome antigen 1 is required for both phosphatidylinositol 3-phosphate and Rab5 binding. critical role of this dual interaction for endosomal localization. J. Biol. Chem. *275*, 3699–3705.

Lawless, C., Holman, S.W., Brownridge, P., Lanthaler, K., Harman, V.M., Watkins, R., Hammond, D.E., Miller, R.L., Sims, P.F.G., Grant, C.M., et al. (2016). Direct and absolute quantification of over 1800 yeast proteins via selected reaction monitoring. Mol. Cell. Proteomics *15*, 1309–1322.

Li, J.J., Bickel, P.J., and Biggin, M.D. (2014). System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ *2*, e270.

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

Lo, C.A., Kays, I., Emran, F., Lin, T.J., Cvetkovska, V., and Chen, B.E. (2015). Quantification of protein levels in single living cells. Cell Rep. *13*, 2634–2644.

Mahen, R., Koch, B., Wachsmuth, M., Politi, A.Z., Perez-Gonzalez, A., Mergenthaler, J., Cai, Y., and Ellenberg, J. (2014). Comparative assessment of fluorescent transgene methods for quantitative imaging in human cells. Mol. Biol. Cell *25*, 3610–3618.

Martin, K.C., and Ephrussi, A. (2009). mRNA localization: gene expression in the spatial dimension. Cell *136*, 719–730.

Moor, A.E., and Itzkovitz, S. (2017). Spatial transcriptomics: paving the way for tissue-level systems biology. Curr. Opin. Biotechnol. *46*, 126–133.

Otsuka, S., Bui, K.H., Schorb, M., Hossain, M.J., Politi, A.Z., Koch, B., Eltsov, M., Beck, M., and Ellenberg, J. (2016). Nuclear pore assembly proceeds by an inside-out extrusion of the nuclear envelope. ELife. *5*.

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.

Palozola, K.C., Donahue, G., Liu, H., Grant, G.R., Becker, J.S., Cote, A., Yu, H., Raj, A., and Zaret, K.S. (2017). Mitotic transcription and waves of gene reactivation during mitotic exit. Science *358*, 119–122.

Peshkin, L., Wühr, M., Pearl, E., Haas, W., Freeman, R.M., Gerhart, J.C., Klein, A.M., Horb, M., Gygi, S.P., and Kirschner, M.W. (2015). On the relationship of protein and mRNA dynamics in vertebrate embryonic development. Dev. Cell *35*, 383–394.

Peterson, V.M., Zhang, K.X., Kumar, N., Wong, J., Li, L., Wilson, D.C., Moore, R., McClanahan, T.K., Sadekova, S., and Klappenbach, J.A. (2017). Multiplexed quantification of proteins and transcripts in single cells. Nat. Biotechnol. *35*, 936–939.

Politi, A.Z., Cai, Y., Walther, N., Hossain, M.J., Koch, B., Wachsmuth, M., and Ellenberg, J. (2018). Quantitative mapping of fluorescently tagged cellular proteins using FCS-calibrated four-dimensional imaging. Nat. Protoc. *13*, 1445–1464.

Rämö, P., Sacher, R., Snijder, B., Begemann, B., and Pelkmans, L. (2009). CellClassifier: supervised learning of cellular phenotypes. Bioinformatics *25*, 3028–3030.

Reichmann, J., Nijmeijer, B., Hossain, M.J., Eguren, M., Schneider, I., Politi, A.Z., Roberti, M.J., Hufnagel, L., Hiiragi, T., and Ellenberg, J. (2018). Dual-spindle formation in zygotes keeps parental genomes apart in early mammalian embryos. Science *361*, 189–193.

Roberts, B., Haupt, A., Tucker, A., Grancharova, T., Arakaki, J., Fuqua, M.A., Nelson, A., Hookway, C., Ludmann, S.A., Mueller, I.A., et al. (2017). Systematic gene tagging using CRISPR/Cas9 in human stem cells to illuminate cell organization. Mol. Biol. Cell *28*, 2854–2874.

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.

Shamir, M., Bar-On, Y., Phillips, R., and Milo, R. (2016). SnapShot: timescales in cell biology. Cell *164*, 1302–1302.e1.

Singh, A., and Bokes, P. (2012). Consequences of mRNA transport on stochastic variability in protein levels. Biophys. J. *103*, 1087–1096.

Snijder, B., Sacher, R., Rämö, P., Damm, E.M., Liberali, P., and Pelkmans, L. (2009). Population context determines cell-to-cell variability in endocytosis and virus infection. Nature *461*, 520–523.

Snijder, B., Sacher, R., Rämö, P., Liberali, P., Mench, K., Wolfrum, N., Burleigh, L., Scott, C.C., Verheije, M.H., Mercer, J., et al. (2012). Single-cell analysis of population context advances RNAi screening at multiple levels. Mol. Syst. Biol. *8*, 579.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods *14*, 865–868.

Stoeger, T., Battich, N., Herrmann, M.D., Yakimovich, Y., and Pelkmans, L. (2015). Computer vision for image-based transcriptomics. Methods *85*, 44–53.

Stumpf, C.R., Moreno, M.V., Olshen, A.B., Taylor, B.S., and Ruggero, D. (2013). The translational landscape of the mammalian cell cycle. Mol. Cell *52*, 574–582.

Sturrock, M., Li, S., and Shahrezaei, V. (2017). The influence of nuclear compartmentalisation on stochastic dynamics of self-repressing gene expression. J. Theor. Biol. *424*, 55–72.

Tanenbaum, M.E., Stern-Ginossar, N., Weissman, J.S., and Vale, R.D. (2015). Regulation of mRNA translation during mitosis. Elife *4*, https://doi.org/10.7554/eLife.07957.

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.

Taniguchi, Y., Choi, P.J., Li, G.W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X.S. (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science *329*, 533–538.

van der Maaten, H. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

van der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. *15*, 3221–3245.

Van Gassen, S., Callebaut, B., Van Helden, M.J., Lambrecht, B.N., Demeester, P., Dhaene, T., and Saeys, Y. (2015). FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. Cytometry A *87*, 636–645.

Vera, M., Biswas, J., Senecal, A., Singer, R.H., and Park, H.Y. (2016). Single-cell and single-molecule analysis of gene expression regulation. Annu. Rev. Genet. *50*, 267–291.

Wachsmuth, M., Conrad, C., Bulkescher, J., Koch, B., Mahen, R., Isokane, M., Pepperkok, R., and Ellenberg, J. (2015). High-throughput fluorescence correlation spectroscopy enables analysis of proteome dynamics in living cells. Nat. Biotechnol. *33*, 384–389.

Walther, N., Hossain, M.J., Politi, A.Z., Koch, B., Kueblbeck, M., Ødegård-Fougner, Ø., Lampe, M., and Ellenberg, J. (2018). A quantitative map of human condensins provides new insights into mitotic chromosome architecture. J. Cell Biol. *217*, 2309–2328.

Washburn, M.P., Koller, A., Oshiro, G., Ulaszek, R.R., Plouffe, D., Deciu, C., Winzeler, E., and Yates, J.R. (2003). Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U S A *100*, 3107–3112.

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al. (2014). Mass-spectrometry-based draft of the human proteome. Nature *509*, 582–587.

Xiong, L.P., Ma, Y.Q., and Tang, L.H. (2009). Attenuation of transcriptional bursting in mRNA transport. Phys. Biol. *7*, 016005.

Zitouni, S., Nabais, C., Jana, S.C., Guerrero, A., and Bettencourt-Dias, M. (2014). Polo-like kinases: structural variations lead to multiple functions. Nat. Rev. Mol. Cell Biol. *15*, 433–452.

Zopf, C.J., Quinn, K., Zeidman, J., and Maheshri, N. (2013). Cell-cycle dependence of transcription dominates noise in gene expression. PLoS Comput. Biol. *9*, e1003161.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
c-Jun rabbit monoclonal antibody (C.238.2)	ThermoFisher	Cat #MA5-15172; RRID: AB_10979794
VPS35 goat polyclonal antibody	Novus Biologicals	Cat #IMG 3575; RRID:AB_614351
Purified mouse anti-human EEA1 antibody clone 14/EEA1 (RUO)	BD Transduction Laboratories	Cat #610456; RRID:AB_397829
Purified mouse anti-human CD71 antibody clone M-A712 (RUO)	BD Transduction Laboratories	Cat #555534; RRID:AB_395918
Purified mouse anti-human CD107a antibody clone H4A3 (RUO)	BD Transduction Laboratories	Cat #555798; RRID:AB_396132
Actin mouse monoclonal antibody ab3280 [ACTN05 (C4)]	Abcam	Cat #ab3280; RRID:AB_303668
Chemicals, Peptides, and Recombinant Proteins		
Probe against TFRC (human) type1	ThermoFisher	VA1-12208-VC,#VX-06
Probe against LAMP1 (human) type1	ThermoFisher	VA1-12096-VC,#VX-06
Probe against EEA1 (human) type1	ThermoFisher	VA1-12193-VC,#VX-06
Probe against VPS35 (human) type1	ThermoFisher	VB1-3034829-VC,#VX-01
Probe against ACTB (human) type1	ThermoFisher	VA1-10351-VC,#VX-06
Probe against EGFP type1 (previously available from Affymetrix)	ThermoFisher	VF1-10141
Probe against JUN (human) type1	ThermoFisher	VA1-12265-VC,#VX-06
Probe against HPRT1 (human) type1	ThermoFisher	VA1-11124-VC,#VX-06
Probe against dapB (*E Coli* K12) type 1 (previously available from Affymetrix against *E Coli* K12, currently against Baccilus Subtilis)	ThermoFisher	VF1-11712-VC,#VX-06
Epidermal growth factor (EGF), Human recombinant	Millipore	Cat #01-107
Alexa Fluor 647 NHS Ester (Succinimidyl Ester)	ThermoFisher	Cat A20106
GeneJuice Transfection Reagent	Millipore	Cat #70967
Bovine Serum Albumin, protease free, suitable for hybridization	Sigma Aldrich	Cat #B4287
Critical Commercial Assays		
Click-iT EdU Alexa Fluor 647 Imaging Kit	ThermoFisher	Cat #C10340
ViewRNA ISH Tissue Assay Kit	ThermoFisher	Cat #QVT0050
ViewRNA Chromogenic Signal Amplification Kit (1-plex)	ThermoFisher	Cat #QVT0201
96 well microplates, flat bottom, clear, black	Greiner	Cat #655090
Deposited Data		
Table S1	This manuscript	https://doi.org/10.17632/5nwct8htjt.2
Table S2	This manuscript	https://doi.org/10.17632/5nwct8htjt.2
Table S3	This manuscript	https://doi.org/10.17632/5nwct8htjt.2
Table S4	This manuscript	https://doi.org/10.17632/5nwct8htjt.2
Table S5	This manuscript	https://doi.org/10.17632/5nwct8htjt.2
Table S6	This manuscript	https://doi.org/10.17632/5nwct8htjt.2
Camera linearity measurements	This manuscript	https://doi.org/10.17632/5nwct8htjt.2
Experimental Models: Cell Lines		
Paxillin hiPSC line	AllenCell	Cat #AICS-0005
Sec61 translocon beta subunit hiPSC line	AllenCell	Cat #AICS-0010

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Alpha tubulin hiPSC line	AllenCell	Cat #AICS-0012
Nuclear Lamin B1 hiPSC line	AllenCell	Cat #AICS-0013
Fibrillarin hiPSC line	AllenCell	Cat #AICS-0014
Actin beta hiPSC line	AllenCell	Cat #AICS-0016
Desmoplakin hiPSC line	AllenCell	Cat #AICS-0017
Tight junction protein ZO1 hiPSC line	AllenCell	Cat #AICS-0023
Myosin heavy chain 10 hiPSC line	AllenCell	Cat #AICS-0024
HeLa cell line (parental)	Ellenberg lab (EMBL)	(Mahen et al., 2014)
ZFN PLK1-mEGFP #24 HeLa line	Ellenberg lab (EMBL)	(Cai et al., 2018)
CRISPR NCAPD3-mEGFP #16 HeLa line	Ellenberg lab (EMBL)	(Walther et al., 2018)
ZFN CEP192-mEGFP #15 HeLa line	Ellenberg lab (EMBL)	(Cai et al., 2018)
ZFN mEGFP-NCAPH #9 HeLa line	Ellenberg lab (EMBL)	(Walther et al., 2018)
CRISPR NCAPD2-mEGFP #272-78 HeLa line	Ellenberg lab (EMBL)	(Walther et al., 2018)
CRISPR mEGFP-NCAPH2 #1 HeLa line	Ellenberg lab (EMBL)	(Walther et al., 2018)
CRISPR TPR-mEGFP #171 HeLa line	Ellenberg lab (EMBL)	(Cai et al., 2018)
CRISPR mEGFP-NUP214 #2-12 HeLa line	Ellenberg lab (EMBL)	(Cai et al., 2018)
CRISPR mEGFP-RANBP2/NUP358 #97 HeLa line	Ellenberg lab (EMBL)	(Koch et al., 2018)
ZFN SMC4-mEGFP #82-68 HeLa line	Ellenberg lab (EMBL)	(Walther et al., 2018)
ZFN AURKB-mEGFP #H24	Ellenberg lab (EMBL)	(Mahen et al., 2014)
ZFN mEGFP-NUP107 #26-31	Ellenberg lab (EMBL)	(Otsuka et al., 2016)
Recombinant DNA		
pEGFP-C1-EEA1	Addgene	Plasmid #42307 Lawe et al. (2000)
Software and Algorithms		
CellProfiler	Kamentsky et al. (2011)	www.cellprofiler.org
R 3.4.1	R Development Core Team	https://www.r-project.org
Matlab 2014a	MathWorks	www.mathworks.com
BH-tSNE	van der Maaten (2014)	R implementation: https://github.com/jkrijthe/Rtsne
Self-organizing map (SOM) algorithm and R package 'FlowSOM'	Van Gassen et al. (2015)	https://bioconductor.org/packages/release/bioc/html/FlowSOM.html R implementation: https://github.com/lmweber/FlowSOM-Rtsne-example/blob/master/FlowSOM_Rtsne_example.R.
CellClassifier (Pelkmans)	Rämö et al. (2009)	https://github.com/pelkmanslab/CellClassificationPelkmans; https://www.pelkmanslab.org/?page_id=63
Spot detection and segmentation (Pelkmans)	Battich et al. (2013)	https://github.com/pelkmanslab/ImageBasedTranscriptomics
Spot localization pattern analysis (Pelkmans)	Battich et al. (2013)	https://github.com/pelkmanslab/locpatterns
Population context features calculation (Pelkmans iBRAIN)	Snijder et al. (2009, 2012)	https://github.com/pelkmanslab/iBRAINShared/tree/master/iBRAIN/CreatePopulationContext

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Lucas Pelkmans (lucas.pelkmans@imls.uzh.ch).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell Lines Source and Cultivation

Endogenously mEGFP-tagged HeLa cell lines (Kyoto) were obtained from Jan Ellenberg (EMBL, Heidelberg, Germany) and endogenously mEGFP-tagged hiPS cells lines were obtained from the Allen Cell Institute (Seattle, Washington, USA) (http://www.allencell.org/cell-catalog.html).

HeLa cells were cultivated and seeded for experiments in plastic Greiner 96 well plates. hiPS cell lines were cultivated according to the protocol provided by the Allen Cell Institute, including coating of the plates and the seeding of the cells for the experiment, in plastic 96 well plates (Greiner).

(https://catalog.coriell.org/0/PDF/Allen/iPSC/AICS_SOP_WTC_CellCulture.pdf). For image-based transcriptomics and protein quantification 2500 single cells were seeded per well and cultivated for 3 days (2 nights), until they established population context, and confluency ~90%.

METHOD DETAILS

Experimental Design Details

Datasets represented in the figures for all HeLa and hiPS cells measurements contain two replicates. Table S1. also contains additional two replicates (total 4 replicates) of quantifications for mEGFP tagged HeLa cell lines. Two datasets that belong to mEGFP tagged HeLa cells are two biological replicates, each of which contains two technical replicates, as it is written within the Table S1.

Number of single cells used for generating the plots or performing statistical tests is written within the figures, and is also available within the Supplementary Tables, that contain measurements that accompany the figures and are used for generating the plots.

Replicate dataset for each gene was obtained by seeding cells in a two separate wells of a 96-well plate, representing therefore technical replicate (for both HeLa and hiPS cells). For all mEGFP HeLa cell lines, biological replicate was obtained by seeding each cell line in a different 96-well plate on a different day, several weeks apart, using different aliquots of frozen cells. No sample size estimations and no blinding were performed at any stage of the study. Further information on statistical tests is provided in the respective STAR Methods section.

Labelling of Cells with EdU

HeLa cells were labelled with EdU available from the Click-iT EdU Alexa Fluor 647 Imaging Kit, Thermo Fisher Scientific. Briefly, dilution of EdU was prepared in the cell culture media DMEM containing no serum, and dispensed into the plates containing growing cells, using Biomek liquid handling robot. Subsequently, cells were returned to the incubator for 15 additional minutes. After 15 min, media was aspirated, cells fixed, and processed for image-based transcriptomics and antibody staining.

Image-Based Transcriptomics and Antibody Labelling

Prior to combining image-based transcriptomics with imaging of GFP tagged cell lines or antibody labeling, we imaged all GFP tagged cell lines, subsequently we quantified the measured intensities in green channel, from which we subtracted the background as described in section below. Briefly, cells were grown for 3 days in 96-well plates, fixed in 4% PFA, permeabilized in 0.2% Triton X100 for 5 min, and incubated with DAPI for 10 min (for the purpose of nuclei segmentation). Such obtained intensity distributions of GFP lines was subsequently compared to the intensities obtained after imaging same cells, using same settings of microscope, which have been also subjected to the whole procedure of mRNA labeling using Affymetrix (now ThermoFisher) kit for bDNA based single molecule FISH. We observed no significant difference in intensity distributions measured in green channel, if we compared several proteins that localize in both cytoplasm in nucleus, or primarily in cytoplasm or nucleus, concluding that smFISH procedure is suitable to be combined with the imaging of GFP or labeling of protein epitopes, and obtained relative protein quantities are accurate.

Image-based transcriptomics and image processing was performed using open source software CellProfiler and our custom made Matlab modules available at https://github.com/pelkmanslab/CellProfilerPelkmans/tree/master/Modules, and https://github.com/pelkmanslab/ImageBasedTranscriptomics/tree/master/CellProfiler/Modules (Battich et al., 2013; Stoeger et al., 2015). Briefly, the modules have been at first tested locally on several images and obtained segmentation images manually inspected. Upon evaluating the good segmentation, pipeline was run on the whole image dataset, using ScienceCloud computational infrastructure of University of Zurich (UZH), or computational cluster Brutus (ETH). The example of the CellProfiler pipeline that contains modules used to analyze the images is within the Supplementary Data (ExamplePipeline.mat).

For all cell lines final dilution of the protease was 1:16000. Protease was first diluted 1:8000 in 1xPBS and dispensed in volume of 30uL, prior to which the volume within the well was aspirated to 30uL, such as is recommended in the protocol for bDNA smFISH (ThermoFisher Scientific, previously Affymetrix). Cells were seeded in 96-well plates and transcript of genes endogenously tagged with EGFP (in both HeLa and hiPS cells) were labelled in separate wells using primary probe against EGFP sequence and branched DNA single-molecule fluorescence *in-situ* hybridization using ViewRNA reagents (Affymetrix). All wash steps were performed on the Biomek robotic liquid handler. Imaging was done using automated confocal microscope, CellVoyager 7000 (Yokogawa) with the enhanced CSU-X1 spinning disc (Microlens enhanced dual Nipkow disc confocal scanner, wide view type) and a 40X Olympus

objective of 0.95 NA and Neo sCMOS cameras (Andor, 2.560 x 2.560 pixels). 12 Z slices with the distance of 1uM were imaged and maximum projection images (MIP) used for the further image analysis.

Genes that were not endogenously tagged (*TFRC, LAMP1, EEA1*, *VPS35* and *ACTB*) were labelled with gene specific probes Type 1 using the probes from previously published library: EEA1 (VA1-112193); TFRC (VA1-112208); LAMP1 (VA1-112096); VPS35 (VA1-112008); ACTB (VA1-10351-01); JUN (VA1-12265); FLY EGFP (VF1-10141); HPRT1 (VA1-11124-01); dapB (VF1-10272) (Battich et al., 2013). The protein was stained using following primary antibodies: TFRC (purified mouse anti-human CD71 antibody, BD Biosciences), LAMP1 (purified mouse anti-human CD107a antibody, BD Biosciences), EEA1 (purified mouse anti-human EEA1 antibody, clone 16, RUO, BD Biosciences), VPS35 (polyclonal goat anti-human VPS35 antibody, Novus Biologicals), ACTB (mouse monoclonal anti-actin antibody, ab3280, Abcam), c-JUN (rabbit monoclonal antibody anti-c-JUN (C.238.2)). Antibody labelling was performed after the sm-FISH protocol, cells were preblocked for 30 min. on room temperature in 2% BSA/ 1xPBS solution. Primary antibodies were applied for 2h on room temperature, and subsequently washed 4x with 1xPBS. Secondary antibodies were applied for 1h on room temperature and washed 5x with 1xPBS. Subsequently cells were stained with DAPI for 10 min., washed 3x with 1xPBS than stained with Succinimidyl Ester stain for 5 min. and finally washed 4x with 1xPBS.

Due to the quenching of EGFP, sm-FISH and Alexa fluorophores by buffer used for EdU Click-iT Kit, cells were first imaged for DAPI, sm-FISH, protein and Succinimidyl Ester stains. Subsequently, cells we stained for EdU, and the whole plate was reimaged using same setting on the microscope. Images were realigned based on the segmentations of the nuclear DAPI signal obtained from the first imaging cycle (using module plugged in CellProfiler pipeline), and subsequently the signal that belongs to EdU was quantified for each single cell.

Measurement and Quantification of Nuclear Transcripts

Nuclear transcripts have been visualized by applying the fixation protocol of 2% [v/v] glacial acetic acid in 4% PFA for 30 min (Battich et al., 2013). Transcripts were stained using same protocol as for the cytoplasmic bDNA sm-FISH, and the number of transcripts at the transcription sites were determined by dividing the total intensity of the burst by the average intensity of surrounding single spots. Subsequently the number of single spots was joined to the number of estimated spots within the burst giving number of total estimated spots per single nuclei (Battich et al., 2015).

EGF Stimulation Experiment

HeLa cells were grown in full growth media DMEM/10%FBS for 48 hr, subsequently the media was exchanged using Biomek robotic liquid handler, where cells were at first washed 3 times in DMEM without serum, and subsequently grown over night in the incubator in serum depleted media. Next morning cells were stimulated with the EGF (Millipore) at a final concentration of 20 ng/mL. Cells were stimulated in a retrograde fashion so that the fixation of the whole plate could be performed at once.

Transient Transfection of cDNA

HeLa cells were seeded in 96-well plate and cultivated for 36h prior to the transfection. cDNA coding for EGFP-EEA1 was transfected using GeneJuice reagent (Novagen) according to the manufacturer protocol and grown for another 24 hr. Cells were subsequently fixed and stained with bDNA sm-FISH probe against EGFP sequence.

QUANTIFICATION AND STATISTICAL ANALYSIS

Feature Extraction

Area, shape, intensities, and texture (at a scale of 5 pixels) of cells and nuclei were extracted using open source software CellProfiler (Carpenter et al., 2006). Correction of uneven illumination and subtraction of camera dependent invariant background was performed using custom CellProfiler modules as previously described (Battich et al., 2013; Stoeger et al., 2015), https://github.com/pelkmanslab. Briefly, large number of acquired images per channel can be used to learn pixel-wise illumination and signal gain biases. For each pixel, standard deviation and mean intensity value is calculated for a given channel. Illumination bias is than corrected by performing Z-scoring per pixel and reverting the values to the intensity values. Population context features were measured using previously published module implemented within the iBRAIN pipeline that calculates Local Cell Density and Distance to Edge for each single cell, after completion of the CellProfiler pipeline. Code for generation of population context features can be found on our Github repository https://github.com/pelkmanslab/iBRAINShared/tree/master/iBRAIN/CreatePopulationContext (Snijder et al., 2012). Number of directly adjacent cells and size of extracellular space with overlap to other cells was extracted using custom CellProfiler module, extending the cell outline by 10 pixels.

Cells that had multiple nuclei, border cells, and missegmented cells were discarded using supervised machine learning (SVM) tool CellClassifier (Rämö et al., 2009), available at https://www.pelkmanslab.org/?page_id=63. Briefly, images with overlayed segmentations of cells and their nuclei have been loaded in Matlab GUI of CellClassifier and classifier was manually trained by selecting cells with wrongly segmented nucleus as a class 1, and correctly segmented nucleus as class 2. Subsequently classifier was tested on a randomly selected images and applied to the whole dataset. Features of nuclear shape and DAPI intensity were used for the classification. Similarly, features of cellular shape were used to train classifier and recognize correctly segmented cells. Classification into four cell cycle stages was performed similarly, using SVM to classify cells in S phase (based on the intensity and texture features of the EdU staining) and M phase (based on the nuclei shape and DAPI intensity) and using Gaussian mixture model to distinguish G1

and G2 cells. Briefly, cells classified in S phase and M phase were excluded from the dataset, and histogram of integrated nuclear DAPI intensity consisted of two separated Gaussians, where the median of second Gaussian was at the double of the intensity of the median of the first Gaussian. Low-intensity Gaussian were G1 cells, and high-intensity Gaussian were G2 cells, sorted than subsequently into two classes by *k*-means clustering (Gut et al., 2015).

hiPS cells grow in colonies, where central regions contain cells overlapping in 3D. Such regions were not accurately segmented in 2D, and cells that have been inaccurately segmented we excluded using SVM classifier. For the same reason, we did not calculate and use population context features for linear models (local cell density and distance to edge), for hiPSC datasets.

Background Subtraction for mEGFP Tagged HeLa Cells

The expression from the endogenous locus results in lower intensities in the green channel in contrast to the antibody coupled fluorophores, requiring therefore longer exposure times (700ms), that led to a detection of background. We have therefore performed background subtraction. Precisely, in the same experimental plate we have seeded parental HeLa line (non-tagged) as a negative control for the sm-FISH labeling with EGFP and DapB probes. Such cells have been imaged with same setting for all channels. Subsequently the intensities obtained after microscope background subtraction have been observed to linearly scale with the cell area feature, summing the plate and the cell-based background into one unique single cell value. Matlab function ''fitlm'' was used to fit linear model using cell area as a predictor for the integrated background intensity in parental HeLa cells, and obtained model was used to predict the background for all measured single cells across the plate. Predicted values have been than subtracted from the total measured intensities and obtained values used for the further analysis, as a background corrected relative protein quantities (AU).

Selection of Genes for the Transcript and Protein Quantification

Initial gene selection contained several genes that have been measured, and subsequently discarded due to the low intensity in the green channel, that upon background subtraction led to a large number of single cells having negative values for the protein intensity. Also, if their final mean intensity per cell was lower than 3 standard deviations from mean of the intensity in background subtracted negative control, we considered them to be too close to the background and they were not further analyzed. Those are *NCAPH2* (HeLa cell line) and *PAX* and *DSP* (hiPS cell lines).

Generation of Components for Transcript and Protein Abundance Prediction

Dataset was first Winsorized at the 0.5% and 99.5% percentiles, subsequently all the features were Z-scored on per-well and again per-gene basis (for the analysis where both replicates have been pooled together). Generalized linear models using transcript abundance as a predictor for the protein abundance were built using ''fitlm'' function of Matlab with robust fit option on. Multivariate linear regression models were built using ''plsregress'' function of Matlab, where number of components was optimized on per-gene basis using 100 cross validations and based on the decrease in estimated mean squared prediction error (MSE). For majority of the multivariate linear models (PLSR) models number of required components was in the range of 5-10. Additionally, optimal number of components led to a good data prediction without overfitting of the training set, that was estimated on the similarity of the explained variance for both training and test dataset.

Cell Cycle Significance Calculation

To examine the significance of the association (contingency) between the cell cycle classification and 'outlier' classification for cells classified as outliers, we used Fisher's exact test, specifically function 'FisherExactTest' within package FisherExactTest from Matlab (https://ch.mathworks.com/matlabcentral/fileexchange/24379-fisher-s-exact-test-with-n-x-m-contingency-table).

To calculate the fraction of outliers in un-expected cell-cycle stage, we have first estimated the number of expected outliers in each cell-cycle stage based on the measured cell cycle stage proportions and those calculated for the whole population of single cells, for each gene separately. Subsequently, we subtracted the differences in expected and measured number of outliers in each cell-cycle stage and used sum of their absolute values divided by the total number of outliers (for each gene) as a measure of proportion.

Embedding of Single Cells in Two Dimensional bh-tSNE Feature Space

Bh-tSNE algorithm for dimensionality reduction was used within the R wrapper package 'Rtsne', https://github.com/jkrijthe/Rtsne, that can be installed from CRAN or github repository. For BH-tSNE embedding we used cells coming from both replicates for each gene. Used features were Z-score normalized for all datasets (all EGFP tagged HeLa cells, all antibody labeled HeLa cells, and all hiPS cells). Each point within the BH-tSNE plots represents a single cell.

Features used for tSNE embedding are represented in the Figure S4.

Comparison of Distributions Using Kolmogorov-Smirnov Statistics and F-Test

The distributions of the replicates or measured and predicted datasets were compared using Matlab function ''kstest2''. To compare multiple sample distributions using statistic of the maximum difference in probability of cumulative distributions extended Kolmogorov-Smirnov test was used, available in open source ''KS statistic'' package (https://ch.mathworks.com/matlabcentral/fileexchange/47900-ks-statistic-zip). To calculate *p* values for F-test, function 'vartest2' of Matlab was used, with the default boundary of significance of 5%.

Classification of Cells Based on the Transcript Spatial Pattern

Probabilities to belong to a specific spatial class was calculated based on the set of spatial features extracted using our custom CellProfiler module. Briefly, we first calculated the primary set of features using module MeasureLocalizationOfSpots.m that is available on our Github https://github.com/pelkmanslab/ImageBasedTranscriptomics/tree/master/CellProfiler/Modules. Subsequently, for every cell and set of primary spatial features, cellular features were obtained by the custom MeasureChildren.m CellProfiler module, available at https://github.com/pelkmanslab/ImageBasedTranscriptomics/tree/master/CellProfiler/Modules (Battich et al., 2013). Further, only cellular features describing mean and sd of spatial features for each single cell were used to perform hierarchical clustering of cells using Euclidean distance space and Ward's linkage method. Cells were grouped in 50 bins, and 1000 samplings of 100 cells was performed using script available at https://github.com/pelkmanslab/locpatterns. To classify clusters from those different samplings into spatial pattern types the centroid for each cluster was computed (using 7 centroids), and distance to a number of randomly sampled cluster centroids measured. The localization type of the cell was than defined as that of its closest centroid, and probability as a fraction of times a cell was defined to belong to a particular localization type. To discretize the class assignment based on the probabilities, threshold of 0.6 was used, where each cell was assigned to a class 1, class 2 or class 3, only when the probability was higher than the threshold (Figures 5D–5F). Such discretization was used mainly for visualization purpose, while all other plots represent calculations based on exact probabilities.

Binning of Cells Based on Cellular Features

Binning of cells in JUN experiment was performed as following: both replicates have been used, 12000 cells taken randomly from each timepoint (apart from the timepoint 210 min.) and cells have been clustered based on their DNA content, nuclear and cellular size and local cell density features, using Self Organizing Map algorithm, and R package 'FlowSOM'

(https://bioconductor.org/packages/release/bioc/html/FlowSOM.html) (Van Gassen et al., 2015). The implementation was used as described in https://github.com/lmweber/FlowSOM-Rtsne-example/blob/master/FlowSOM_Rtsne_example.R. All the features have been normalized by mean Z-score normalization across all the single cells. Default setting of FlowSOM initially created 100 clusters, after which we discarded clusters if they contained less than 20 cells from any of the timepoints, leaving 79 clusters. Upon obtaining index that defined the cluster assignment for each cell, we calculated the median values for the features of interest, such as total DNA content, cell and nuclear size, local cell density, total protein content as well as mRNA count in the nucleus and cytoplasm, and protein content. Upon obtaining median values we compared the distributions of 79 bins to distributions of all single cells. Also, upon extracting the bin assignment indices, we have randomized them and calculated the medians of all features based on random grouping of cells. When we compared KS statistics for bins based on SOM, or bins with randomly assigned cells, we observed that SOM assigned bins resemble close the distribution of the whole population, which was not the case for randomly assigned cells.

Using medians for mRNA quantity in the nucleus and cytoplasm, and protein intensity we generated pseudo time-traces (Figure 4), and we calculated the deltas that described the change in quantities of all three molecular species in time. Precisely, calculation of deltas (changes) in nuclear mRNA (export), cytoplasmic mRNA (import and decay), as well as deltas (changes) of protein quantities (translation) was calculated as an absolute value difference (between values measured at two timepoints for which delta is to be calculated), and divided by time in minutes (time difference between the two timepoints for which delta is to be calculated) for each cluster and expressed as a change in quantity per minute:

$$\Delta x = \left| \frac{\min(x) - \max(x)}{time(minutes)} \right|.$$

BH-tSNE embedding of SOM obtained clusters was done using median bin values of cell autonomous features including those used for generation of SOMs ('Intensity_SubsBlue_Nuclei_1_IntegratedIntensity',
'Intensity_SubsBlue_Nuclei_2_MeanIntensity',
'Intensity_SubsFarRed_Cells_1_IntegratedIntensity','Intensity_SubsFarRed_Nuclei_1_IntegratedIntensity',
'Texture_5_SubsFarRed_Cells_1_AngularSecondMoment',
'Texture_5_SubsFarRed_Nuclei_1_AngularSecondMoment',
'AreaShape_Nuclei_1_Area', 'AreaShape_Nuclei_2_Eccentricity'
'AreaShape_Cells_1_Area','AreaShape_Cells_2_Eccentricity',
'DistanceToEdge_Nuclei_1_LCD_Border-4_Size864_Sigma144_Shrink18_TotPSF401',
'LocalCellDensity_Nuclei_1_LCD_Border-4_Size864_Sigma144_Shrink18_TotPSF401'). Subsequently median values of all other features and calculated deltas have been projected on the BH-tSNE plots.

Analysis of Signal Cross-Correlation

Analysis of *JUN* experiment was performed using Signal Processing Toolbox of Matlab. Precisely, the lags between signals of nuclear and cytoplasmic transcript abundance, and transcript and protein abundance and the cross-correlation of signals was calculated using "xcorr" function with "coeff" option (https://ch.mathworks.com/help/signal/ref/xcorr.html). Lags at which maximum cross-correlation was observed (time at which two peaks overlapped) were than transformed to time based on where the peak of each quantity occurs.

Cell Systems, Volume *7*

Supplemental Information

Multivariate Control of Transcript to Protein Variability in Single Mammalian Cells

Doris Popovic, Birgit Koch, Moritz Kueblbeck, Jan Ellenberg, and Lucas Pelkmans

Figure S1.

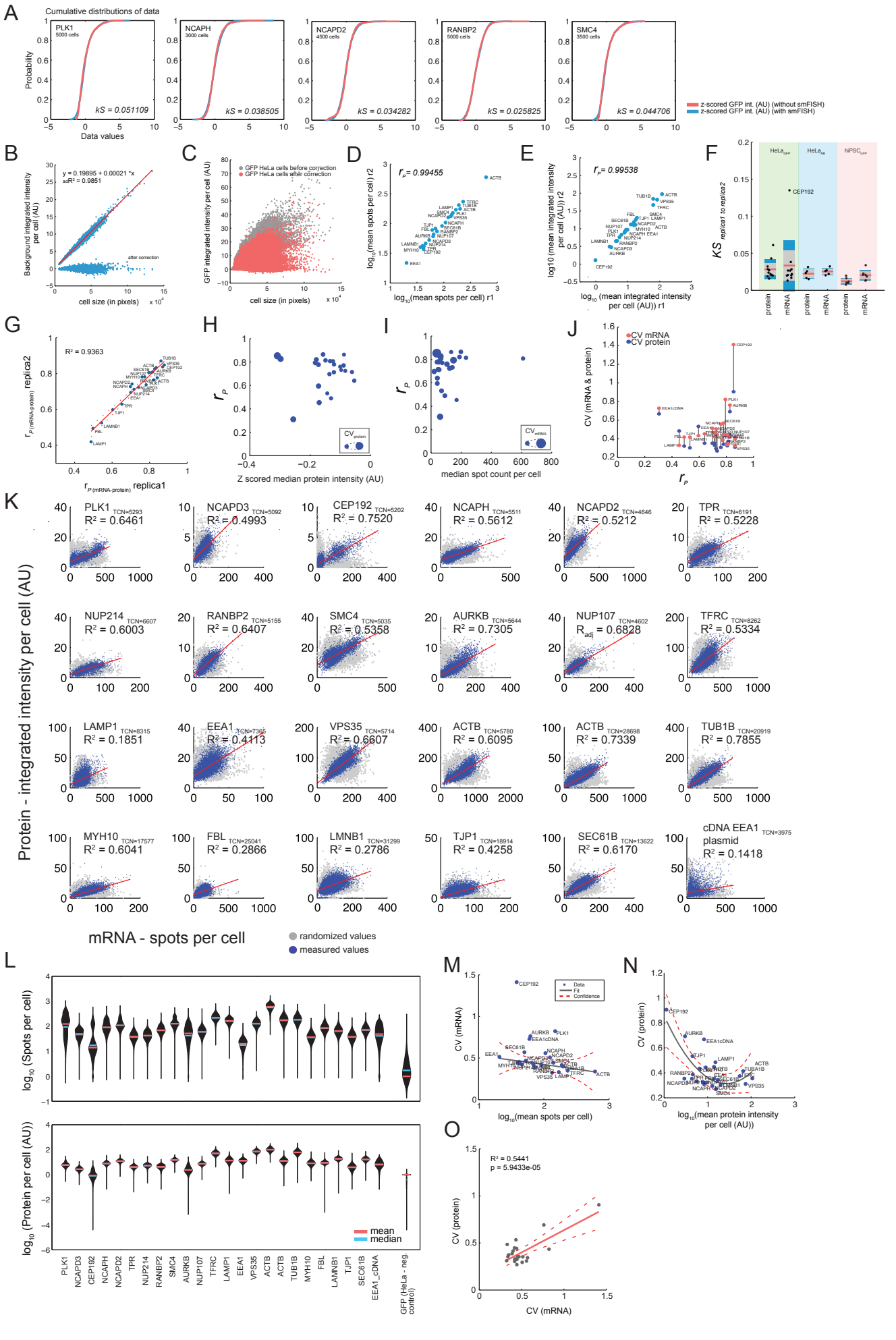

A
Cumulative distributions of data
PLK1
5000 cells
NCAPH
3000 cells
NCAPD2
4500 cells
RANBP2
5000 cells
SMC4
3500 cells
Probability
Data values
kS = 0.051109
kS = 0.038505
kS = 0.034282
kS = 0.025825
kS = 0.044706
z-scored GFP int. (AU) (without smFISH)
z-scored GFP int. (AU) (with smFISH)
B
Background integrated intensity per cell (AU)
cell size (in pixels)
after correction
C
GFP integrated intensity per cell (AU)
GFP HeLa cells before correction
GFP HeLa cells after correction
cell size (in pixels)
D
log10(mean spots per cell) r2
log10(mean spots per cell) r1
E
log10 (mean integrated intensity per cell (AU)) r2
log10 (mean integrated intensity per cell (AU)) r1
F
KS replica1 to replica2
protein
mRNA
G
replica2
replica1
H
Z scored median protein intensity (AU)
I
median spot count per cell
J
CV mRNA
CV protein
CV (mRNA & protein)
K
Protein - integrated intensity per cell (AU)
mRNA - spots per cell
randomized values
measured values
L
log10 (Spots per cell)
log10 (Protein per cell (AU))
mean
median
M
CV (mRNA)
log10(mean spots per cell)
Data
Fit
Confidence
N
CV (protein)
log10(mean protein intensity per cell (AU))
O
R² = 0.5441
p = 5.9433e-05
CV (protein)
CV (mRNA)

Figure S1. Image-Based Quantification of mRNA and Protein Levels and Cell- to-Cell Variability in Their Abundance, Related to Figure 1.

(A) Cumulative probability distributions of integrated background corrected single cell intensities in green channel of few representative GFP endogenously tagged HeLa cell lines non- or subjected to the sm-FISH protocol. Kolmogorov-Smirnov statistics was calculated on Z-scored values and represented on the plots.

(B) Background subtraction based on linear fit of measured intensity in green channel in parental HeLa cell line. Liner model obtained was used to predict the background values based on cell size, after subtraction of the background the values of protein intensity in negative control are distributed around the zero.

(C) Correction of the measured intensities for all the single cells HeLa cells tagged with mEGFP.

(D) Pearson's correlation of two replicates of measured cytoplasmic mRNA count.

(E) Pearson's correlation of two replicates of measured protein quantity (integrated protein intensity per cell (AU)).

(F) Kolmogorov-Smirnov statistic comparing the measured shapes of single-cell protein and transcript abundance distributions of two technical replicates. Green color: all genes measured in endogenously tagged HeLa cell lines, blue: all genes measured using antibody labeling in HeLa cells, red: all genes measured in endogenously tagged hiPS cells.

(G) Correlation of calculated mRNA-protein Pearson's correlation coefficient of two technical replicates, for all genes that were measured in HeLa and hiPS cells.

(H) Scaling of measured protein quantities with the Pearson's correlation coefficient of mRNA and protein for all the genes. Size of the dot represents the coefficient of variation in protein and mRNA, respectively.

(I) Scaling of cytoplasmic measured mRNA quantities with the Pearson's correlation coefficient of mRNA and protein for all the genes. Size of the dot represents the coefficient of variation in protein and mRNA, respectively.

(J) Scaling of mRNA-protein correlation with the coefficient of variation in protein and mRNA abundance.

(K) Scatter plots between mRNA and protein abundance. Each plot has depicted R^2 of the linear fit using mRNA as predictor for the protein quantity, and total number of measured cells (TCN). Linear fit was performed on pooled cells coming from both of the replicates, that are represented by the blue color dots. Grey color dots represent randomized values of protein quantity for the given mRNA count.

(L) Distributions of cytoplasmic mRNA count (upper) and protein quantity (lower) for all the genes measured in HeLa cells and hiPS cells.

(M) Coefficient of variation plotted against the mean spot count per cell.

(N) Coefficient of variation plotted against the mean integrated protein intensity per cell (AU).

(O) Correlation of cell-to-cell variability (CV) in mRNA and protein abundance across all genes.

Figure S2.

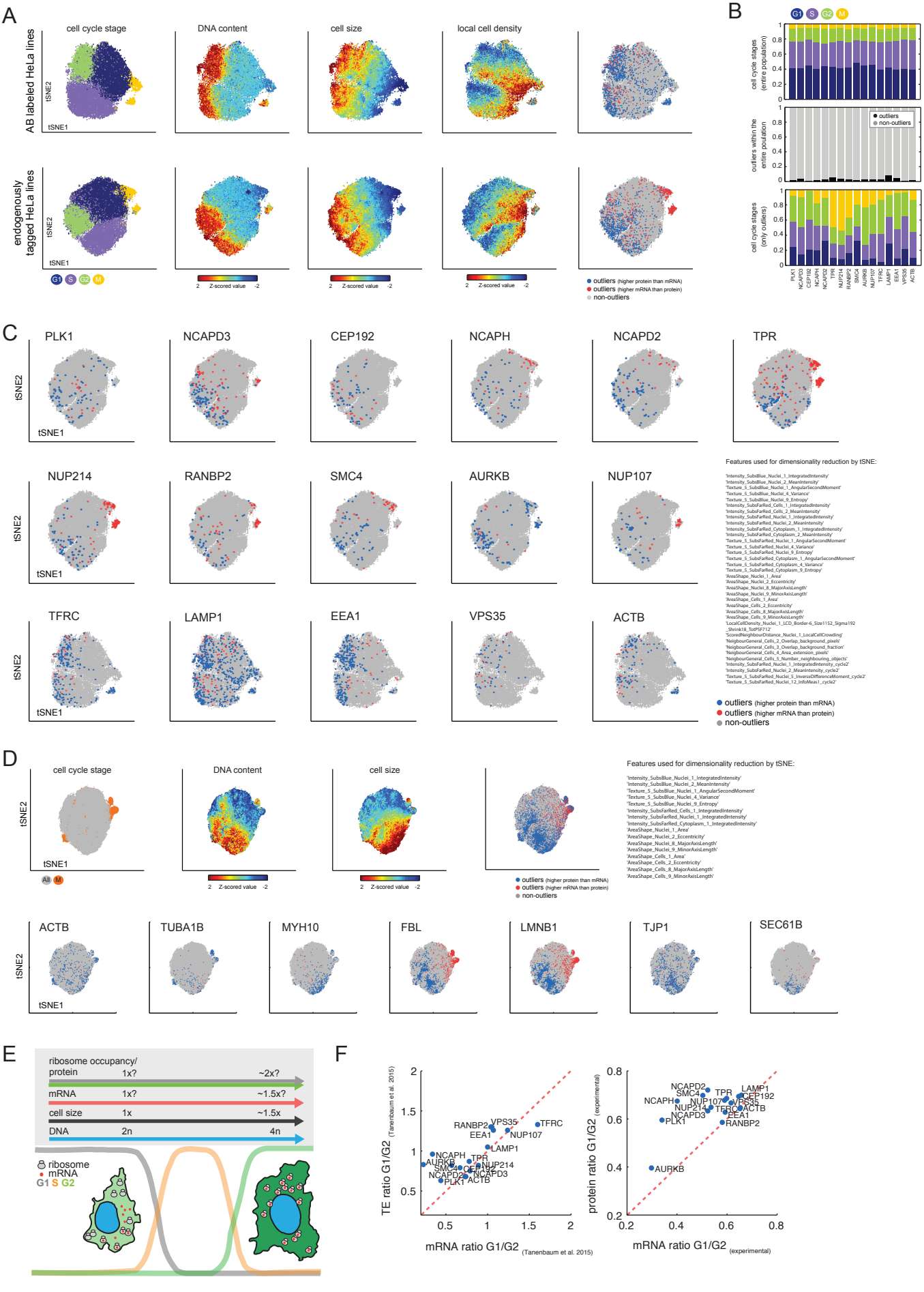

A
cell cycle stage
DNA content
cell size
local cell density
AB labeled HeLa lines
endogenously tagged HeLa lines
tSNE1
tSNE2
G1 S G2 M
2 Z-scored value -2
outliers (higher protein than mRNA)
outliers (higher mRNA than protein)
non-outliers
B
cell cycle stages (entire population)
outliers within the entire population
outliers
non-outliers
cell cycle stages (only outliers)
PLK1 NCAPD3 CEP192 NCAPH NCAPD2 TPR NUP214 RANBP2 SMC4 AURKB NUP107 TFRC LAMP1 EEA1 VPS35 ACTB
C
PLK1
NCAPD3
CEP192
NCAPH
NCAPD2
TPR
NUP214
RANBP2
SMC4
AURKB
NUP107
TFRC
LAMP1
EEA1
VPS35
ACTB
Features used for dimensionality reduction by tSNE:
'Intensity_SubsBlue_Nuclei_1_IntegratedIntensity'
'Intensity_SubsBlue_Nuclei_2_MeanIntensity'
'Texture_S_SubsBlue_Nuclei_1_AngularSecondMoment'
'Texture_S_SubsBlue_Nuclei_4_Variance'
'Texture_S_SubsBlue_Nuclei_9_Entropy'
'Intensity_SubsFarRed_Cells_1_IntegratedIntensity'
'Intensity_SubsFarRed_Cells_2_MeanIntensity'
'Intensity_SubsFarRed_Nuclei_1_IntegratedIntensity'
'Intensity_SubsFarRed_Nuclei_2_MeanIntensity'
'Intensity_SubsFarRed_Cytoplasm_1_IntegratedIntensity'
'Intensity_SubsFarRed_Cytoplasm_2_MeanIntensity'
'Texture_S_SubsFarRed_Nuclei_1_AngularSecondMoment'
'Texture_S_SubsFarRed_Nuclei_4_Variance'
'Texture_S_SubsFarRed_Nuclei_9_Entropy'
'Texture_S_SubsFarRed_Cytoplasm_1_AngularSecondMoment'
'Texture_S_SubsFarRed_Cytoplasm_4_Variance'
'Texture_S_SubsFarRed_Cytoplasm_9_Entropy'
'AreaShape_Nuclei_1_Area'
'AreaShape_Nuclei_2_Eccentricity'
'AreaShape_Nuclei_8_MajorAxisLength'
'AreaShape_Nuclei_9_MinorAxisLength'
'AreaShape_Cells_1_Area'
'AreaShape_Cells_2_Eccentricity'
'AreaShape_Cells_8_MajorAxisLength'
'AreaShape_Cells_9_MinorAxisLength'
'LocalCellDensity_Nuclei_1_LCD_Border-6_Size1152_Sigma192_Shrink18_TotPSF712'
'ScondNeighbourDistance_Nuclei_1_LocalCellCrowding'
'NeighbourGeneral_Cells_2_Overlap_background_pixels'
'NeighbourGeneral_Cells_3_Overlap_background_fraction'
'NeighbourGeneral_Cells_4_Area_extension_pixels'
'NeighbourGeneral_Cells_5_Number_neighbouring_objects'
'Intensity_SubsFarRed_Nuclei_1_IntegratedIntensity_cycle2'
'Intensity_SubsFarRed_Nuclei_2_MeanIntensity_cycle2'
'Texture_S_SubsFarRed_Nuclei_5_InverseDifferenceMoment_cycle2'
'Texture_S_SubsFarRed_Nuclei_12_InfoMeas1_cycle2'
outliers (higher protein than mRNA)
outliers (higher mRNA than protein)
non-outliers
D
cell cycle stage
DNA content
cell size
Features used for dimensionality reduction by tSNE:
'Intensity_SubsBlue_Nuclei_1_IntegratedIntensity'
'Intensity_SubsBlue_Nuclei_2_MeanIntensity'
'Texture_S_SubsBlue_Nuclei_1_AngularSecondMoment'
'Texture_S_SubsBlue_Nuclei_4_Variance'
'Texture_S_SubsBlue_Nuclei_9_Entropy'
'Intensity_SubsFarRed_Cells_1_IntegratedIntensity'
'Intensity_SubsFarRed_Nuclei_1_IntegratedIntensity'
'Intensity_SubsFarRed_Cytoplasm_1_IntegratedIntensity'
'AreaShape_Nuclei_1_Area'
'AreaShape_Nuclei_2_Eccentricity'
'AreaShape_Nuclei_8_MajorAxisLength'
'AreaShape_Nuclei_9_MinorAxisLength'
'AreaShape_Cells_1_Area'
'AreaShape_Cells_2_Eccentricity'
'AreaShape_Cells_8_MajorAxisLength'
'AreaShape_Cells_9_MinorAxisLength'
outliers (higher protein than mRNA)
outliers (higher mRNA than protein)
non-outliers
ACTB
TUBA1B
MYH10
FBL
LMNB1
TJP1
SEC61B
E
ribosome occupancy/protein
1x?
~2x?
mRNA
1x?
~1.5x?
cell size
1x
~1.5x
DNA
2n
4n
ribosome
mRNA
G1 S G2
F
TE ratio G1/G2 (Tanenbaum et al. 2015)
mRNA ratio G1/G2 (Tanenbaum et al. 2015)
protein ratio G1/G2 (experimental)
mRNA ratio G1/G2 (experimental)
RANBP2 VPS35 TFRC EEA1 NUP107 LAMP1 NCAPH TPR AURKB SMC4 CEP192 NUP214 NCAPD2 NCAPD3 PLK1 ACTB
NCAPD2 SMC4 TPR LAMP1 CEP192 NCAPH NUP107 VPS35 NUP214 TFRC ACTB NCAPD3 EEA1 PLK1 RANBP2 AURKB

Figure S2. Explaining Variability in Protein Abundance and Cytoplasmic mRNA Abundance in Single Cells. Related to Figure 2.

(A) BH-tSNE embedding of single HeLa cells and the spread of the cells classified as “outliers”. Upper row of plots: all HeLa cells labeled with antibodies pooled together, lower row of plots: all mEGFP tagged HeLa cells pooled together.

(B) Upper bar-plot: classification of all the HeLa cells in different stages of a cell-cycle, as projected on BH-tSNE in (A), middle bar-plot: percentage of “outliers” within the population of all measured single cells, lower bar-plot: classification of “outlier” fraction in different cell cycle stages.

(C) Outliers projected on BH-tSNE plots as described in Figure 2B. Text on the right: cellular features used for BH-tSNE based dimensionality reduction.

(D) BH-tSNE embedding of hiPS cells. Outliers for separate genes are projected on lower tSNE plots. Text on the right: Cellular features used for tSNE based dimensionality reduction.

(E) Schematic representation of changes that can impact the protein-mRNA ratio during the cell cycle, leading to the appearance of cell-cycle specific outliers in mRNA-protein scatter plots. mRNA and protein quantities can change depending on genome content, cellular size, or translational efficiency of ribosomes.

(F) Ratio of mean translational efficiency of cells in G2 and G1 cells plotted against the ratio of the mean mRNA content in G1 and G2 cells, calculated from the data provided in work of Tanenbaum et al. (2015). In case when mRNA content is higher in G2 cells (x axis values lower than 1), translational efficiency in G2 cells can be higher leading to higher protein levels than mRNA (eg. PLK1).

(G) Ratio of mean protein levels (integrated intensity per cell (AU)) of cells in G2 and G1 cells plotted against the ratio of the mean mRNA content in G1 and G2 cells, calculated from data obtained in this study. Calculated ratios for both datasets (F and G) were represented within the Figure 2F.

Figure S3.

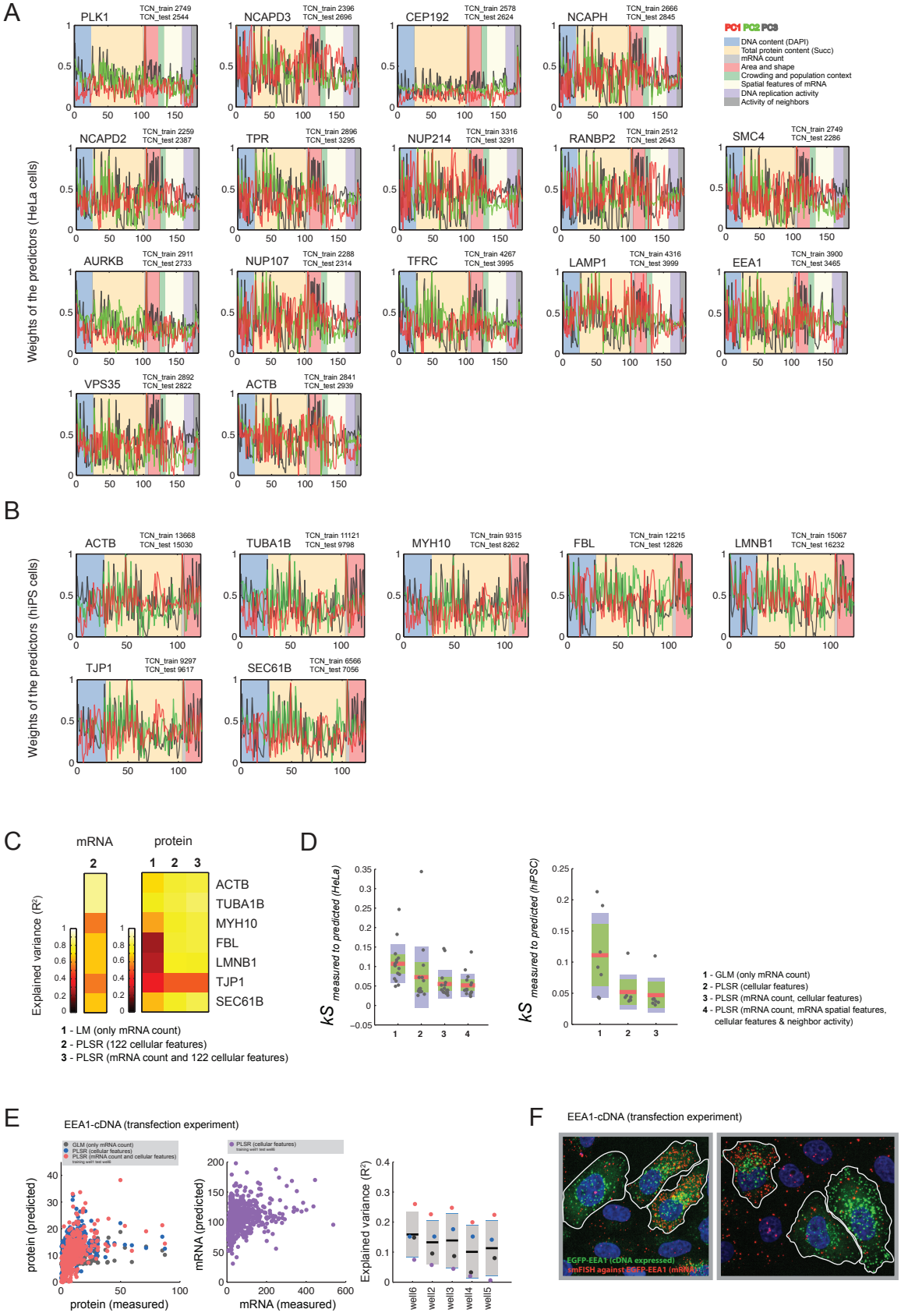

A
PLK1
NCAPD3
CEP192
NCAPH
NCAPD2
TPR
NUP214
RANBP2
SMC4
AURKB
NUP107
TFRC
LAMP1
EEA1
VPS35
ACTB
Weights of the predictors (HeLa cells)
PC1 PC2 PC3
DNA content (DAPI)
Total protein content (Succ)
mRNA count
Area and shape
Crowding and population context
Spatial features of mRNA
DNA replication activity
Activity of neighbors
B
ACTB
TUBA1B
MYH10
FBL
LMNB1
TJP1
SEC61B
Weights of the predictors (hiPS cells)
C
mRNA
protein
Explained variance (R²)
ACTB
TUBA1B
MYH10
FBL
LMNB1
TJP1
SEC61B
1 - LM (only mRNA count)
2 - PLSR (122 cellular features)
3 - PLSR (mRNA count and 122 cellular features)
D
ks measured to predicted (HeLa)
ks measured to predicted (hiPSC)
1 - GLM (only mRNA count)
2 - PLSR (cellular features)
3 - PLSR (mRNA count, cellular features)
4 - PLSR (mRNA count, mRNA spatial features, cellular features & neighbor activity)
E
EEA1-cDNA (transfection experiment)
protein (predicted)
protein (measured)
mRNA (predicted)
mRNA (measured)
Explained variance (R²)
well6
well2
well3
well4
well5
F
EEA1-cDNA (transfection experiment)
EGFP-EEA1 (cDNA expressed)
smFISH against EGFP-EEA1 (mRNA)

Figure S3. Explaining Variability in Protein Abundance and Cytoplasmic mRNA Abundance in Single Cells. Related to Figure 3.

(A) Examples of weights for the first 3 principal components of full PLSR model used to explain variability in protein abundance in HeLa cells. (Model 3 in Figure 3B). Above the plots is the total cell number (TCN) used as a training set for the models (technical replicate 1) and as a test set for the models (technical replicate 2). Right side of the plots represents the features used to create principal components, colormap of the legend corresponds to the colormap on the plots. Weights are min-max normalized prior to the plotting. Total number of principal components for every model together with the KS statistics is provided in Table 4.

(B) Examples of weights for the first 3 components of PLSR model used to explain variability in protein abundance in hiPS cells. (Model 3 in Figure S3C). Total cell number (TCN) used as a training dataset (technical replicate 1) and test dataset (technical replicate 2) is written above each plot.

(C) Left: Explained variance of cytoplasmic mRNA using multivariate linear model (PLSR) based on cellular features reduced to principal components. Right: Explained variance of protein content based on linear model using only mRNA as predictor variable (LM) or various cellular features reduced to the principal components for multivariate linear models (PLSR).

(D) Kolmogorov-Smirnov statistic comparing the measured and predicted shapes of single-cell protein abundance. Left: KS distances for models on HeLa cells. Right: KS distances for models on hiPS cells. The exact values are provided in Table 4.

(E) Scatter plot for measured and predicted protein content (left) or mRNA content (right) for EEA1 gene expressed from ectopically introduced plasmid into HeLa cells (transfection). Very right: Explained variance for all the PLSR models across six tested sets of transfected HeLa cells. Calculated values are provided in Table 4.

(F) Example images of HeLa cells transfected with the cDNA coding for EGFP-EEA1.

Figure S4.

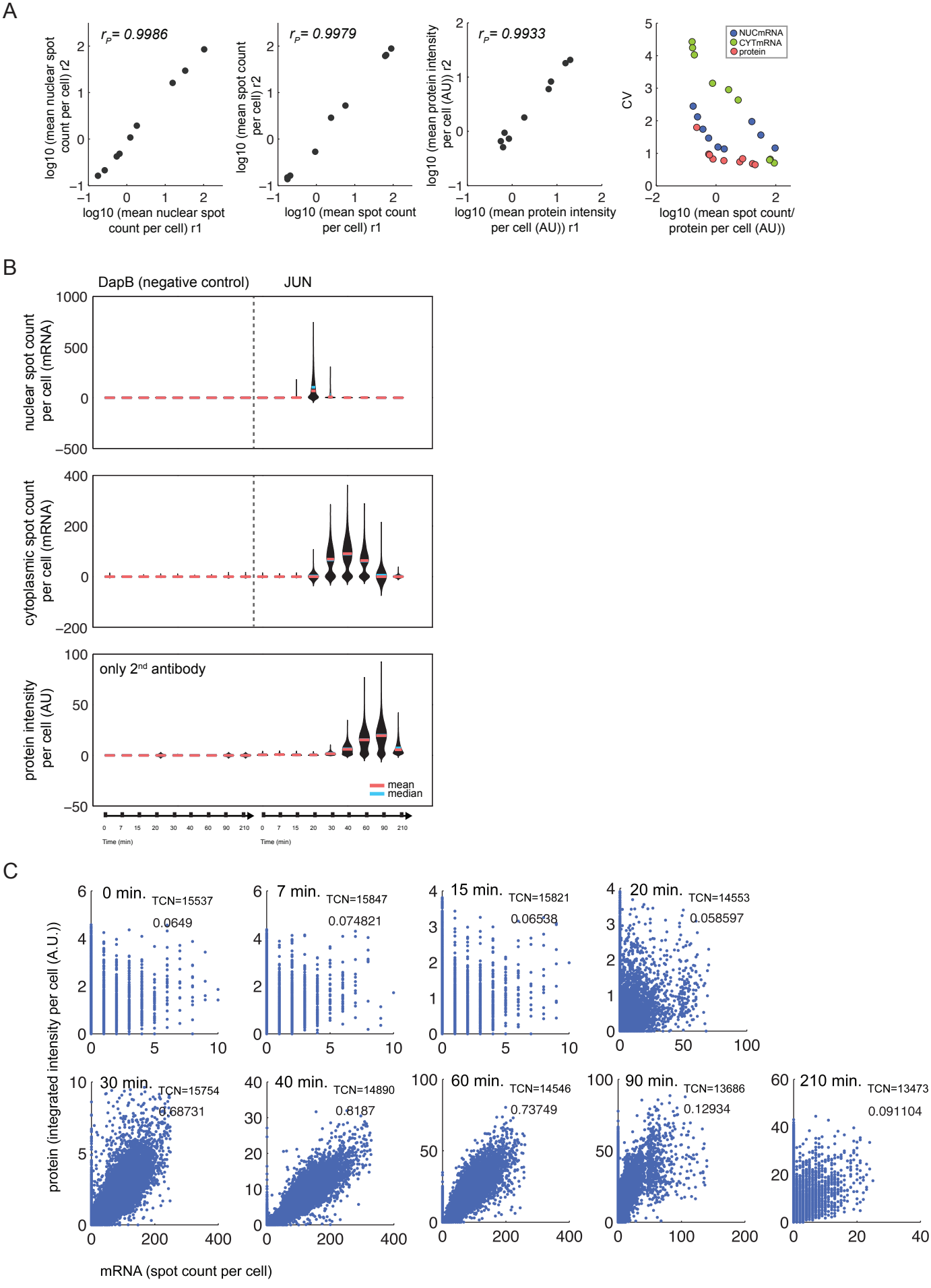

A
r_P= 0.9986
log10 (mean nuclear spot count per cell) r2
log10 (mean nuclear spot count per cell) r1
r_P= 0.9979
log10 (mean spot count per cell) r2
log10 (mean spot count per cell) r1
r_P= 0.9933
log10 (mean protein intensity per cell (AU)) r2
log10 (mean protein intensity per cell (AU)) r1
NUCmRNA
CYTmRNA
protein
CV
log10 (mean spot count/ protein per cell (AU))
B
DapB (negative control)
JUN
nuclear spot count per cell (mRNA)
cytoplasmic spot count per cell (mRNA)
only 2nd antibody
protein intensity per cell (AU)
mean
median
Time (min)
C
protein (integrated intensity per cell (A.U.))
mRNA (spot count per cell)
0 min. TCN=15537 0.0649
7 min. TCN=15847 0.074821
15 min. TCN=15821 0.06538
20 min. TCN=14553 0.058597
30 min. TCN=15754 0.68731
40 min. TCN=14890 0.8187
60 min. TCN=14546 0.73749
90 min. TCN=13686 0.12934
210 min. TCN=13473 0.091104

Figure S4. mRNA-protein correlations in EGF gene induction system. Related to Figure 4.

(A) Pearson's correlation of *JUN* protein and transcript abundance in the cytoplasm and the nucleus for two technical replicates of EGF stimulation. Very right: coefficient of variation for all three molecular species: nuclear mRNA, cytoplasmic mRNA and protein.

(B) Distributions of *JUN* transcript abundance in the nucleus and cytoplasm, and protein abundance across all the timepoints.

(C) Correlation between *JUN* cytoplasmic mRNA abundance and protein abundance for different timepoints of EGF stimulation in HeLa cells. Both technical replicates are pooled together and plotted, total cell number (TCN) together with the R^2 are written on the plots.

Figure S5.

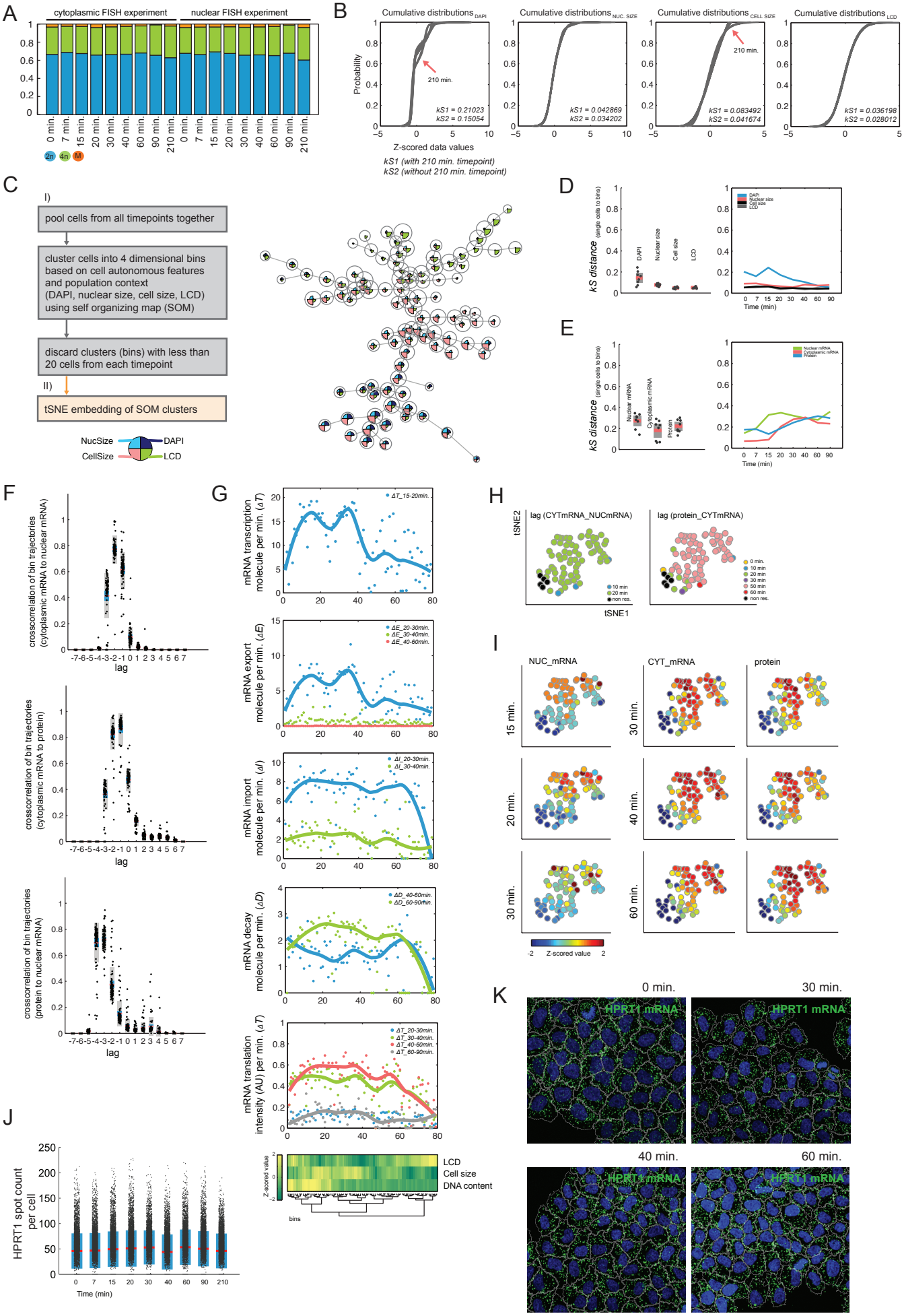

A
cytoplasmic FISH experiment
nuclear FISH experiment
0 min.
7 min.
15 min.
20 min.
30 min.
40 min.
60 min.
90 min.
210 min.
2n
4n
M
B
Cumulative distributions DAPI
Cumulative distributions NUC. SIZE
Cumulative distributions CELL SIZE
Cumulative distributions LCD
Probability
Z-scored data values
210 min.
kS1 = 0.21023
kS2 = 0.15054
kS1 = 0.042869
kS2 = 0.034202
kS1 = 0.083492
kS2 = 0.041674
kS1 = 0.036198
kS2 = 0.028012
kS1 (with 210 min. timepoint)
kS2 (without 210 min. timepoint)
C
I)
pool cells from all timepoints together
cluster cells into 4 dimensional bins based on cell autonomous features and population context (DAPI, nuclear size, cell size, LCD) using self organizing map (SOM)
discard clusters (bins) with less than 20 cells from each timepoint
II)
tSNE embedding of SOM clusters
NucSize
CellSize
DAPI
LCD
D
kS distance (angle cells to bins)
DAPI
Nuclear size
Cell size
LCD
Time (min)
E
kS distance (angle cells to bins)
Nuclear mRNA
Cytoplasmic mRNA
Protein
Time (min)
F
crosscorrelation of bin trajectories (cytoplasmic mRNA to nuclear mRNA)
lag
crosscorrelation of bin trajectories (cytoplasmic mRNA to protein)
lag
crosscorrelation of bin trajectories (protein to nuclear mRNA)
lag
G
mRNA transcription molecule per min. (αT)
ΔT_15-20min.
mRNA export molecule per min. (αE)
ΔE_20-30min.
ΔE_30-40min.
ΔE_40-60min.
mRNA import molecule per min. (αI)
ΔI_20-30min.
ΔI_30-40min.
mRNA decay molecule per min. (αD)
ΔD_40-60min.
ΔD_60-90min.
mRNA translation intensity (AU) per min. (αT)
ΔT_20-30min.
ΔT_30-40min.
ΔT_40-60min.
ΔT_60-90min.
Z-scored value
bins
LCD
Cell size
DNA content
H
tSNE2
lag (CYTmRNA_NUCmRNA)
lag (protein_CYTmRNA)
10 min
20 min
non res.
0 min.
10 min
20 min
30 min
50 min
60 min
non res.
tSNE1
I
NUC_mRNA
CYT_mRNA
protein
15 min.
30 min.
20 min.
40 min.
30 min.
60 min.
-2
Z-scored value
2
J
HPRT1 spot count per cell
Time (min)
K
0 min.
30 min.
40 min.
60 min.
HPRT1 mRNA

Figure S5. mRNA-protein correlations of *JUN* in EGF gene induction system and variability in spatial positioning of mRNA. Related to Figure 4 and Figure 5.

(A) Bar-plots representing the proportions of cells with single or double genome content and mitotic cells within both technical replicates of EGF stimulation experiment pooled together.

(B) Cumulative distributions for all the single cells for all of the timepoints for features used to cluster the cells using self-organizing maps. Only upon 210 minutes of EGF stimulation cumulative distributions for nucleus and cell size change (as pointed by the red arrow on the plots). Kolmogorov-Smirnov statistics was calculated to compare the similarity of distributions of all 4 cell phenotypic features and is written within the plots.

(C) Scheme of the workflow used for binning of single cells. Right: SOM tree representing 100 clusters of single cells. Each cluster pie-chart represents the values of 4 cell phenotypic features for cells that are within the cluster.

(D) Kolmogorov-Smirnov statistic comparing the shapes of features for all the single cells used for SOM clustering and shape of distribution obtained upon clustering (based on calculated medians for each of the used feature). Left: bar-plots representing the KS distances of single cells and clusters for all timepoints, based on SOM clustering. Right: bar-plots representing the KS distances of single cells and clusters for all timepoints, based on randomly assigned cluster number. Distributions obtained for clusters where cells are randomly clustered are not any more similar to the distributions observed within the whole population of cells.

(E) Kolmogorov-Smirnov statistic comparing the shapes of nuclear mRNA, cytoplasmic mRNA and protein quantities for all the single cells used for clustering and shape of distribution obtained upon clustering (based on calculated medians for each of the used feature). Green line: values obtained when comparing clusters of randomly assigned single cells to the whole populations, grey line: values obtained when comparing clusters of SOM based clusters to the whole populations.

(F) Cross-correlation of trajectories obtained by SOM based clustering of single cells as described in (D) and (E)

(G) Hierarchical clustering of SOM-based clusters, based on median DNA content, cell size, and local cell density features. Deltas in nuclear mRNA production, cytoplasmic mRNA appearance and decay and protein appearance are plotted above the clustergram. Lines represent “rloess” fit, for purpose of visualization of trends in changes of delta with the change in cell size, and population context.

(H) Time delays for nuclear and cytoplasmic mRNA and cytoplasmic mRNA and protein for all the clusters. Clusters are same as represented in the Figure 4.

(I) Z-scored median quantities of nuclear mRNA, cytoplasmic mRNA and protein during the time of their expression. At all timepoints larger cells have higher amplitudes of response in all three species.

(J) mRNA count of HPRT1 does not change upon starvation and EGF stimulation.

(K) Representative images for *HPRT1* mRNA in HeLa cells at 0, 30, 40 and 60 minutes of EGF stimulation.