Genome-wide parallel quantification of mRNA and protein levels and turnover in mammalian cells

Björn Schwanhäusser[1], Dorothea Busse[1], Na Li[1], Gunnar Dittmar[1], Johannes Schuchhardt[2], Jana Wolf[1], Wei Chen[1], Matthias Selbach[1]

1 Max Delbrück Centrum for Molecular Medicine, Robert-Rössle-Str. 10, D-13092 Berlin, Germany

2 MicroDiscovery GmbH, Marienburger Str. 1, D-10405 Berlin, Germany

Correspondence:

Jana Wolf, Max Delbrück Centrum for Molecular Medicine, Robert-Rössle-Str. 10, D-13092 Berlin, Germany, Tel.: +49 30 9406 2641, Fax.: +49 30 9406 2394, email: jana.wolf@mdc-berlin.de

Wei Chen, Max Delbrück Centrum for Molecular Medicine, Robert-Rössle-Str. 10, D-13092 Berlin, Germany, Tel.: +49 30 9406 2995, Fax.: +49 30 9406 3068, email: wei.chen@mdc-berlin.de

Matthias Selbach, Max Delbrück Centrum for Molecular Medicine, Robert-Rössle-Str. 10, D-13092 Berlin, Germany, Tel.: +49 30 9406 3574, Fax.: +49 30 9406 2394, email: matthias.selbach@mdc-berlin.de

Running title: mRNA and protein levels and half-lives

Summary

Gene expression is a multistep process that involves transcription, translation and turnover of mRNAs and proteins. Although it is one of the most fundamental processes of life, the entire cascade has never been quantified on a genome-wide scale. Here, we simultaneously measured mRNA and protein abundance and turnover by parallel metabolic pulse labeling for more than 5,000 genes in mammalian cells. While mRNA and protein levels correlated better than previously thought, corresponding half-lives showed no correlation. Employing a quantitative model we obtain the first genome-scale prediction of synthesis rates of mRNAs and proteins. We find that the cellular abundance of proteins is predominantly controlled at the level of translation. Genes with similar combinations of mRNA and protein stabilities shared functional properties, suggesting that half-lives evolved under energetic and dynamic constraints. Quantitative information about all stages of gene expression obtained in this study provides a rich resource and helps understanding the underlying design principles.

Gelöscht: different

Introduction

The four fundamental cellular processes involved in gene expression are transcription, mRNA degradation, translation and protein degradation. It is now clear that each step of this cascade is controlled by gene-regulatory events[1-3]. While each individual process has been intensively studied[4-7], little is known about how the combined effect of all regulatory events shapes gene expression. The fundamental question of how genomic information is processed at different levels to obtain a specific cellular proteome has therefore remained unanswered. Genome-wide quantitative data about the flux of information from genes to proteins is not available for any organism.

Towards a quantitative description of gene expression numerous previous studies compared steady-state mRNA and protein levels and arrived at the conclusion that the correlation is poor[8]. However, the available data suffers from several limitations. First, most studies are limited to a few hundred genes, mainly due to the technical challenges involved in large scale protein identification and quantification. For example, the largest mammalian protein copy number dataset comprises only 512 genes [9]. Second, protein levels measured in one experiment are typically compared to mRNA levels determined in a different experiment performed at a different time in a different lab, making it difficult to interpret why the correlation is low. Third, mRNA levels are measured using microarrays which are less accurate than recent mRNA sequencing methods [10]. Fourth, many studies were performed in bacteria or yeast and thus do not represent regulatory mechanisms specific for higher eukaryotes. Finally, mRNA and protein levels result from coupled processes of synthesis and degradation. Therefore, analysis of mRNA and protein levels alone cannot provide sufficient information to understand gene expression comprehensively. mRNA and protein turnover can be measured with drugs to inhibit transcription or translation [11], but this has severe side-effects. Studies based on artificial fusion proteins are also problematic since tagging can affect protein stability [12].

To overcome these limitations we sought to quantify cellular mRNA and protein expression levels and turnover in parallel in a population of unperturbed mammalian cells. Pulse labeling with radioactive nucleosides or amino acids is regarded as the gold standard method to determine mRNA and protein half-lives [13]. Recently, variants of this approach based on non-radioactive tracers have been established [14-16]. In *stable isotope labeling by amino acids in cell culture* (SILAC) cells are cultivated in a medium containing heavy stable-isotope versions of essential amino acids [17]. When non-labeled (i.e. light) cells are transferred to heavy SILAC growth medium, newly synthesized proteins incorporate the heavy label while pre-existing proteins remain in the light form. This strategy can be used to measure protein turnover [18] or relative changes in protein translation [19]. Similarly, newly synthesized RNA can be labeled with the nucleoside analog 4-thiouridine (4sU). 4sU containing mRNA can be biotinylated and affinity purified. Comparing the newly synthesized and pre-existing fraction allows for global quantification of mRNA half-lives [16,20].

Parallel pulse labeling of proteins and mRNAs

We used parallel metabolic pulse labeling with amino acids and 4sU to simultaneously measure protein and mRNA turnover in a population of exponentially growing non-synchronized mouse fibroblasts (Fig. 1 A). Protein samples were harvested at three time points and analyzed by liquid chromatography and online tandem mass spectrometry (LC-MS/MS) on a high performance instrument (LTQ-Orbitrap-Velos). We identified and quantified proteins with the MaxQuant software package [21]. During five days of data acquisition we measured 1,471,375 fragment spectra that resulted in 229,985 peptide identifications (84,924 unique peptide sequences, false discovery rate (FDR) < 1%, see Supplementary Methods on ´Processing of mass spectrometry data´). These peptides were assigned to 6,445 unique proteins (FDR < 1%). 5,279 of these proteins were quantified by at least three heavy to light (H/L) peptide ratios (Fig. 1 B). Tissue-specific amino acid precursor pools and recycling rates, a pervasive problem for *in vivo* pulse labeling experiments[15,22], did not appreciably affect our results (Fig. S1). We also tested if protein synthesis rates are uniform over time. In case of constant incorporation rates the logarithm of H/L ratios should increase linearly with time (Fig. 1 C). 93 % showed excellent linear correlation indicated by a variability of the linear regression slope smaller than 1 % (two and three time point measurements, Fig. 1 D).

Thus, our data does not seem to be affected by non-uniform incorporation rates or by recycling. Also, protein abundance did not influence H/L ratio measurements (Fig. S 2). In total, we obtained a confident set of 5,028 protein half-lives calculated from the slope of the regression line (see Supplementary Methods). Cycloheximide-chase experiments for selected proteins spanning a representative range of half-lives agreed well with half-lives determined by pulsed labeling and mass spectrometry in all cases (Fig. 1 E).

In parallel, we pulse labeled newly synthesized RNA for 2 h with 4sU. RNA samples were fractionated into the newly synthesized and pre-existing fractions. Both fractions and the total unfractionated RNA sample were analyzed by mRNA sequencing on an Illumina Genome Analyzer. In total, we obtained 80,709,361 sequencing reads in all three samples, 55,046,553 (68%) of which could be mapped to the mouse genome. In all three samples, transcripts were quantified based on the number of reads mapped on their exonic region divided by transcript length and the total number of reads obtained [10]. We calculated mRNA half-lives based on the ratios of newly synthesized RNA/total RNA ratio and the preexisting RNA/total RNA using the previously published approach[16].

Importantly, this procedure compensates for different RNA yields during the fractionation process. To assess the reproducibility of protein and mRNA half-lives we performed an independent biological replicate of the entire large-scale experiment (see below).

Proteins were on average five times more stable (median half-live 46 h) than mRNAs (9 h) and spanned a bigger dynamic range (Fig. 2 A). Since very long (> 200 h) and very short (<30 min) protein half-lives cannot be accurately quantified from our three time points the true dynamic range of protein stabilities may be even higher (see Supplementary Methods). Intriguingly, we found no correlation between protein and mRNA half-lives (Fig. 2 C, $R^2 = 0.02$, $R_s = 0.16$, both at log-log scale). Thus, many stable proteins have unstable mRNAs and *vice versa*.

Absolute cellular mRNA and protein copy numbers

We calculated absolute cellular mRNA copy numbers based on the number of sequencing reads in the unfractionated sample in conjunction with information on cellular mRNA content [10]. Absolute protein copy numbers were inferred from mass spec data [23,24]. To this end, we used the sum of peak intensities of all peptides matching to a specific protein. When divided by the number of theoretically observable peptides, this value provides an accurate proxy for protein levels ('intensity-based absolute quantification' or iBAQ, see Supplementary Methods). As for half-lives, reproducibility of protein and mRNA copy numbers was assessed by performing an independent biological replicate (see below).

Levels of detected proteins spanned ~5 orders of magnitude (Fig. 2 B). Since relatively few proteins had less than 100 copies per cell we reasoned that some low abundant proteins escaped detection. Indeed, comparing mRNA levels of detected and not detected proteins revealed a moderate detection bias (Fig. S3). We therefore restricted our analysis to the set of genes that were identified at both the mRNA and protein level. In this subset, proteins were on average ~900 times more abundant than their corresponding transcripts. Despite a huge spread mRNAs and protein levels were clearly correlated (Fig. 2 D, $R^2 = 0.41$, $R_s = 0.62$, both at log-log scale). An attempt to further improve this correlation by non-linear transformation resulted only in a marginal increase (R^2=0.44, Fig. S4). It appears that for our data set, this is about the maximum correlation between RNA and protein that can be achieved without making use of additional information. This correlation is considerably higher than in any previous study in mammals[8,9]. For example, the recent study by Vogel and co-workers found an R^2 of 0.29 for a set of 512 mostly abundant proteins.

Our data therefore suggests that the often claimed poor correlation between mRNAs and proteins can partially be explained by non-parallel sample acquisition and/or imprecise measurements. Collectively, our data indicates that mRNA and protein levels correlate better than previously thought.

Reproducibility

To investigate the experimental noise we performed a second independent large-scale experiment and measured mRNA and protein levels and half-lives again. The overall correlation of half-lives and levels between both replicates was good (Fig. S5 and Supplementary Table 1 for more detailed error estimates).

To test if experimental noise affects the observed correlation between mRNAs and proteins we successively discarded genes with the highest variability between both replicates. For the remaining fraction we investigated correlation of mRNA and protein levels again. Removing less consistent data points did not increase correlation between mRNA and protein levels (Fig. 2 F). Similar results were obtained for half-lives (Fig. 2 E). Therefore, noise has little impact on the observed correlation between mRNA and protein levels and half-lives.

To exclude systematic errors we sought to quantify absolute mRNA and protein copy numbers using independent methods. For mRNA copy numbers we employed the NanoString technology which captures and counts individual transcripts without enzymatic reactions or bias[25]. Correlation between Illumina sequencing and NanoString data was high (R = 0.79, see also Fig S6 A). Absolute protein quantification was validated by spike-in experiments using a mixture of 48 proteins with known concentrations (Fig. S6 B). iBAQ values correlated well with known absolute protein amounts over at least four orders of magnitude and had a higher precision and accuracy than alternative measures of absolute protein abundance (data not shown)[23,24].

A quantitative model of gene expression allows genome-wide prediction of transcription and translation rates

Our data allows calculating average synthesis rates of mRNAs and proteins for thousands of genes employing a mathematical model (Fig. 3 A and Supplementary Methods). The experimental data is based on a population of non-synchronized cells. Therefore, our estimated rates provide an average over the population and time. They do not describe gene expression in single cells which requires single cell measurements[26].

Average cellular transcription rates predicted by the model spanned two orders of magnitude with a median of about two mRNA molecules per hour (Fig. 3 B). An extreme example was mdm2 with more than 500 mRNAs per hour, consistent with the extrachromosomal amplification of this gene in NIH3T3 cells. Since this is the first genome-scale estimate of mammalian transcription rates we cannot compare it with existing data. A microscopic study on the cytomegalovirus (CMV) promoter reported transcription termination rates of 5.8 to 8.7 mRNAs per hour [27]. These values are above the median of our predictions as perhaps expected for a rather strong promoter system.

Next, we calculated translation rate constants, i.e. how many proteins are made from each mRNA template per hour (Fig. 3 C). We find a median translation rate constant of about 40 proteins per mRNA per hour. Several proteins involved in translational regulation such as the translation initiation factor eIF4G, fragile X syndrome related protein Fxr2 and Tuberin had extremely low rate constants, i.e. were translationally repressed. Plotting translation rate constants against protein levels revealed that abundant proteins are translated about 100 times more efficiently than low abundant ones (Fig. 3 D). Hence, different translation efficiencies seem to contribute to the higher dynamic range of proteins compared to mRNAs (Fig. 2 B). Intriguingly, translation rate constants saturated at around 180 protein copies / (mRNA*h). This is unlikely a signal saturation artifact since we did not observe dynamic range compression of protein levels (Fig. S6 B). Alternatively, the observation can be interpreted as a maximal translation rate constant. To our knowledge, the maximal translation rate constant in mammals is not known. Based on Davidson and co-workers the estimated maximal translation rate constant in sea urchin embryos is 140 copies / (mRNA*h)[1] which is surprisingly close to the prediction of our model for mouse fibroblasts. We also assessed degradation and synthesis rates for mRNAs and proteins by actinomycin D and cycloheximide treatment, respectively (Supplementary Methods). For high turnover proteins and mRNAs we obtained results consistent with pulse labeling data (Fig. S6 C-F).

Gelöscht: estimated

Gelöscht: in sea urchin embryos

Estimating the impact of post-transcriptional, translational and post-translational control on protein abundance

A long standing question is how much protein abundance is controlled at the transcriptional, post-transcriptional, translational and post-translational level. Until now, this has mainly been addressed indirectly by analyzing mRNA and protein sequence features. For example, features related to translation initiation (e.g. Shine-Dalgarno, Kozak and 3' UTR sequences), elongation (e.g. codon bias) and protein stability (e.g. degrons) have been analyzed and reported to partially correlate with protein/mRNA ratios in bacteria, yeast and mammals[9,28]. We also observed sequence features characteristic of mRNA and protein stability and found that mRNAs with long 3' UTRs are on average less stable (Fig. S7). In addition, the density of AU-rich elements (AREs) and binding motifs of specific RNA-binding proteins (Pumilio2) correlated negatively with mRNA stability (Fig. S8). Moreover, we observed that intrinsically unstructured proteins tend to have shorter half-lives, and we identified amino acids overrepresented in unstable proteins (Fig. S9).

Sequence features are at best indirect proxies for the regulatory mechanisms controlling protein abundance. How much efficiencies of different steps in the gene expression cascade contribute to variance of cellular protein copy numbers can only be revealed by direct parallel genome-scale measurements of mRNA and protein levels and half-lives which were not available previously. In our data the coefficient of determination (R^2) between mRNA and protein copy numbers was 0.41 (Fig. 2). If we assume absence of technical and biological noise, this means that ~40% of the variance in protein levels are explained by different mRNA levels – considerably more than previously thought (Fig. 4 A). Most of these 40% are due to different transcription rates while mRNA stability plays a smaller role. Considering translation rate constants dramatically boosts R^2 to 0.95 and thus the correlation to 95%. Although this is an over-fit (see below), the analysis shows that translation rate constants play the dominant role for control of protein levels. Unexpectedly, the impact of protein degradation is rather small.

In the above analysis the same experimental data was used to calculate synthesis rates and to estimate their impact on protein levels. To avoid this over-fit and to assess reliability of the model predictions we performed the same analysis with data from the biological replicate experiment. In this replicate experiment the coefficient of determination between mRNA and protein levels was 0.37 (Fig. 4 B). We then used the model including the estimated parameters from the first experiment to predict protein levels from mRNA levels in the replicate data. Predicted protein levels agreed very well with measured protein levels (R^2 = 0.85, Fig. 4 C). Therefore, the model explains ~85% of the variability in protein copy numbers in an independent experiment. The correlation is very similar to the direct comparison of protein levels in both experiments (R^2 = 0.84, Fig. S5 D). We conclude that (*i*) technical and biological noise in our data is low and that (*ii*) the model faithfully predicts protein levels from mRNA levels in mouse fibroblasts. It also indicates that the estimated impact of transcription, mRNA stability, translation and protein stability on protein abundance is reproducible. In the replicate experiment mRNA levels and translation rate constants combined can explain 75% of the variability in protein levels. We also assessed how much of the efficiencies of the various steps in gene expression are retained in a different cell type and organism. To this end, we quantified mRNA and protein abundance in the human breast cancer cell line MCF7 by RNA-seq and mass spectrometry, respectively. 2,030 human genes from the MCF7 dataset had orthologs in the mouse fibroblast data. We then used rates from the mouse fibroblast model to predict protein levels from mRNA levels in human breast cancer cells. In MCF7 cells, the model predicted ~60% of the variability in protein levels (Fig. 4 A).

Although the fraction explained by the model is smaller than in mouse fibroblasts, this indicates that translation and degradation rates are to some extent independent of the cell type and conserved between mouse and human.

Genes with similar combinations of mRNA and protein half-lives share functional properties

It is well-known that degradation of proteins is critically involved in many cellular processes including cell cycle progression, signal transduction and apoptosis [7]. Similarly, mRNA stability is important for the temporal order of gene induction[20,29,30]. Genes may have evolved specific combinations of mRNA and protein half-lives under functional constraints[20,29-31]. We therefore asked if genes with specific combinations of mRNA and protein stability have distinct biological functions. We grouped genes according to their combinations of mRNA and protein half-lives and used gene ontology to find enriched biological processes (Fig. 5 A, see Supplementary Table 2 for a complete list of GO terms with Benjamini-Hochberg adjusted p-values).

Genes with stable mRNAs and stable proteins were enriched in constitutive cellular processes like translation (i.e. ribosomal proteins), respiration and central metabolism (glycolysis, citric acid cycle). Hence, many 'house-keeping' genes tend to have stable mRNAs and proteins. In lower organisms, energy costs keep transcription and translation rates under selective pressure[32]. We therefore reasoned that energy constraints may explain why 'housekeeping' genes tend to have stable mRNAs and proteins. Based on the model, we calculated the theoretical energy required to maintain cellular mRNAs and protein levels by recycling from their building blocks (nucleotide monophosphates and amino acids, respectively) in terms of high energy phosphates. This scenario corresponds to non-dividing cells in which the overall amount of mRNAs and proteins stays constant. Therefore, the metabolic cost of synthesizing amino acids and nucleotides is not considered. mRNA synthesis costs were calculated for primary transcripts (i.e. including introns). The calculation is a conservative estimate since the energy needed for splicing, folding, transport etc. is not known and therefore not included. We found that protein synthesis consumes more than 90% of the energy while less than 10% is needed for transcription. 80% of the energy for translation is required to synthesize 20% of all proteins. Hence, protein synthesis follows the Pareto principle ("80/20 rule") with a small fraction of proteins consuming most of the energy. If gene expression was optimized under energetic constraints abundant proteins are expected to be more stable than less abundant ones. This was indeed the case (Fig. 5 B, $p<10^{-15}$, Wilcoxon test). This is not necessarily expected since the overall contribution of protein stability to protein levels is very small (Fig. 4 A). Consistent with the energy constraint abundant proteins were also significantly shorter (Fig. 5 C). Shuffling protein half-lives and lengths markedly increased theoretical energy consumption (Fig. 5 D). Similar results are obtained for mRNAs but their impact on overall costs is small. Collectively, these observations are consistent with the idea that mammalian gene expression evolved under energy constraints.

The subset of genes with unstable mRNAs and proteins was strongly enriched in transcription factors, signaling genes, chromatin modifying enzymes and genes with cell cycle-specific functions (Fig. 5 A). Thus, many regulatory genes have low mRNA and protein half-lives. Since mRNAs and proteins are information carriers, their degradation can be interpreted as a built-in timer which controls persistence of genetic information[33]. Transcription of genes with short mRNA and protein half-lives has therefore only a short-term impact on the protein level. In this scenario it makes intuitive sense that many transcription factors, cell cycle genes and chromatin modifiers have short mRNA and protein half-lives. However, it must be stressed that our data cannot provide information about individual cells or molecules and should only be interpreted at the cell population level.

The group of genes with stable proteins but unstable mRNAs was strongly enriched in terms related to processing of mRNAs, tRNAs and non-coding RNAs. This shows that many mammalian RNA-binding proteins are stable while their encoding transcripts tend to be short-lived, as noted recently for yeast[34]. Since many RNA-binding proteins bind their own message[35], this observation is indicative of a post-transcriptional negative feedback-loop for RNA-binding proteins. Consistently, we found that unstable mRNAs are enriched for binding motifs of RNA-binding proteins (Fig. S8).

Finally, the subset of genes with stable mRNAs and unstable proteins was rich in extracellular proteins. This is expected, since secreted proteins have a short cellular half-life. Additionally, this group contains proteins involved in cellular homeostasis, defense response and proteolysis. For example, this set contains two ferritin proteins which are rapidly up-regulated in response to iron[36]. Interestingly, ferritins are text book examples of translationally regulated genes. Since translational regulation is not dependent on mRNA half-lives, genes with stable mRNAs can still be dynamically regulated as long as their protein half-lives are short. It is tempting to speculate that other homeostasis genes in this group are regulated at the level of translation.

Discussion

Although gene expression is one of the most fundamental processes in biology it has never been quantified comprehensively. While it is now clear that regulation occurs at multiple levels, the flow of information from genes to proteins has not yet been investigated on a genome-wide scale[2,6]. Here, we used parallel metabolic pulse labeling, mass spectrometry and next generation sequencing to provide the first analysis of mRNA and protein levels and half-lives for thousands of genes. We also report the first estimate of average transcription and translation rate constants predicted from our data. Our work provides a first global overview of mammalian gene expression dynamics from beginning (transcription) to the end (protein degradation). We provide novel insights on the steps that control protein abundance and shed new light on the underlying design principles. In the future, additional methods like sequencing of nascent transcripts and ribosome profiling may further refine this picture[37].

We found that mRNA levels explain around 40% of the variability in protein levels. This fraction is higher than in any previous study on mammals and does not seem to be affected by technical noise (Fig. 2 F). In yeast, mRNA and protein levels show a much higher correlation with mRNA levels already explaining 73% (R^2=0.73) of the variability in protein levels[8,9,23]. One reason may be that higher eukaryotes show a higher degree of translational and post-translational regulation. We found that in mouse fibroblasts translation efficiency is the single best predictor of protein levels. Hence, protein abundance seems to be predominantly regulated at the ribosome, highlighting the importance of translational control for gene expression[5]. Whether this observation is valid in other cell types is not known. A recent study on embryonic stem cells revealed that changes in protein levels are not accompanied by changes in corresponding mRNAs, although this study did not discern translational and post-translational control[38]. It is also not clear how much translation rate constants change under different conditions. In fact, our observation that the mouse model can to some degree predict levels of orthologous proteins in MCF7 cells suggests that translation efficiency is partially ‘hard-coded’ in the genome and not subject to change.

Compared to translational control, protein stability seems to play a minor role for cellular protein abundance in our system. This might be surprising since protein degradation is involved in regulation of many cellular processes such as cell cycle progression[7]. From the global perspective, the dominance of translational regulation makes sense given the high energy costs associated with protein synthesis. Interestingly, the study by Maier and co-workers on a model bacterium comes to similar conclusions (see accompanying submission by Maier et al., 2010). However, it should also be stressed that our dataset represents average values derived from a population of dividing, non-synchronized cells. At the single cell level, the role of protein degradation for protein abundance may be higher. Similarly, protein degradation may be more important upon perturbation.

Gene expression may follow certain design principles for optimal evolutionary fitness. Intriguingly, we found that genes with certain combinations of mRNA and protein half-lives share common functions, suggesting they evolved under similar constraints. One of these constraints may be energy efficiency[32]. Consistently, we observed that the theoretical energy needed for gene expression is much lower than random. A second constraint may be the ability of genes to respond quickly to a stimulus. We find that many transcription factors and genes with cell-cycle specific function have unstable mRNAs and proteins, predisposing them to rapid transcriptional and/or translational regulation. In addition, genes with stable mRNAs but unstable proteins can be regulated quickly at the level of translation. These observations are consistent with the idea that many fast responding genes have short protein and/or mRNA half-lives[20,30,31,39]. The global picture is that most mRNAs and especially proteins tend to be stable unless genes need to respond quickly to a stimulus. Due to the trade-off between dynamic regulation and energy efficiency this may be an optimal design. Another design principle emerges from the striking observation that many mammalian RNA-binding proteins are stable but encoded by unstable transcripts, as also seen in yeast[34].

Finally, our data provides a rich resource for the scientific community that can be mined in many ways that are beyond the scope of this study. For example, we provide by far the largest dataset on protein copy numbers which contains valuable information for modeling of cellular processes and stoichiometry of protein complexes[24]. Half-lives of proteins and mRNAs can be used to search for properties of unstable mRNAs or proteins, and we provide a first analysis of characteristic sequence features (Fig. S7 and S8). Genome-scale quantitative data on absolute mRNA and protein levels and half-lives will certainly help to understand the complex relationships between thousands of genes and their products in biological systems.

Acknowledgements

We thank Nikolaus Rajewsky and Lars Dölken for fruitful discussions and Christian Sommer for excellent technical assistance. This work was supported by funding from the Helmholtz Association and the German Ministry of Education and Research (BMBF) to MS. WC is supported by funding for the Berlin Institute of Medical Systems Biology by the BMBF and the Senate of Berlin [grant number 315362A]. JW is supported by the ForSys-programme of the German Ministry of Education and Research [grant number 0315289], DB by the Helmholtz Alliance on Systems Biology, MSBN and NL by the China Scholarship Council CSC.

Availability of the data

The data is freely available from the authors upon request.

Author contributions

MS conceived, designed and supervised the experiments. BS performed wet-lab experiments, mass spectrometry and proteomic data analysis. DB and JW developed and employed the mathematical model. NL performed RNA-seq experiments. WC designed and supervised RNA-Seq experiments. BS, DB, JS, WC and MS analyzed genome-wide data. GD helped in cycloheximide chase experiments and data analysis. BS, DB, JS, JW, WC and MS interpreted the data. MS wrote the manuscript.

Figure legends

Fig. 1: Parallel quantification of mRNA and protein turnover and levels. (A) Mouse fibroblasts were pulse labeled with heavy amino acids (SILAC, left) and the nucleoside 4-thiouridine (4sU, right). Protein and mRNA turnover was quantified by mass spectrometry and next generation sequencing, respectively. **(B)** Mass spectra of peptides from a high and low turnover protein reveal increasing heavy to light (H/L) ratios over time. **(C)** Protein half-lives were calculated from log H/L ratios at all three time points using linear regression. **(D)** Variability of linear regression slopes assessed by leave-one-out cross validation was small. **(E)** Comparison of protein half-lives measured by SILAC and traditional cycloheximide-chase experiments.

Formatiert: Schriftart: Nicht Fett

Formatiert: Schriftart: Nicht Fett

Gelöscht: Most proteins showed an excellent fit.

Fig. 2: mRNA and protein levels and half-lives. Histograms of mRNA (blue) and protein (red) half-lives **(A)** and levels **(B)**. Proteins were on average 5 times more stable and 900 times more abundant than mRNAs and spanned a higher dynamic range. While mRNA and protein levels correlated significantly, correlation of half-lives was virtually absent **(C,D)**. Consecutive removal of genes with highest deviation between biological replicates did not significantly increase correlations of mRNA and protein half-lives **(E)** or levels **(F)**.

Fig. 3: Quantitative model of gene expression in growing cells (A) mRNAs are synthesized with the rate v_{sr} and degraded with a rate constant k_{dr}. Proteins are translated and degraded with rate constants k_{sp} and k_{dp}, respectively. **(B)** Calculated mRNA transcription rates show a uniform distribution. **(C)** Calculated translation rate constants are not uniform. **(D)** Translation rates of abundant proteins saturate between approx 120 and 240 proteins/(mRNA*h). Red line shows the locally weighted fit (LOWESS). Dashed lines indicate 95% confidence intervals of the LOWESS maximum value calculated by bootstrapping.

Fig. 4: Impact of regulation at different levels on protein abundance

Gelöscht: abundance

(A) According to the model, protein levels are best explained by translation rates, followed by transcription rates. mRNA and protein stability is less important (left bar). **(B)** In a second, independent biological experiment mRNA levels explained 37% of protein levels in NIH3T3 cells (middle bar in A). **(C)** Using the model to predict protein levels from measured mRNA levels boosts predictive power to 85% (middle bar in A). The mouse fibroblast model can to some extent predict protein levels from mRNA levels of human orthologs in MCF7 cells (right bar in A). Error bars show 95% confidence intervals estimated by bootstrapping.

Fig. 5: Functional characteristics of genes with different mRNA and protein half-lives (A) Genes were grouped according to their combination of mRNA and protein half-lives and analyzed for enriched gene ontology terms. A heat map of enrichment p-values reveals functional similarities of genes with similar combinations of half-lives. **(B, C)** Abundant proteins are significantly more stable and shorter than less abundant ones ($p<10^{-15}$, Wilcoxon test). **(D)** Theoretical energy consumption of gene expression. Randomizing protein half-lives or lengths enhances energy costs. Error bars show 95% confidence intervals determined by multiple randomizations and bootstrapping.

References

[1] Ben-Tabou de-Leon, S. and Davidson, E. H., *Dev Biol* **325** (2), 317 (2009).

[2] Komili, S. and Silver, P. A., *Nat Rev Genet* **9** (1), 38 (2008).

[3] Alon, U., *Nat Rev Genet* **8** (6), 450 (2007).

[4] Kim, H. D., Shay, T., O'Shea, E. K., and Regev, A., *Science* **325** (5939), 429 (2009); Kouzarides, T., *Cell* **128** (4), 693 (2007); Ambros, V., *Nat Med* **14** (10), 1036 (2008).

[5] Gebauer, F. and Hentze, M. W., *Nat Rev Mol Cell Biol* **5** (10), 827 (2004); Sonenberg, N. and Hinnebusch, A. G., *Cell* **136** (4), 731 (2009).

[6] Keene, J. D., *Nat Rev Genet* **8** (7), 533 (2007).

[7] Mayer, R. J., *Nat Rev Mol Cell Biol* **1** (2), 145 (2000); Kirkpatrick, D. S., Denison, C., and Gygi, S. P., *Nat Cell Biol* **7** (8), 750 (2005); Elsasser, S. and Finley, D., *Nat Cell Biol* **7** (8), 742 (2005); Hershko, A. and Ciechanover, A., *Annu Rev Biochem* **67**, 425 (1998); King, R. W., Deshaies, R. J., Peters, J. M., and Kirschner, M. W., *Science* **274** (5293), 1652 (1996).

[8] de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M., and Vogel, C., *Mol Biosyst* **5** (12), 1512 (2009); Maier, T., Guell, M., and Serrano, L., *FEBS Lett* **583** (24), 3966 (2009).

[9] Vogel, C. et al., *Mol Syst Biol* **6**, 400 (2010).

[10] Mortazavi, A. et al., *Nat Methods* **5** (7), 621 (2008).

[11] Belle, A. et al., *Proc Natl Acad Sci U S A* **103** (35), 13004 (2006); Raghavan, A. et al., *Nucleic Acids Res* **30** (24), 5529 (2002); Yang, E. et al., *Genome Res* **13** (8), 1863 (2003).

[12] Yen, H. C. et al., *Science* **322** (5903), 918 (2008).

[13] Kenney, F. T., *Science* **156** (774), 525 (1967); Puckett, L., Chambers, S., and Darnell, J. E., *Proc Natl Acad Sci U S A* **72** (1), 389 (1975).

[14] Gouw, J. W., Krijgsveld, J., and Heck, A. J., *Mol Cell Proteomics* **9** (1), 11 (2010).

[15] Beynon, R. J. and Pratt, J. M., *Mol Cell Proteomics* **4** (7), 857 (2005).

[16] Dolken, L. et al., *RNA* **14** (9), 1959 (2008).

[17] Mann, M., *Nat Rev Mol Cell Biol* **7** (12), 952 (2006).

[18] Doherty, M. K. et al., *J Proteome Res* **8** (1), 104 (2009); Milner, E., Barnea, E., Beer, I., and Admon, A., *Mol Cell Proteomics* **5** (2), 357 (2006); Pratt, J. M. et al., *Mol Cell*

Proteomics **1** (8), 579 (2002); Lam, Y. W., Lamond, A. I., Mann, M., and Andersen, J. S., *Curr Biol* **17** (9), 749 (2007).

[19] Schwanhausser, B., Gossen, M., Dittmar, G., and Selbach, M., *Proteomics* **9** (1), 205 (2009); Selbach, M. et al., *Nature* **455** (7209), 58 (2008).

[20] Friedel, C. C. et al., *Nucleic Acids Res* **37** (17), e115 (2009).

[21] Cox, J. and Mann, M., *Nat Biotechnol* **26** (12), 1367 (2008).

[22] Price, J. C. et al., *Proc Natl Acad Sci U S A* **107** (32), 14508 (2010); Wu, C. C. et al., *Anal Chem* **76** (17), 4951 (2004); Hellerstein, M. K., *Metab Eng* **6** (1), 85 (2004).

[23] Lu, P. et al., *Nat Biotechnol* **25** (1), 117 (2007).

[24] Malmstrom, J. et al., *Nature* **460** (7256), 762 (2009).

[25] Geiss, G. K. et al., *Nat Biotechnol* **26** (3), 317 (2008).

[26] Rosenfeld, N. et al., *Science* **307** (5717), 1962 (2005); Geva-Zatorsky, N. et al., *Cell* **140** (5), 643 (2010).

[27] Darzacq, X. et al., *Nat Struct Mol Biol* **14** (9), 796 (2007).

[28] Arava, Y., Boas, F. E., Brown, P. O., and Herschlag, D., *Nucleic Acids Res* **33** (8), 2421 (2005); Wu, G., Nie, L., and Zhang, W., *Curr Microbiol* **57** (1), 18 (2008).

[29] Elkon, R., Zlotorynski, E., Zeller, K. I., and Agami, R., *BMC Genomics* **11**, 259.

[30] Hao, S. and Baltimore, D., *Nat Immunol* **10** (3), 281 (2009).

[31] Legewie, S., Herzel, H., Westerhoff, H. V., and Bluthgen, N., *Mol Syst Biol* **4**, 190 (2008).

[32] Wagner, A., *Mol Biol Evol* **22** (6), 1365 (2005).

[33] Pedraza, J. M. and Paulsson, J., *Science* **319** (5861), 339 (2008).

[34] Mittal, N., Roy, N., Babu, M. M., and Janga, S. C., *Proc Natl Acad Sci U S A* **106** (48), 20300 (2009).

[35] Hogan, D. J. et al., *PLoS Biol* **6** (10), e255 (2008).

[36] Hentze, M. W., Muckenthaler, M. U., and Andrews, N. C., *Cell* **117** (3), 285 (2004).

[37] Core, L. J., Waterfall, J. J., and Lis, J. T., *Science* **322** (5909), 1845 (2008); Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., and Weissman, J. S., *Science* **324** (5924), 218 (2009); Churchman, L. S. and Weissman, J. S., *Nature* **469** (7330), 368 (2011).

[38] Lu, R. et al., *Nature* **462** (7271), 358 (2009).

[39] Rosenfeld, N., Elowitz, M. B., and Alon, U., *J Mol Biol* **323** (5), 785 (2002).