Pervasive discordance between mRNA and protein expression during embryonic stem cell differentiation

Patrick R. van den Berg[1], Bogdan Budnik[2], Nikolai Slavov[3,*], Stefan Semrau[1,*,†]

[1] Leiden Institute of Physics, Leiden University, Leiden, Zuid-Holland, 2333 CC, The Netherlands

[2] Mass Spectrometry and Proteomics Resource Laboratory, Harvard University, Cambridge, Massachusetts, MA 02138, USA

[3] Department of Bioengineering, Northeastern University, Boston, Massachusetts, MA 02115,

[*]corresponding author, †, lead contact

Summary

During *in vitro* differentiation, pluripotent stem cells undergo extensive remodeling of their gene expression. While studied extensively at the transcriptome level, much less is known about protein dynamics, which can differ significantly from their mRNA counterparts. Here, we present genome-wide dynamic measurements of mRNA and protein levels during differentiation of embryonic stem cells (ESCs). We reveal pervasive discordance, which can be largely understood as a dynamic imbalance due to delayed protein synthesis and degradation. Through a combination of systematic classification and kinetic modeling, we connect modes of regulation at the protein level to the function of specific gene sets in differentiation. We further show that our kinetic model can be applied to single-cell transcriptomics data to predict protein levels in differentiated cell types. In conclusion, our comprehensive data set, easily accessible through a web application, is a valuable resource for the discovery of protein-level regulation in ESC differentiation.

Keywords

embryonic stem cells, *in vitro* differentiation, gene regulation, transcriptomics, proteomics, kinetic modeling, integration with single-cell transcriptomics, web application for data exploration

Introduction

Much of the medical potential of pluripotent stem cells is due to their ability to differentiate *in vitro* into all tissue types of the adult body (Soldner and Jaenisch, 2012). While tremendous progress has been made in guiding cells through successive lineage decisions, the gene regulatory mechanisms underlying these decisions remain largely unknown. This gap in knowledge hampers the streamlining and acceleration of differentiation protocols. A large body of work has focused on transcriptional regulation, charting transcriptome changes during differentiation, most recently down to the single-cell level (Klein et al., 2015; Loh et al., 2016) (Semrau et al., 2016). These studies assumed implicitly that mRNA levels are a good proxy for protein levels. Mounting evidence suggests that this is not a good assumption for mammalian systems, where mRNA and protein levels were found to correlate only moderately (Lu et al., 2009) (Kristensen et al., 2013; Peshkin et al., 2015; Schwanhäusser et al., 2011). Where the discordance between protein and mRNA expression originates and what the biological function might be are long-standing and controversially discussed issues (Liu et al., 2016; Vogel and Marcotte, 2012). Here we study the relationship between mRNA and protein expression in the context of *in vitro* differentiation, a highly dynamic process in which gene regulation at the protein level likely plays an important role (Sampath et al., 2008).

Results

Measurement of transcriptome and proteome dynamics during retinoic acid driven differentiation

We used retinoic acid (RA) differentiation of mESCs as a generic model for *in vitro* differentiation. Previously, we characterized this differentiation assay in detail at the transcriptional level by single-cell RNA-seq (Semrau et al., 2016). In particular, we have shown that within 96 h of RA exposure, mESCs bifurcate into an extraembryonic endoderm-like and an ectoderm-like cell type (XEN and ECT respectively). Here we collected samples of the mixed population during an RA differentiation time course as well as the two final, FACS-purified differentiated cell types at 96 h (Fig. 1a). For each time point or cell type we quantified poly(A) RNA by RNA-seq and protein expression by tandem mass tag (TMT) labeling followed by tandem mass spectrometry (MS/MS). In total we obtained both RNA and protein expression for 7459 genes (Supplementary Fig. 1a). Protein levels were quantified with low technical error (Supplementary Fig. 1a) and high reproducibility between protein fold changes measured in biological replicates (Pearson's r = 0.92, Supplementary Fig. 1b). Moreover, the XEN-like cells measured here were similar to embryo derived XEN cells in their proteome (Mulvey et al., 2015) (r = 0.65, Supplementary Fig. 1c).

Correlation between mRNA and protein levels is moderate

To explore the relationship between mRNA and protein levels we first correlated the two expression levels across genes for individual time points or cell types (sample-wise correlation). In mESCs (0h time point) Pearson correlation between mRNA and protein was 0.57 (Fig. 1b). Similar values have been reported in other mammalian systems (de Sousa Abreu et al., 2009; Jovanovic et al., 2015; Schwanhäusser et al., 2011). Sample-wise correlation was approximately the same for all samples, including the purified differentiated cell types (Fig. 1c). Low mRNA-protein correlation was thus not cell state dependent. Importantly, a low sample-wise correlation does not exclude the possibility that relative changes in protein levels during differentiation closely follow relative changes in mRNA levels. To quantify the concordance between mRNA and protein dynamics we calculated their correlation across time for individual genes (gene-wise correlation, Fig. 1d-e). Some genes, like the pluripotency factor *Rex1* (*Zfp42*) indeed exhibited a high correlation between mRNA and protein (r = 0.93 for *Rex1*). However, we also observed many genes with anti-correlated profiles, among which were genes with a role in differentiation, like the transcription factor *Foxn2* (r = -0.85). Numerous genes, like the ribosomal protein *Rps6*, for example, did not exhibit any strong correlation between protein or mRNA (r = 0 for *Rps6*). Correspondingly, the distribution of gene-wise correlations, while peaking close to 1, had a long tail towards -1 (Fig. 1e). This result clearly shows that mRNA dynamics are in general not a good predictor for protein dynamics during differentiation.

Classification by dominant temporal trends visualizes widespread discordance between mRNA and protein

Having discovered that mRNA and protein dynamics are in general dissimilar we wanted to reveal the main trends in expression dynamics and study how they differ between mRNA and protein. To that end we used singular value decomposition (SVD) to decompose an expression profile into a weighted sum of generic profiles, called eigengenes (Fig. 2a). In contrast to other classification methods, SVD allows us to discriminate systematically between the main trend (the dominant eigengene) and smaller, additional fluctuations (Fig. 2b). The first three eigengenes, which corresponded to monotonic, transient or oscillatory trends, explained 76% and 85% of the variance in mRNA and protein expression, respectively (Fig. 2c). mRNA eigengenes were more dynamic than protein eigengenes (Supplementary Fig. 1d), which reflects the buffering of mRNA dynamics by protein synthesis and degradation (Liu et al., 2016) (Jovanovic et al., 2015). Classification of all genes by their dominant mRNA and protein eigengenes (which reflect the main temporal trends) revealed widespread discordance (Fig. 2d). While there was a statistically significant enrichment of genes with similar dominant mRNA and protein eigengenes (p-value < 1E-5), the majority of genes (60%) had discordant mRNA and protein dynamics.

A simple kinetic model explains the mRNA-protein discordance for the majority of genes

The temporal delay between mRNA and protein eigengenes (Fig. 2a) sparked the hypothesis that the delay inherent to protein synthesis and degradation might cause much of the observed discordance. To pursue this hypothesis we modeled protein turnover using a simple birth-death process with constant protein synthesis and degradation rates (Tchourine et al., 2014) (Peshkin et al., 2015) (Methods, Fig. 3a). In our model the synthesis rate k_s lumps all processes related to protein production (translation initiation, elongation, etc.) while the degradation rate k_d represents all processes leading to a reduction in protein levels (dilution due to cell division, active degradation, etc.). To avoid over-fitting, we also considered simpler models, which correspond to cases in which a protein is only synthesized, only degraded or completely constant (Fig. 3b). To select among these models, we employed the Bayesian Information Criterion (BIC), a score that penalizes the fit according to the number of parameters (Methods). To reveal whether there is a connection between a certain model and specific molecular functions, we performed GO term enrichment analysis. This analysis revealed that the "degradation only" model was enriched for genes with a role in blastocysts development and inner cell mass proliferation (Supplementary Fig. 2a). These genes are likely involved in preserving the pluripotent state, as exemplified by the pluripotency factor *Nanog*. Degradation of the corresponding proteins is crucial for the timely exit from pluripotency. GO term enrichment analysis also showed that the "synthesis only" model was enriched for genes involved in neuron development and mesenchymal cell development. These genes thus likely have specific functions in differentiated cell types and hence must be synthesized quickly to ensure proper function. An example of such a gene is *Lamb1*, which is highly expressed in XEN cells. This analysis shows that the different regulatory modes identified by our model correspond to specific functions in differentiation.

We next wanted to evaluate the validity of our model by comparison with relevant data sets from the literature. Protein half-lives (Supplementary Fig. 2b) calculated from the degradation rates were in the same range as previously reported values for other systems (Peshkin et al., 2015; Schwanhäusser et al., 2011). Synthesis rates were positively correlated with translational efficiencies determined from ribosome profiling in mESCs (Supplementary Fig. 2c) (Ingolia et al., 2011). The inferred kinetic rates are thus biologically meaningful.

In order to assess how far our kinetic model can explain the observed protein-mRNA discordance we calculated the correlation between measured and predicted protein levels (Fig. 3c). These correlations were sharply peaked close to one, which means that our simple model is able to explain a large portion of the observed mRNA-protein discordance. This discordance is likely only transient since protein-to-mRNA ratios differed most from their equilibrium value ($k_{eq} = k_s/k_d$) in the beginning but approached it over time (Supplementary Fig. 2d). This observation supports our conclusion that the observed mRNA-protein discordance during differentiation is largely a transient, dynamic imbalance caused by delayed protein synthesis and degradation.

The CDS/ 3'UTR mRNA expression ratio is a modulator of the synthesis rate

We next sought to further refine our kinetic model and explore whether we could find predictors of protein abundance. In that respect we were intrigued by a recent report that connected the ratio of mRNA expression from the coding sequence (CDS) and 3' untranslated region (UTR) to protein abundance (Kocabas et al., 2015). In our data sets, the CDS/3'UTR mRNA expression ratio *w* also had a non-trivial relationship with protein levels (Supplementary Fig. 2e). Consequently, we included *w* in our model as a modulator of the synthesis rate (Fig. 3d, Methods). Again, using the BIC to determine whether using an additional free parameter is warranted by the improvement of the fit, we found that 492 genes were fit optimally by the extended kinetic model (Fig. 3e). In the cases where it was optimal the extended model provided a substantial improvement over the basic model (Fig. 3f). For roughly half of those genes, *w* has a positive effect on protein synthesis and a negative effect on the other half (Supplementary Fig. 2f). While the molecular mechanism relating *w* to the protein synthesis rate is not yet known, our analysis shows that *w* is an interesting predictor that should be explored in future studies of protein dynamics.

Failure of the kinetic model reveals dynamically regulated genes

Despite its success in explaining the mRNA-protein discordance overall, our kinetic model does not fit the dynamics of all quantified proteins. We identified 1232 genes with a poor mRNA-protein correlation that is not appreciably improved by any of the kinetic models (Supplementary Fig. 3a). Due to the buffering of mRNA dynamics when synthesis and degradation rates are constant, the model fails in particular when the protein profile is more dynamic than the mRNA profile (Supplementary Fig. 3b). Importantly, the genes that are not fit well by our model are very similar to the full data set in their protein reliabilities (medians: 0.970 versus 0.972) and measurement errors (median SEM: 0.121 versus 0.115). Hence, technical noise is in general not the reason for the lack of a good fit. Rather, the model fails due to the assumption that kinetic rates are constant. Consequently, we consider genes that are not fit well by the model to be dynamically regulated. We sought to find sets of such genes that potentially share regulatory features. To this end we again used the classification by dominant eigengenes (Supplementary Fig. 3c). As an example, we focused on a class of genes with relatively simple dynamics: monotonically increasing mRNA and a transient increase in protein expression (highlighted in Supplementary Fig. 3c). Notably, we discovered that genes belonging to the MAPK pathway were enriched in this particular class (ConsensusPathDB, adjusted p-value = 1.8E-3, Supplementary Fig. 3d). This suggests that genes of the MAPK pathway, which is highly relevant for the differentiation of mESCs (Kunath et al., 2007), are regulated dynamically at the protein level. This analysis exemplifies that we can systematically identify sets of genes that are dynamically regulated at the protein level, likely by common mechanisms.

Sets of genes with different functions in differentiation show distinct regulatory modes

We next wanted to concentrate further on the regulation of gene sets that are relevant for embryonic stem cell differentiation. To that end, we defined sets of markers for the pluripotent state, XEN cells, and ECT cells based on differential mRNA expression (Supplementary Fig. 4a), which were confirmed by GO term enrichment (Supplementary Fig. 4b). As a fourth gene set we considered ribosomal proteins since it has been shown previously that the translational state changes dramatically during differentiation (Sampath et al., 2008). For these 4 gene sets we calculated the average mRNA and protein profiles, correlation between mRNA and protein, classification by dominant eigengene and inferred synthesis and degradation rates for the genes that are fit optimally by the full kinetic model (Fig. 4a). This analysis of gene sets is also available on the companion website. Pluripotency markers were in general down regulated at the mRNA level (per definition) but also at the protein level. Correspondingly, we found this set to be enriched in the “degradation only” kinetic model while the “synthesis only” model is underrepresented (Supplementary Fig. 4c). This observation is consistent with the fact that pluripotency genes have to be down-regulated quickly to allow for a timely exit from pluripotency. Nevertheless, there were some genes that showed a substantial increase in protein expression and consequently had a negative correlation between measured mRNA and protein (see Supplementary Fig. 4d for examples). XEN and ECT markers were in general upregulated, where ECT markers came up before XEN markers, as shown by us previously (Semrau et al., 2016). In contrast to the set of pluripotency markers, XEN and ECT genes showed a high level of concordance between mRNA and protein, as immediately obvious from the eigengene classification. Correspondingly, both gene sets were enriched for high correlation between mRNA and protein. Additionally, XEN markers were enriched for the “synthesis only” model (Supplementary Fig. 4b). This might be related to the fact that XEN cells have to produce high levels of extracellular matrix proteins(Mulvey et al., 2015), like laminin (*Lamb1*) or collagen (*Col4a2*) . Consequently, these proteins must be synthesized in a timely manner to ensure the proper function of the XEN cells. All in all, it seems that cell type specific markers defined at the mRNA level could be confirmed at the level of protein and that for these genes protein expression closely follows mRNA expression. Compared to the gene sets discussed so far, ribosomal protein (RP) genes showed a remarkable extent of discordance between mRNA and protein expression. Eigengene classification revealed that many RP genes had protein profiles that were more dynamic than their mRNA counterparts. Correspondingly, RP genes were enriched for low correlation between mRNA and protein (p-value = 3.3E-2). As cells differentiated, the protein levels of RP genes decreased, consistent with reduced cell division rates. The rate of decrease in abundance, however, was RP specific. Thus, it will be interesting to isolate ribosomes and analyze the extent to which these RP dynamics reflect ribosome remodeling and specialization (Slavov et al., 2015). In summary, we have shown that the 4 analyzed gene sets follow distinct regulatory modes that can be related to biological functions.

The kinetic model can be applied to single-cell transcriptomics data to predict protein levels in differentiated cell types

In the experiment presented here, the existence of good antibodies for highly expressed surface markers allowed us to purify differentiated cells at 96 h and profile their proteome. For earlier time points or many other differentiation assays such an approach is difficult or even impossible. By contrast, single-cell transcriptomics methods can be applied to any differentiation system. Hence, we would like to use such data sets to predict protein levels in subpopulations. To that end, we extracted cell type specific mRNA dynamics during differentiation from our earlier single-cell RNA-seq measurement of the system (Semrau et al., 2016). We then applied our kinetic model to this data set to predict protein levels in the differentiated cell types at 96 h (Fig. 4b, Methods). Our prediction was clearly superior to a prediction that only used bulk RNA-seq measurements and protein-to-mRNA ratios (Edfors et al., 2016) (Fig. 4c). We have thus demonstrated that our kinetic model with parameters learned from bulk measurements can be applied to single-cell transcriptomics data to predict cell type specific protein levels.

We finally compared the differentiated cell types directly with each other. Overall, the correlation between mRNA and protein changes was poor and we identified a few outlier genes in particular that showed extreme behavior (Fig. 4d). These outliers had comparable protein expression in XEN and ECT cells (at most 2-fold difference) but mRNA expression was much lower in XEN cells (up to 19-fold). Notably, these outliers are strongly enriched for imprinted genes (hypergeometric test, p-value = 2.3E-10). It is a well-known fact that some imprinted genes are mono-allelically expressed in extra-embryonic tissues (Miri and Varmuza, 2009). Yet, the observed down-regulation goes well beyond a two-fold change expected for mono-allelic expression. This observation demonstrates that our data set can be used to discover significant differences in gene regulation between differentiated cell types.

Discussion

Here we systematically analyzed the dynamics of mRNA and protein levels during mESC differentiation and found widespread discordance. Such discordance has been observed recently in several systems, in particular: *Xenopus* development (Peshkin et al., 2015), *C. elegans* development (Grün et al., 2014), macrophage differentiation (Kristensen et al., 2013) and mESC differentiation (Lu et al., 2009). While this discordance is typically interpreted as a sign of (post) translational regulation (Grün et al., 2014) (Lu et al., 2009), we showed here that a simple model with constant kinetic rates, substantially reduces the discordance for 63% of discordant genes (Supplementary Fig. 3a). The same kinetic model explained protein dynamics of a third of all genes during stress response in yeast (Tchourine et al., 2014) and of 75% of all genes in *Xenopus* development (Peshkin et al., 2015). Consistently, this simple model thus explains discordance for significant proportions of the genome. Genes that were not fit well by the kinetic model, are by our definition dynamically regulated, as constant synthesis and degradation rates are insufficient to describe the observed kinetics. This approach is complementary to the recently developed PECA method that can be used to reveal regulatory events at the mRNA and protein level (Cheng et al., 2016).

Our in-depth analysis of several gene sets revealed that cell type specific genes show a high concordance between mRNA and protein dynamics, while for RP genes the correlation is much lower. Together with previous reports (Kristensen et al., 2013) (Jovanovic et al., 2015) our study supports a model in which mRNA fold changes set the level of newly produced proteins that have crucial, specific function in the new cell state or cell type. Regulation on the level of protein turnover, on the other hand, is used to adapt the existing proteome. Importantly, we also showed that some pluripotency genes, defined as such by being down-regulated at the mRNA level, showed increasing protein expression. This result cautions against defining markers for cell states or cell types solely based on mRNA expression.

In summary, this study provided the first in-depth, integrated analysis of mRNA and protein dynamics during mESC differentiation. All measured data are provided in a convenient web application. We hope that this application will facilitate future studies of specific gene sets or global relationships, for example between sequence features and protein regulation (Vogel et al., 2010).

Author contributions

Conceptualization, S.S. and N.S.; Investigation, P. vd B., S.S., B.B. and N.S.; Resources, B.B.; Formal analysis, P. vd B. and N.S.; Software, P. vd B.; Data curation, P. vd B. and N.S.; Writing – original draft, S.S. and P. vd B.; Writing – review and editing, P. vd B., N.S. and S.S.; Supervision, S.S. and N.S.

Acknowledgements

P. vd B. and S.S. were supported by the Netherlands Organisation for Scientific Research (NWO/OCW), as part of the Frontiers of Nanoscience (NanoFront) program. Data analysis was carried out on the Dutch national e-infrastructure with the support of SURF Foundation. N.S. was supported by a New Innovator Award from the NIGMS of the NIH under Award number DP2GM123497. We would like to thank Rudolf Jaenisch for reagents and valuable advice.

The authors declare no competing financial interests.

References

Alexa, A., Rahnenführer, J., Lengauer, T., 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22, 1600–1607. doi:10.1093/bioinformatics/btl140

Cheng, Z., Teo, G., Krueger, S., Rock, T.M., Koh, H.W., Choi, H., Vogel, C., 2016. Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. Mol Syst Biol 12, 855–855. doi:10.15252/msb.20156423

de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M., Vogel, C., 2009. Global signatures of protein and mRNA expression levels. Mol Biosyst 5, 1512–1526. doi:10.1039/B908315D

Edfors, F., Danielsson, F., Hallström, B.M., 2016. Gene-specific correlation of RNA and protein levels in human cells and tissues. Molecular Systems …. doi:10.15252/msb.20167325

Grün, D., Kirchner, M., Thierfelder, N., Stoeckius, M., Selbach, M., Rajewsky, N., 2014. Conservation of mRNA and Protein Expression during Development of C. elegans. Cell Reports 6, 565–577. doi:10.1016/j.celrep.2014.01.001

Ingolia, N.T., Lareau, L.F., Weissman, J.S., 2011. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. Cell 147, 789–802. doi:10.1016/j.cell.2011.10.002

Jovanovic, M., Rooney, M.S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E.H., Fields, A.P., Schwartz, S., Raychowdhury, R., Mumbach, M.R., Eisenhaure, T., Rabani, M., Gennert, D., Lu, D., Delorey, T., Weissman, J.S., Carr, S.A., Hacohen, N., Regev, A., 2015. Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. - PubMed - NCBI. Science 347, 1259038–1259038. doi:10.1126/science.1259038

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., Kirschner, M.W., 2015. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell 161, 1187–1201. doi:10.1016/j.cell.2015.04.044

Kocabas, A., Duarte, T., Kumar, S., Hynes, M.A., 2015. Widespread Differential Expression of Coding Region and 3' UTR Sequences in Neurons and Other Tissues. Neuron 88, 1149–1156. doi:10.1016/j.neuron.2015.10.048

Kristensen, A.R., Gsponer, J., Foster, L.J., 2013. Protein synthesis rate is the predominant regulator of protein expression during differentiation. Mol Syst Biol 9, 689–689. doi:10.1038/msb.2013.47

Kunath, T., Saba-El-Leil, M.K., Almousailleakh, M., Wray, J., Meloche, S., Smith, A., 2007. FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. Development 134, 2895–2902. doi:10.1242/dev.02880

Liu, Y., Beyer, A., Aebersold, R., 2016. On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell 165, 535–550. doi:10.1016/j.cell.2016.03.014

Loh, K.M., Chen, A., Koh, P.W., Deng, T.Z., Sinha, R., Tsai, J.M., Barkal, A.A., Shen, K.Y., Jain, R., Morganti, R.M., Shyh-Chang, N., Fernhoff, N.B., George, B.M., Wernig, G., Salomon, R.E.A., Chen, Z., Vogel, H., Epstein, J.A., Kundaje, A., Talbot, W.S., Beachy, P.A., Ang, L.T., Weissman, I.L., 2016. Mapping the Pairwise Choices Leading from Pluripotency to Human Bone, Heart, and Other Mesoderm Cell Types. Cell 166, 451–467. doi:10.1016/j.cell.2016.06.011

Lu, R., Markowetz, F., Unwin, R.D., Leek, J.T., Airoldi, E.M., MacArthur, B.D., Lachmann, A., Rozov, R., Ma'ayan, A., Boyer, L.A., Troyanskaya, O.G., Whetton, A.D., Lemischka, I.R., 2009. Systems-level dynamic analyses of fate change in murine embryonic stem cells. Nature 462, 358–362. doi:10.1038/nature08575

Miri, K., Varmuza, S., 2009. Chapter 5 Imprinting and Extraembryonic Tissues—Mom Takes Control, in: International Review of Cell and Molecular Biology. Elsevier, pp. 215–262. doi:10.1016/S1937-6448(09)76005-8

Mulvey, C.M., Schröter, C., Gatto, L., Dikicioglu, D., Fidaner, I.B., Christoforou, A., Deery, M.J., Cho, L.T.Y., Niakan, K.K., Martinez Arias, A., Lilley, K.S., 2015. Dynamic Proteomic Profiling of Extra-Embryonic Endoderm Differentiation in Mouse Embryonic Stem Cells. STEM CELLS 33, 2712–2725. doi:10.1002/stem.2067

Peshkin, L., Wühr, M., Pearl, E., Haas, W., Freeman, R.M., Gerhart, J.C., Klein, A.M., Horb, M., Gygi, S.P., Kirschner, M.W., 2015. On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development. Developmental cell 35, 383–394. doi:10.1016/j.devcel.2015.10.010

Sampath, P., Pritchard, D., Pabon, L., Reinecke, H., Schwartz, S., Morris, D., Murry, C., 2008. A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation. Cell Stem Cell 2, 448–460.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., Selbach, M., 2011. Global quantification of mammalian gene expression control. Nature 473, 337–342. doi:10.1038/nature10098

Semrau, S., Goldmann, J., Soumillon, M., Mikkelsen, T.S., Jaenisch, R., van Oudenaarden, A., 2016. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. bioRxiv 068288. doi:10.1101/068288

Slavov, N., Semrau, S., Airoldi, E., Budnik, B., van Oudenaarden, A., 2015. Differential Stoichiometry among Core Ribosomal Proteins. Cell Reports 13, 865–873. doi:10.1016/j.celrep.2015.09.056

Soldner, F., Jaenisch, R., 2012. iPSC Disease Modeling. Science 338, 1155–1156. doi:10.1126/science.1227682

Tchourine, K., Poultney, C.S., Wang, L., Silva, G.M., Manohar, S., Mueller, C.L., Bonneau, R., Vogel, C., 2014. One third of dynamic protein expression profiles can be predicted by a simple rate equation. Mol Biosyst 10, 2850–2862. doi:10.1039/C4MB00358F

Vogel, C., de Sousa Abreu, R., Ko, D., Le, S.Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R., Marcotte, E.M., Penalva, L.O., 2010. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol Syst Biol 6, 400. doi:10.1038/msb.2010.59

Vogel, C., Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. doi:10.1038/nrg3185

Wall, M., Rechtsteiner, A., Rocha, L., 2003. Singular value decomposition and principal component analysis. A practical approach to microarray data analysis 91–109.

Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., Smith, A., 2008. The ground state of embryonic stem cell self-renewal. Nature 453, 519–523. doi:10.1038/nature06968

Figure captions

Figure 1 mRNA and protein expression correlate poorly during mESC differentiation

(A) Experimental setup. (B) mRNA versus protein expression of 7459 genes in mESCs. Each data point is an individual gene. Red lines indicate contour lines of equal density. (C) Sample-wise Pearson correlation between mRNA and protein for all samples. The solid line indicates the average of all time course samples. The grey area indicates the 5% rejection region for all samples being identical (see Methods). Error bars: SEM. (D) mRNA versus protein expression at all time points for nine example genes. Pearson's correlation r is indicated for each gene. The line and grey area indicate the linear regression fit and 95% CI, respectively. Error bars: SEM. (E) Distribution of the gene-wise Pearson correlation between mRNA and protein. Numbered arrows indicate the position of the examples shown in D. See also Supplementary Figure 1.

Figure 2 Classification of temporal mRNA and protein expression profiles by dominant trends reveals widespread discordance

(A) First six eigengenes of mRNA and protein expression profiles. (B) Reconstruction of mRNA and protein expression profiles from the top three eigengenes of an example gene. (C) Cumulative variance explained by the eigengenes for mRNA and protein profiles. (D) Classification of all genes by their dominant mRNA eigengene (columns) and protein eigengene (rows).
See also Supplementary Figure 1.

Figure 3 Simple kinetic models of protein synthesis and degradation explain mRNA-protein discordance.

(A) Kinetic model. k_s = synthesis rate constant; k_d = degradation rate constant. (B) Example fits of the *full model* (k_s >0, k_d >0) and the three reduced models: *synthesis only* ($k_s > 0, k_d = 0$), *degradation only* ($k_s = 0, k_d > 0$) and *degenerate* ($k_s = k_d = 0$). Percentages indicate the fraction of genes fit best by the respective model. (C) Distribution of Pearson correlation between measured protein expression and mRNA expression or predicted protein expression. (D) Extended kinetic model. $k_s(t)$ = time-dependent synthesis rate. (E) mRNA expression, log ratio of expression from CDS and 3'UTR and protein expression profiles of two example genes with fits of the extended model (solid line) or the basic model (dashed line). (F) Distribution of Pearson correlation between measured protein expression and: mRNA expression, protein expression predicted by the basic model or the extended model. Error bars in (B) and (E): SEM.

See also Supplementary Figures 2 and 3.

Figure 4. Classification and kinetic modelling reveal differences between gene sets involved in differentiation and between differentiated cell types.

(A) Comparison of four gene sets that are relevant for differentiation. Log_2 fold change (L2FC) of mRNA and protein expression are shown for individual genes (colored) and the set average (black). The p-value in the classification matrix is based on picking genes at random from all genes (chi-squared test). (B) mRNA expression of XEN and ECT subpopulations (from single cell data) and the mixed populations (bulk sample). Protein expression in XEN and ECT is predicted by applying the kinetic model to the single cell data. Alternatively, at 96 h we also predicted protein based on the protein-to-mRNA (PTR) ratio. (C) Sum of squared residuals (SSR) of the kinetic model-based prediction compared to the PTR-based prediction for the XEN and ECT marker genes. (D) mRNA and protein expression in XEN cells relative to ECT cells. Outlier genes are highlighted with a dark background and imprinted genes are shown in red (obtained from www.geneimprint.com, Oct-11-2016). Imprinted genes are significantly enriched in the outlier gene set (hypergeometric test: p-value = 2.72e-10).

See also Supplementary Figure 4.

Methods

Cell culture

E14 mouse embryonic stem cells were cultured as previously described (Semrau et al., 2016). Briefly, cells were grown in modified 2i medium (Ying et al., 2008): DMEM/F12 (Life technologies) supplemented with 0.5x N2 supplement, 0.5x B27 supplement, 4mM L- glutamine (Gibco), 20 µg/ml human insulin (Sigma-Aldrich), 1x 100U/ml penicillin/streptomycin (Gibco), 1x MEM Non-Essential Amino Acids (Gibco), 7 µl 2-Mercaptoethanol (Sigma-Aldrich), 1 µM MEK inhibitor (PD0325901,Stemgent), 3 µM GSK3 inhibitor (CHIR99021, Stemgent), 1000 U/ml mouse LIF (ESGRO). Cells were passaged every other day with Accutase (Life technologies) and replated on gelatin coated tissue culture plates (Cellstar, Greiner bio-one).

Differentiation and sample collection

Retinoic acid induced differentiation was carried out exactly as describe before (Semrau et al., 2016). Prior to differentiation cells were grown in 2i medium for at least 2 passages. Cells were seeded at 2.5 × 10^5 per 10 cm dish and grown over night (12 h). Cells were then washed twice with PBS and differentiated in basal N2B27 medium (2i medium without the inhibitors, LIF and the additional insulin) supplemented with 0.25 µM all-trans retinoic acid (RA, Sigma-Aldrich). Spent medium was exchanged with fresh medium after 48 h.

To collect samples, cells were dissociated with Accutase. RNA was extracted from half of the sample (RNeasy, Qiagen) and the purified RNA was stored at -80C until RNA-sequencing was performed. The other half of the sample was flash frozen in liquid nitrogen and stored at -80C until mass spectrometry was performed.

Fluorescence-activated cell sorting

FACS sorting of the differentiated cell types and quantification of the cell type frequencies was carried out exactly as described previously *(Semrau et al., 2016)*.

RNA sequencing and mRNA quantification

Library preparation and RNA sequencing

The libraries for RNA sequencing were prepared under standard conditions using Illumina's TruSeq stranded mRNA sample preparation kit. The libraries were sequenced using Illumina HiSeq 3000 ; 40

basepair long, stranded single-end reads were sequenced at an average read depth of 40 million reads per sample. The data is available through GEO.

Read alignment

An RSEM-reference was created using RSEM v1.2.28 (Li and Dewey, 2011) with the Illumina iGenome GRCm38 reference using the standard settings. Next, the Illumina adapter was trimmed from the reads with *cutadapt* v1.8.3 (Martin, 2011) and low quality bases with sickle v1.33 (Joshi et al., 2011). Finally the reads were aligned with RSEM v1.2.28 (Li and Dewey, 2011) and Bowtie 2 v2.2.6 (Langmead and Salzberg, 2012) using standard settings accept for "*--sampling-for-bam --fragment-length-mean 40"*. The option "*--sampling-for-bam"* was applied so each read appears in the BAM file once. This enabled the estimation of the CDS and 3'UTR counts by *summarizeOverlaps* from the package *GenomicAlignment* v1.8.4 (Lawrence et al., 2013).

Gene quantification

mRNA expression was quantified by several different methods depending on the application. Transcripts per million (TPM) was calculated by *RSEM* and was used when comparing between genes since it is corrected for gene length. The more variance stabilized regularized log counts (rLC) were determined by applying the *rlog* function from *DESeq2* v1.12.3 (Love et al., 2014) on rounded expected counts obtained from *RSEM*. From this regularized counts (rC) were obtained by: rC = 2^{rLC}. rLC and rC are corrected for overdispersion in low-read genes and are therefore used when comparing one gene across multiple samples. CDS and 3'UTR counts were determined by splitting the gene annotation file (GTF) with the *GenomicFeatures* package v1.26.0 (Lawrence et al., 2013) into CDS and 3'UTR for every Ensembl gene ID. Next, the number of reads on the CDS and 3'UTR features from the aligned BAM files were counted with *summarizeOverlaps* with default options. "Union", the default option for *mode*, discards reads, if they overlap with both CDS and 3'UTR. The ratio *w* (CDS / 3'UTR) was only calculated for genes with at least 10 reads for CDS and 3'UTR in every sample.

Differentially expressed genes

Differentially expressed genes (DEGs) were determined by *DESeq2* v1.12.3 (Love et al., 2014) on the rounded expected counts obtained from RSEM at a false discovery rate (FDR) of 10%. The gene set 'pluripotency genes' were DEGs that were down-regulated when comparing the samples 0h (n=2) and 96h (n=2). XEN- and ECT-marker gene sets were DEGs that were up-regulated when comparing the samples 0h (n=2) with XEN (n=1) or ECT (n=1) respectively. Additionally, XEN- and ECT-markers have at least a 2-fold difference in expression between the two cell types.

Mass spectrometry and protein quantification

Sample preparation

Pelleted cells were lysed in 400 μl RIPA buffer, except for the sorted cells, which were lysed in 200 μl RIPA buffer. Volumes of cell lysate corresponding to 100 μg protein per sample were digested with trypsin using a modified FASP protocol (Wiśniewski et al., 2009). Subsequently each sample was labeled with TMT 10-plex reagent (Prod# 90061, Thermo Fisher, San Jose, CA) according to the manufacturer's protocol. All labeled samples were combined into a set-sample.

Mass spectrometry

The labeled set–sample was fractionated by electrostatic repulsion-hydrophilic interaction chromatography chromatography (ERLIC) run on an HPLC 1200 Agilent system using PolyWAX LP column (200x2.1 mm, 5 μm, 30nm, PolyLC Inc, Columbia, MD) and a fraction collector (Agilent Technologies, Santa Clara, CA). Set-samples were fractionated into a total of 40 ERLIC fractions. Each ERLIC fraction was subsequently further separated by online nano-LC and submitted for tandem mass spectrometry analysis to both LTQ OrbitrapElite or Q exactive high field (HF). One third of each fraction was injected from an auto–sampler into the trapping column (75 um column ID, 5 cm length packed with 5 um beads with 20 nm pores, from Michrom Bioresources, Inc.) and washed for 15 min; the sample was eluted to analytic column with a gradient from 2 to 32 % of buffer B (0.1 % formic acid in ACN) over 180 min gradient and fed into LTQ OrbitrapElite or Q exactive HF. The instruments were set to run in TOP 20 MS/MS mode method with dynamic exclusion. After MS1 scan in Orbitrap with 60K resolving power, each ion was submitted to an HCD MS/MS with 60K resolving power and to CID MS/MS scan subsequently. All quantification data were derived from HCD spectra.

Protein quantification

Relative peptide levels were estimated from reporter ion intensities measured at MS2 level. Only peptides with co-isolation below 40 % were used for quantification. The intensities of all peptides belonging to a Uniprot ID were averaged to form mean peptide intensity (MPI) for every protein. When comparing different protein samples mean peptide intensities were normalized to the sample-mean to form protein expression. Standard error of the mean (SEM) was calculated for every protein as follows: 1) for every peptide the intensities were averaged across the samples, 2) the SEM was calculated from these mean-centered peptide intensities for every protein and sample.

Protein reliability

The protein reliability was calculated for genes with at least two peptides quantified. For each gene, the peptides were randomly split into two groups and the MPI was calculated for each group as described above. The correlation between the MPIs of the two peptide groups across the different samples is defined as the reliability of the measurement of that protein.

Transcriptomics and proteomics integration

While transcripts were identified by Ensembl gene IDs, Uniprot IDs were used for proteins. To integrate the two, we mapped 7681 out of 8515 Uniprot IDs to Ensembl gene IDs present in the RNA-seq data using the *idmapping* file from the Uniprot website (15-Sept-2016). An additional set of Uniprot IDs were mapped to Ensembl IDs using *biomaRt* v2.28.0 (Durinck et al., 2009). Some proteins have more than one Ensembl ID mapping to it, therefore 33 Uniprot IDs were removed, Moreover, 92 Uniprot IDs mapped non-uniquely to Ensembl IDs and for these the protein intensities were reevaluated based on Ensembl IDs. Finally, some genes were not considered because they were not detected in all samples. This resulted in a total of 7489 genes based on Ensembl gene IDs, for which we have matched mRNA and protein expression data in all samples. Additionally, we observed 3770 genes with at least 10 mRNA reads in every sample but no detected protein.

Sample-wise correlation

We tested if the sample-wise correlation is constant during the differentiation time course using a resampling approach. For each bootstrap a *pseudo-sample* was constructed consisting of every gene, but with mRNA and protein expression randomly sampled from the different time points. The correlations of 10,000 *pseudo-samples* were calculated to obtain a null distribution. Samples have significantly different correlation if it falls below or above the 0.36 and 99.64 percentiles of the null distribution respectively ($\alpha = 0.05$, Bonferroni correction, grey area in Figure 1c).

Gene-wise correlation

To define a threshold for low gene-wise correlation we applied a shuffling approach (Tchourine et al., 2014). We determined the Pearson correlation for all possible permutations of the mRNA and protein expression for every gene. More than 95% of all Pearson correlation values obtained in this way were lower than 0.7, which we therefore set as the threshold between low and high correlation.

Expression profile classification

mRNA and protein expression were arranged in matrix form rows corresponding to genes and the columns corresponding to time course samples. These matrices were standardized by rows. Next, standard singular value decomposition (SVD) was performed separately for mRNA and protein (Wall et al., 2003). From this analysis, we obtain n eigengenes $\vec{V}_k$ where $k \in 1, \dots, n$ and *n* is the number of time points. Using these eigengenes we can reconstruct the standardized expression of gene i, as follows: $\vec{X}_i = \sum_k M_{ik} \vec{V}_k$, where M_{ik} is the contribution of eigengene k to the standardized expression of gene i. We defined the eigengene with the biggest contribution to $\vec{X}_i$ as the dominant eigengene. To determine if there is an enrichment of genes with concordant mRNA and protein eigengenes, we calculated an empirical p-value based on a null distribution generated by bootstrapped (number of bootstraps = 100,000). This null distribution was constructed under the assumption that the marginal eigengene distributions of mRNA and protein are independent. Moreover, we defined a confident set of genes with a bigger than median fold-change between the contribution of the dominant eigengene and the second most contributing eigengene for both mRNA and protein.

Kinetic models of protein synthesis and degradation

Approximation of mRNA and CDS/3'UTR expression by natural cubic splines

To describe the mRNA, CDS and 3'UTR behavior in the kinetic model of protein synthesis and degradation we approximated the expression with natural cubic splines. These splines were fit on the mRNA expression and on the $\log_2$ fold change (L2FC) of *w*, which we call *ω*. The number of degrees of freedom *p* used for the fits of every gene was 4 for mRNA expression and 3 for *ω* expression. These values were automatically determined as described by Storey *et al.* (Storey, 2005). Briefly, an SVD was performed on the expression matrices of mRNA and *ω* and the first *n* eigengenes that explain at least 60% of the variance were selected. For each of these eigengenes the optimal number of degrees of freedom p_i was selected by leave one out cross validation (LOOCV) and the largest p_i was used as the number of degrees of freedom *p* to fit the natural cubic splines for all the genes of the expression matrix. The nodes of the cubic splines were equally spaced across the time course.

Kinetic rate parameters estimation

We model protein turnover as a birth-death process

$$\frac{dP(t)}{dt} = k_s \cdot R(t) - k_d \cdot P(t)$$

where $P(t)$and $R(t)$ are protein and mRNA expression respectively. The solution of this ordinary differential equation (ODE) is given by:

$$P(t) = P_0 e^{-k_d\,t} + k_s \int_0^{\tau} R(\tau)\, e^{-k_d \cdot (t-\tau)}$$

where P_0 is the protein expression at t = 0 hours. The integral of this equation was estimated numerically in R using the spline fits described above. We fit the model using gene specific parameters P_0, k_s and k_d with the Levenberg – Marquardt non-linear least squares algorithm, which is implemented in the R package *minpack.lm* v1.2-0. Additionally, we fit models where we set $k_d = 0$, $k_s = 0$ or $k_d = k_s = 0$. For each successful fit we determined the Bayesian Information Criterion:

$$BIC = -2\ln(\hat{L}) + k \cdot \ln(n)$$

where $\hat{L}$ is the posterior likelihood of the fit, k is number of parameters in the model and n is the number of time points. $\hat{L}$ is determined by:

$$\hat{L} = \prod_{j=1}^{n} p\big(P(t_j)\big|\hat{\theta}\big)$$

where $\hat{\theta}$ is the vector of inferred model parameters. The probabilities are estimated by assuming a normal distribution around the observed protein expression with a standard deviation equal to the SEM of the peptide intensities. The kinetic model with the lowest BIC was selected as the optimal model.

Additionally, for the subset of genes for which we could determine ω we constructed a model with a time-dependent synthesis rate:

$$\frac{dP(t)}{dt} = k_s(t) \cdot R(t) - k_d \cdot P(t) = \kappa_s \big(1 + \beta\, \omega(t)\big) \cdot R(t) - k_d \cdot P(t)$$

where κ_s describes the constant synthesis rate and β parameterizes the time-dependent modulation of the synthesis rate by ω. The solution of this ODE:

$$P(t) = P_0 e^{-k_d\,t} + \int_0^{\tau} \Big((\kappa_s + \beta\, \omega(t)\,) \cdot R(\tau) \cdot e^{-k_d \cdot (t-\tau)} \Big)$$

was fit to the data in the same manner as above.

95% confidence region estimation

To estimate the 95% confidence intervals (CIs) for $\mathrm{k_s}$ and $\mathrm{k_d}$ we applied Wilk's theorem:

$$\ln\left(L(\theta)\right) \geq \ln\left(L(\hat{\theta})\right) - \frac{1}{2}\chi^2_{1,1-\alpha}$$

where α is 0.05 and $\chi^2_{1,1-\alpha}$is the value at which the cumulative chi-squared distribution with 1 degree of freedom reaches 0.95. We varied k_s and k_d around the obtained fit $\hat{\theta}$ to find the edges where Wilk's theorem holds. These edges where determined at 24 directions in the k_s - k_d solution plane to obtain a crude 95% confidence region. The projection of this region on k_s and k_d defined $CI^{95\%}_{k_s}$ and $CI^{95\%}_{k_d}$, their respective 95% CIs. Note that these intervals are typically much larger than the intervals obtained when searching one parameter at a time. Genes with the full model (as determined by BIC), and with a small $CI^{95\%}_{k_s}$ and $CI^{95\%}_{k_d}$ (each spanning less than a 10-fold range) were defined as the high-confidence gene set. Additionally, for genes in this set we determined the protein half-life τ_p as

$$\tau_p = \frac{\ln 2}{k_d}$$

Protein prediction of sorted populations

We applied our kinetic model to single-cell transcriptomics data of RA driven differentiation, which we obtained previously (Semrau et al., 2016). We determined the mean expression of all cells, as well as XEN and ECT subpopulations starting from the lineage bifurcation at 36 h. All three datasets thus have identical expression up to 36 h. We then scaled the subpopulation data to the bulk data measured here for every gene in the following way: 1) We standardized the single cell time course data using the mean and standard deviation of the pooled single cell data, and 2) we scaled the standardized single cell data to the bulk data using the mean and standard deviations of the bulk time course. Next, we fit a natural cubic spline to the single cell data as before and applied the kinetic model using P_0, k_s and k_d learned from the bulk mRNA and protein measurements. We evaluated the model performance by calculating the residuals between the predicted XEN and ECT protein expression at 96 h and the bulk measurements of protein in the purified cell types.

An alternative way of predicting protein expression is by simply multiplying a gene's protein-to-mRNA ratio (PTR) with the gene's mRNA expression. We defined the PTR as the mean protein expression divided by the mean mRNA expression during the time course. We predicted the protein expression of the XEN and ECT populations at 96 h using the bulk mRNA of the respective sorted populations. We used the sorted bulk data rather than the single cell data, because it is more accurate and we therefore expect this to perform better. Like with the single cell predictions, we evaluated model performance using the residuals of the PTR-predictions relative to the measured protein expression of the sorted bulk data.

Ribosomal protein gene list

The list of RPs was compiled as all Swiss-Prot proteins curated as ribosomal proteins in their descriptions.

Eigengene dynamics

We quantified the dynamics of the eigengene profiles as the mean of the squared second derivatives (roughness). The second derivatives were estimated numerically from three unequally spaced points by this formula:

$$\frac{d^2y}{dx^2} = \frac{2y_1}{(x_2 - x_1)(x_3 - x_1)} - \frac{2y_2}{(x_3 - x_2)(x_2 - x_1)} + \frac{2y_3}{(x_3 - x_2)(x_3 - x_1)}$$

where x_1, x_2 and x_3 are adjacent time points and y_1, y_2 and $y_3$are the respective eigengene intensities.

GO term enrichment

GO term enrichment was performed with the R package *topGO* v2.24.0 (Alexa et al., 2006) with the *classic* algorithm. The genes were ranked using Fisher's exact test and deemed significant with an FDR of 10%.

Accession numbers

The RNA-seq data has been deposited in GEO (ID: GSE9563). The raw MS data has been deposited in MassIVE (ID: MSV000080461). A web application complementing this publication, which allows convenient access to all data can be found here:

https://home.physics.leidenuniv.nl/~semrau/proteomics/

user name:

password:

Additional references

Alexa, A., Rahnenführer, J., Lengauer, T., 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22, 1600–1607. doi:10.1093/bioinformatics/btl140

Durinck, S., Spellman, P.T., Birney, E., Huber, W., 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols 4, 1184–1191. doi:10.1038/nprot.2009.97

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359. doi:10.1038/nmeth.1923

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., Carey, V.J., 2013. Software for Computing and Annotating Genomic Ranges. PLoS Comp Biol 9, e1003118–10. doi:10.1371/journal.pcbi.1003118

Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 1. doi:10.1186/1471-2105-12-323

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 31. doi:10.1186/s13059-014-0550-8

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, pp. 10–12. doi:10.14806/ej.17.1.200

Storey, J.D., 2005. Significance analysis of time course microarray experiments. Proceedings of the National Academy of Sciences 102, 12837–12842. doi:10.1073/pnas.0504609102

Wiśniewski, J.R., Zougman, A., Nagaraj, N., Mann, M., 2009. Universal sample preparation method for proteome analysis. Nat Methods 6, 359–362. doi:10.1038/nmeth.1322

Figure 1

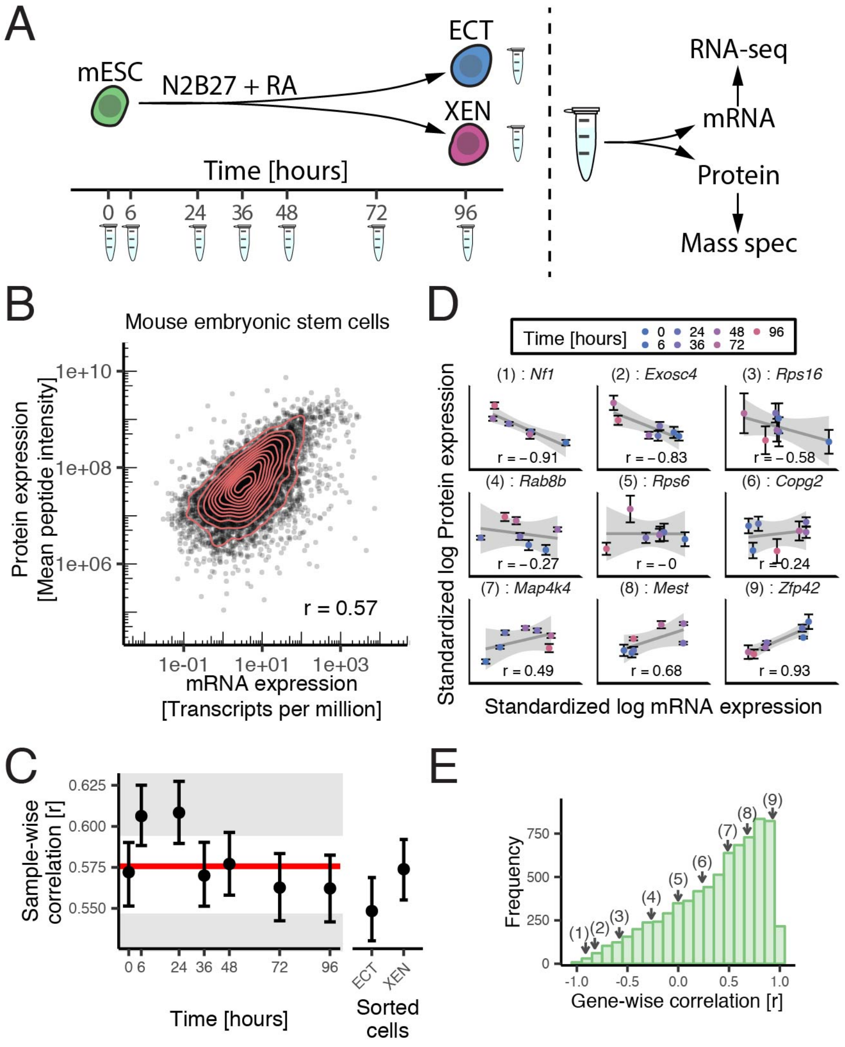

A
mESC
N2B27 + RA
ECT
XEN
Time [hours]
0 6 24 36 48 72 96
RNA-seq
mRNA
Protein
Mass spec
B
Mouse embryonic stem cells
Protein expression [Mean peptide intensity]
1e+10
1e+08
1e+06
r = 0.57
1e-01 1e+01 1e+03
mRNA expression [Transcripts per million]
C
Sample-wise correlation [r]
0.625
0.600
0.575
0.550
0 6 24 36 48 72 96
Time [hours]
ECT XEN
Sorted cells
D
Time [hours] 0 6 24 36 48 72 96
Standardized log Protein expression
(1) : Nf1
r = −0.91
(2) : Exosc4
r = −0.83
(3) : Rps16
r = −0.58
(4) : Rab8b
r = −0.27
(5) : Rps6
r = −0
(6) : Copg2
r = 0.24
(7) : Map4k4
r = 0.49
(8) : Mest
r = 0.68
(9) : Zfp42
r = 0.93
Standardized log mRNA expression
E
Frequency
750
500
250
0
(1) (2) (3) (4) (5) (6) (7) (8) (9)
-1.0 -0.5 0.0 0.5 1.0
Gene-wise correlation [r]

Figure 2

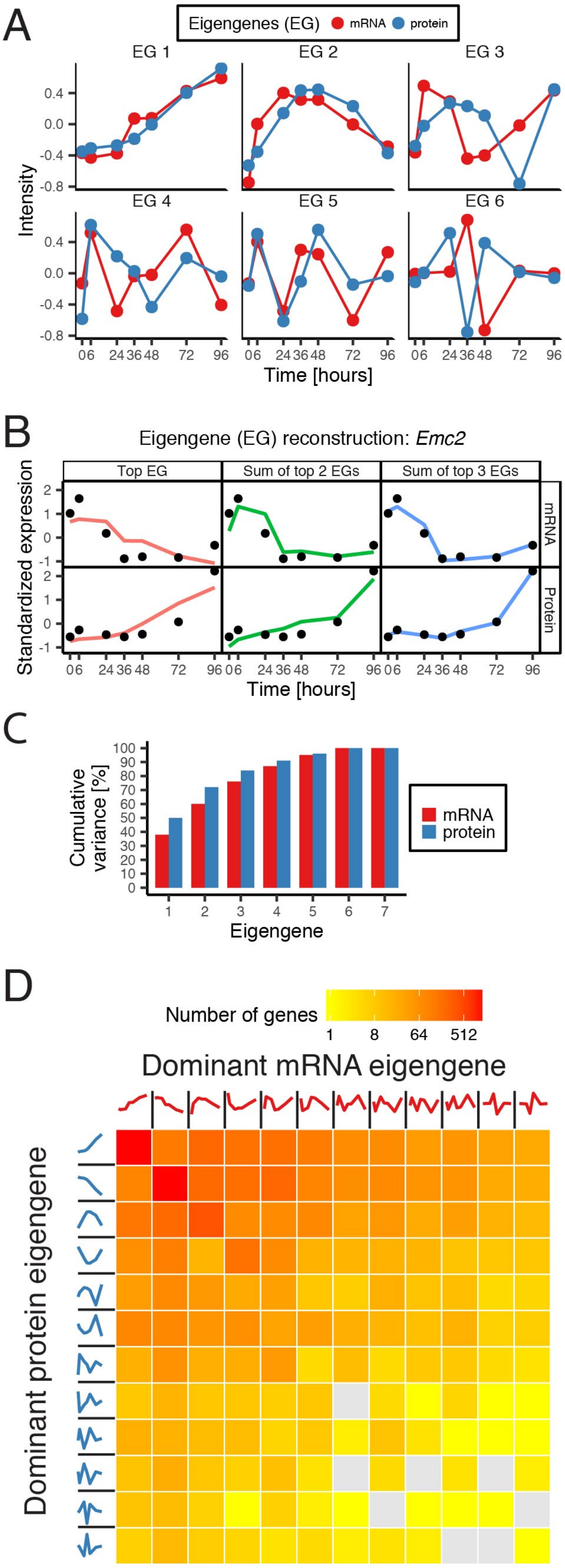

A
Eigengenes (EG) ● mRNA ● protein
EG 1
EG 2
EG 3
EG 4
EG 5
EG 6
Intensity
0.4
0.0
-0.4
-0.8
06 24 36 48 72 96
Time [hours]
B
Eigengene (EG) reconstruction: *Emc2*
Top EG
Sum of top 2 EGs
Sum of top 3 EGs
mRNA
Protein
Standardized expression
2
1
0
-1
Time [hours]
C
Cumulative variance [%]
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7
Eigengene
mRNA
protein
D
Number of genes
1
8
64
512
Dominant mRNA eigengene
Dominant protein eigengene

Figure 3

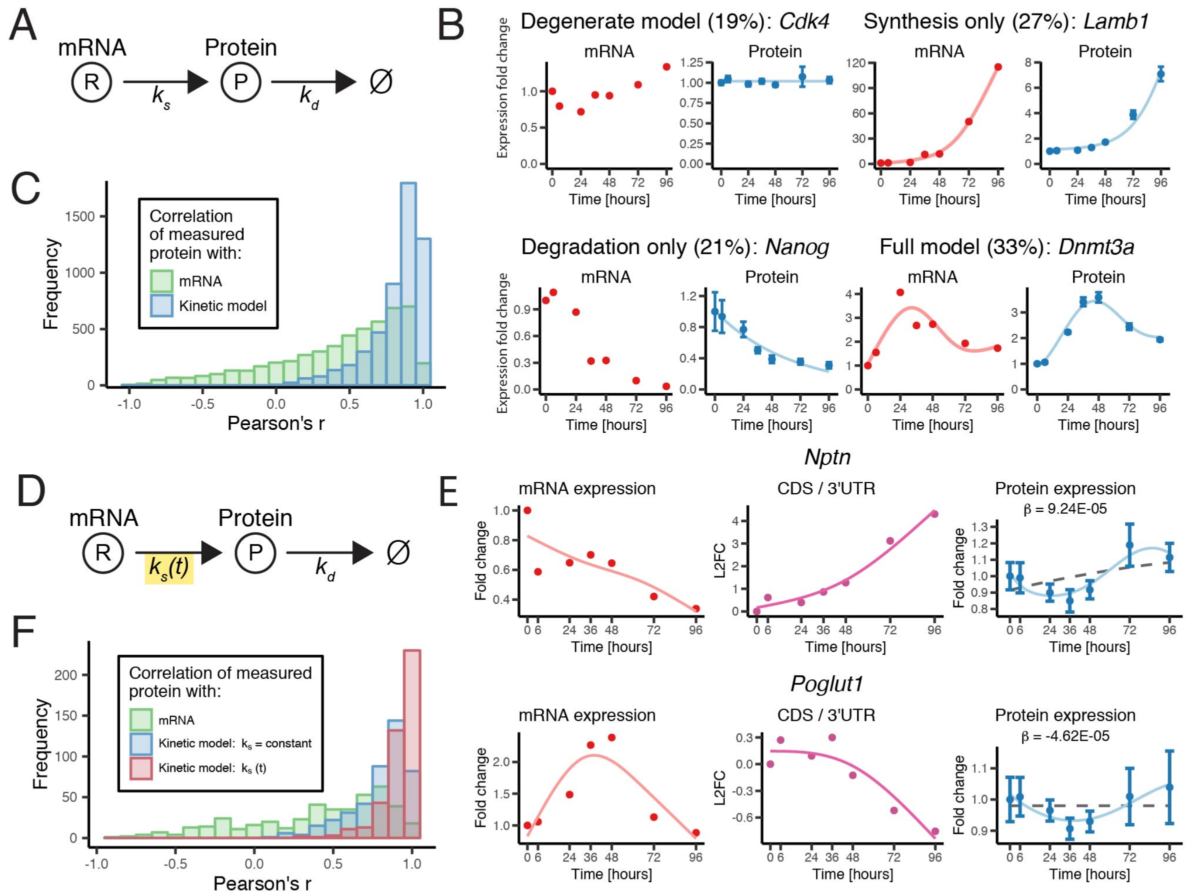
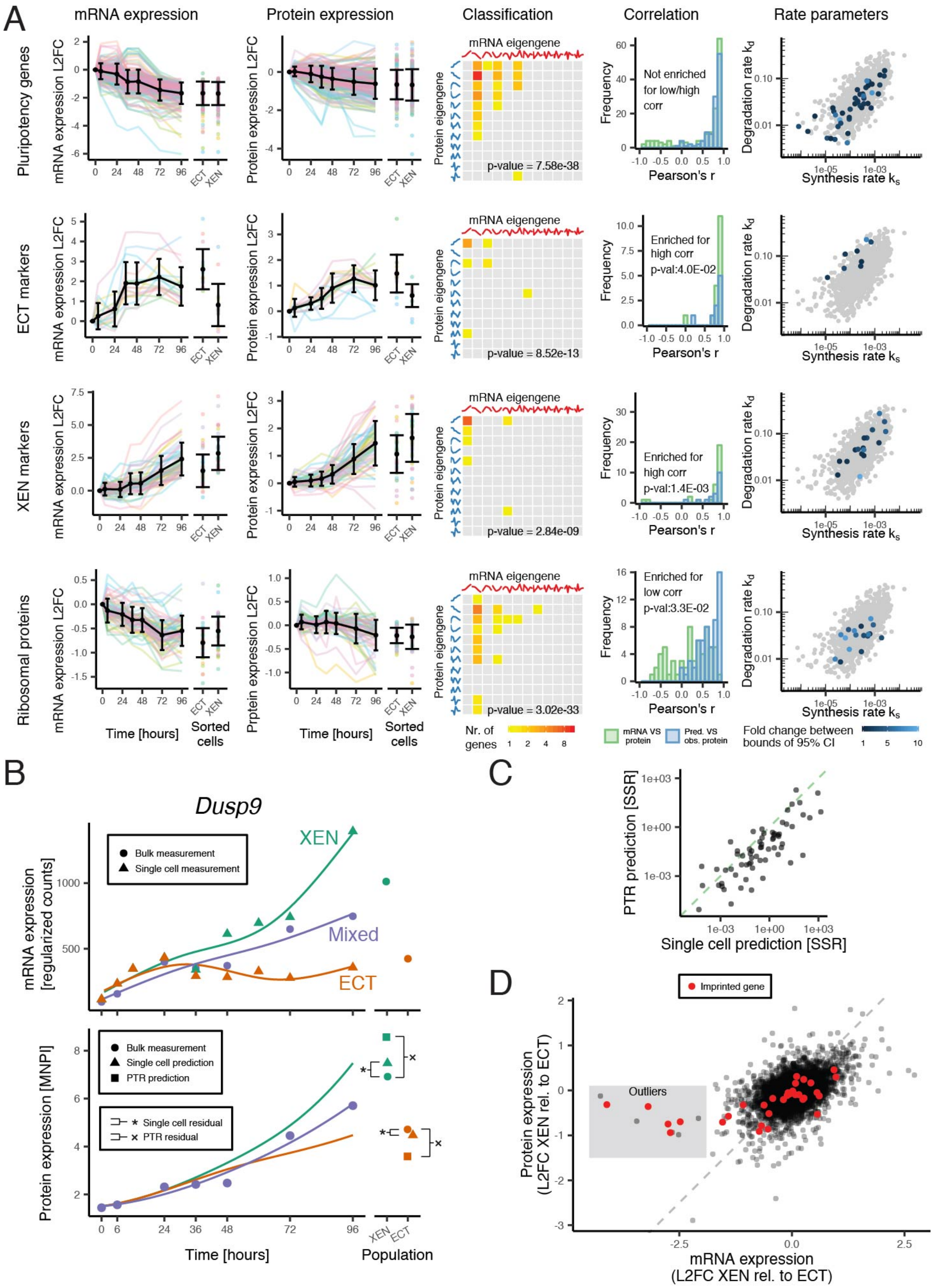
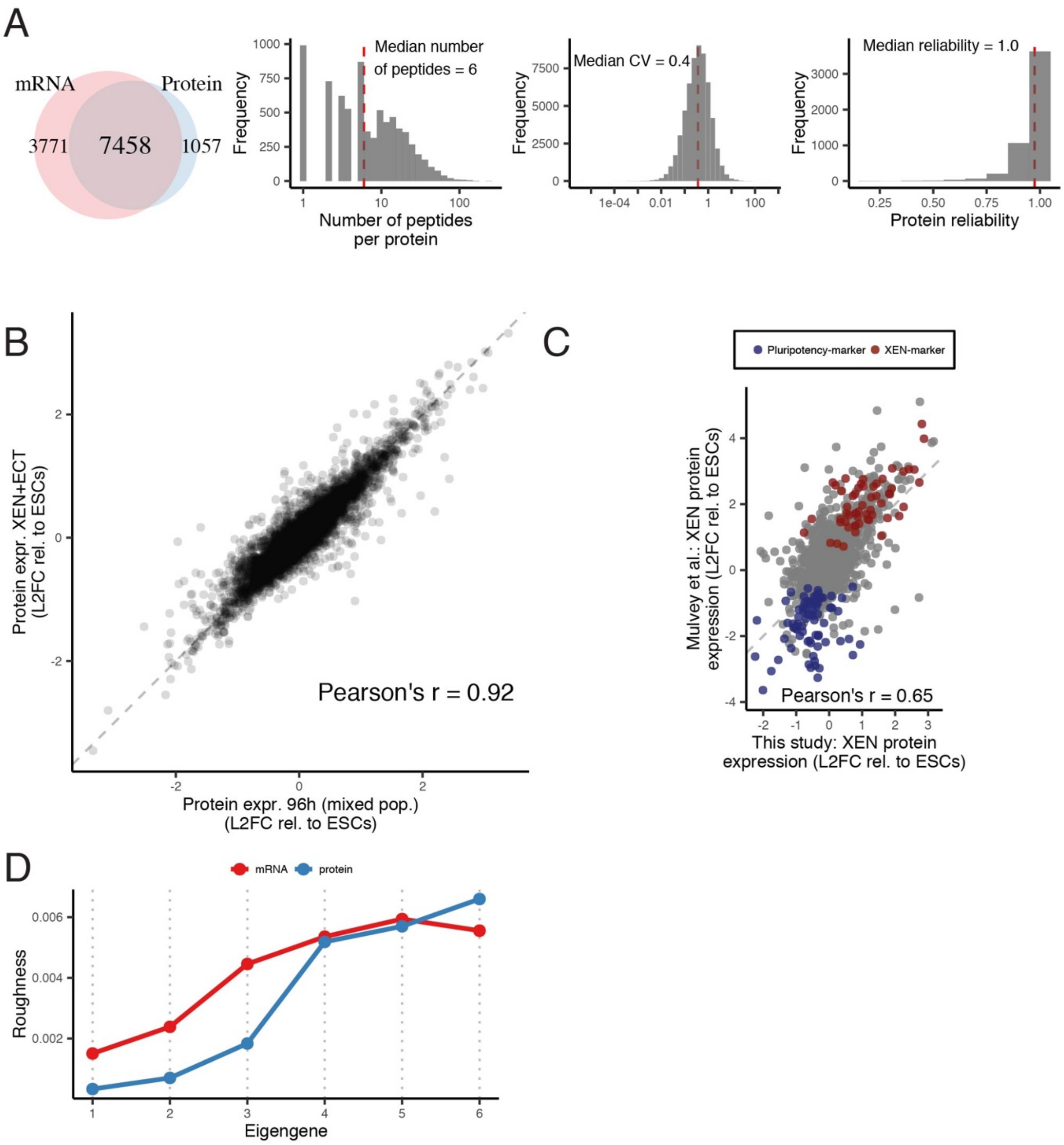

A
mRNA
Protein
R
k_s
P
k_d
Ø
B
Degenerate model (19%): *Cdk4*
Synthesis only (27%): *Lamb1*
Degradation only (21%): *Nanog*
Full model (33%): *Dnmt3a*
mRNA
Protein
Expression fold change
Time [hours]
C
Correlation of measured protein with:
mRNA
Kinetic model
Frequency
Pearson's r
D
mRNA
Protein
R
$k_s(t)$
P
k_d
Ø
E
Nptn
Poglut1
mRNA expression
CDS / 3'UTR
Protein expression
β = 9.24E-05
β = -4.62E-05
Fold change
L2FC
Time [hours]
F
Correlation of measured protein with:
mRNA
Kinetic model: k_s = constant
Kinetic model: k_s (t)
Frequency
Pearson's r

Figure 4


A
mRNA expression
Protein expression
Classification
Correlation
Rate parameters
Pluripotency genes
ECT markers
XEN markers
Ribosomal proteins
mRNA expression L2FC
Protein expression L2FC
Prptein expression L2FC
mRNA eigengene
Protein eigengene
p-value = 7.58e-38
p-value = 8.52e-13
p-value = 2.84e-09
p-value = 3.02e-33
Frequency
Pearson's r
Not enriched for low/high corr
Enriched for high corr p-val:4.0E-02
Enriched for high corr p-val:1.4E-03
Enriched for low corr p-val:3.3E-02
Degradation rate k_d
Synthesis rate k_s
Time [hours]
Sorted cells
ECT
XEN
Nr. of genes
mRNA VS protein
Pred. VS obs. protein
Fold change between bounds of 95% CI
B
Dusp9
mRNA expression [regularized counts]
Bulk measurement
Single cell measurement
XEN
Mixed
ECT
Protein expression [MNPI]
Single cell prediction
PTR prediction
Single cell residual
PTR residual
Time [hours]
Population
C
PTR prediction [SSR]
Single cell prediction [SSR]
D
Imprinted gene
Outliers
Protein expression (L2FC XEN rel. to ECT)
mRNA expression (L2FC XEN rel. to ECT)

Supplementary Figures

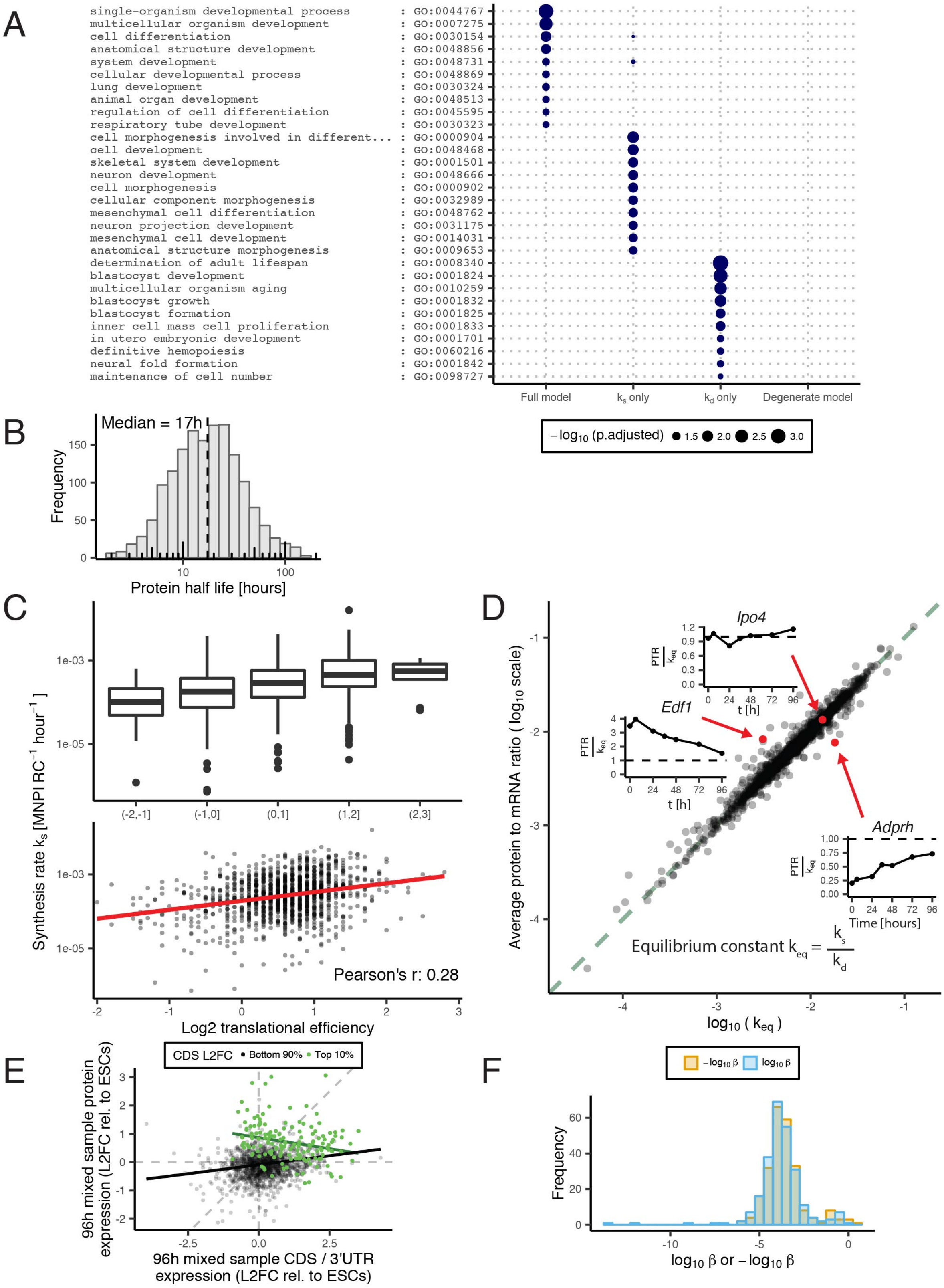


Supplementary Figure 1. Related to figures and 1 and 2. Protein quantification using TMT labeling is robust and reproduces previous results on embryo-derived XEN cells. mRNA eigengenes are more dynamic than protein eigengenes.

(A) From left to right: Venn diagram of the number of genes with quantified mRNA and protein levels (see Methods), distribution of the number of peptides used to quantify protein expression, distribution of the coefficient of variation (CV, SD/mean) of the mean-centered peptide intensities, distribution of the gene-wise protein reliability. The 7459 genes in the intersection are detected in all mRNA and protein samples.
(B) Protein expression of the 96 h sample (consisting of both XEN and ECT cells) compared with a sample mixed *in silico* from the independently generated purified XEN and ECT cell samples. L2FC: $\log_2$ fold-change.
(C) Protein expression in XEN cells relative to ESCs as measured in this study compared with *in vivo* derived XEN cells measured by Mulvey et al. (2015). Pluripotency- and XEN-marker gene sets were defined using a support vector machine learning algorithm. The pluripotency

set is significantly enriched in genes that are downregulated in our data (p-value = 4.0E-4) and the XEN-marker gene set is enriched in genes that are upregulated (p-value = 1.4E-4, gene set enrichment analysis).

(D) Roughness of mRNA and protein expression eigengenes. The roughness of a profile is defined as the average squared second derivative.



Supplementary Figure 2. Related to figure 3. The kinetic models can be related to biological functions and the inferred kinetic rates are biologically meaningful.

(A) Union of the top 10 significantly enriched *cellular differentiation* GO terms for genes fit best by each of the four kinetic models. False discovery rate = 10%.

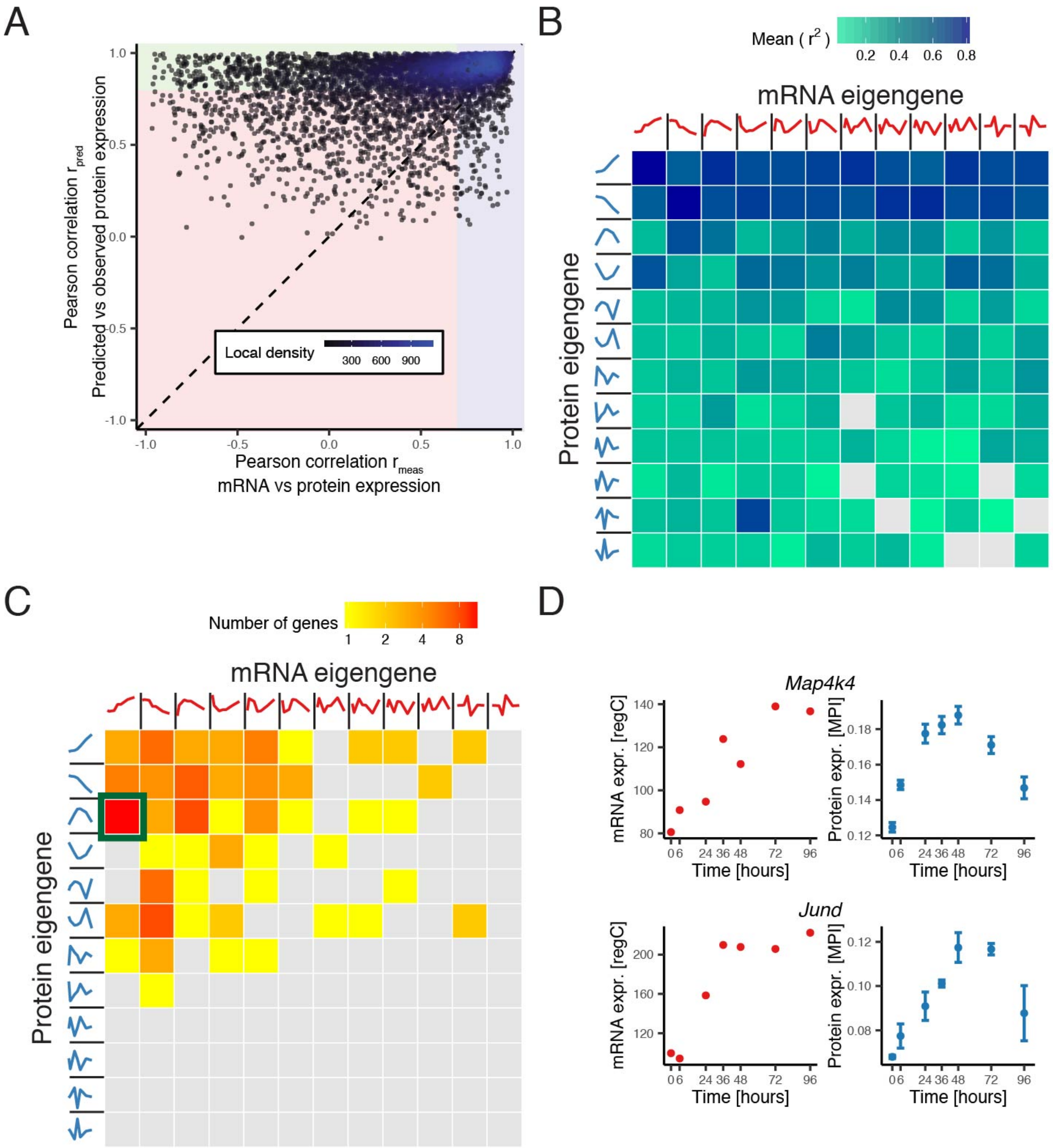
(B) Protein half life distribution for 1554 genes that were fit best by the full model (according to the BIC) and have precise estimates of the rates (upper and lower bound of the 95% confidence intervals (CIs) fall within a 10-fold range)

(C) Translational efficiency (TE) in mESCs from Ingolia et al. (2011) versus our synthesis rates. We show the rates for 1284 genes (intersection between data from Ingolia et al. (2011) and the1554 genes shown in B). Boxplots represent the binned TE with whiskers indicating 1.5 x IQR.

(D) Log_{10} protein to mRNA ratio (PTR) versus equilibrium constant ($k_{eq} = k_s / k_d$) for the 1554 genes described in B. Each data point is an individual gene. Genes that are at equilibrium (PTR = k_{eq}) are on the 1:1 line (green). Inserts: PTR relative to k_{eq} across time are shown for three example genes that are above, approximately on and below the 1:1 line.

(E) Ratio of CDS and 3'UTR expression versus protein expression in the 96h sample relative to ESCs. The genes with the highest CDS expression fold change are indicated in green. Solid lines indicate linear regression fits. CDS = coding DNA sequence, 3'UTR = 3' untranslated region.

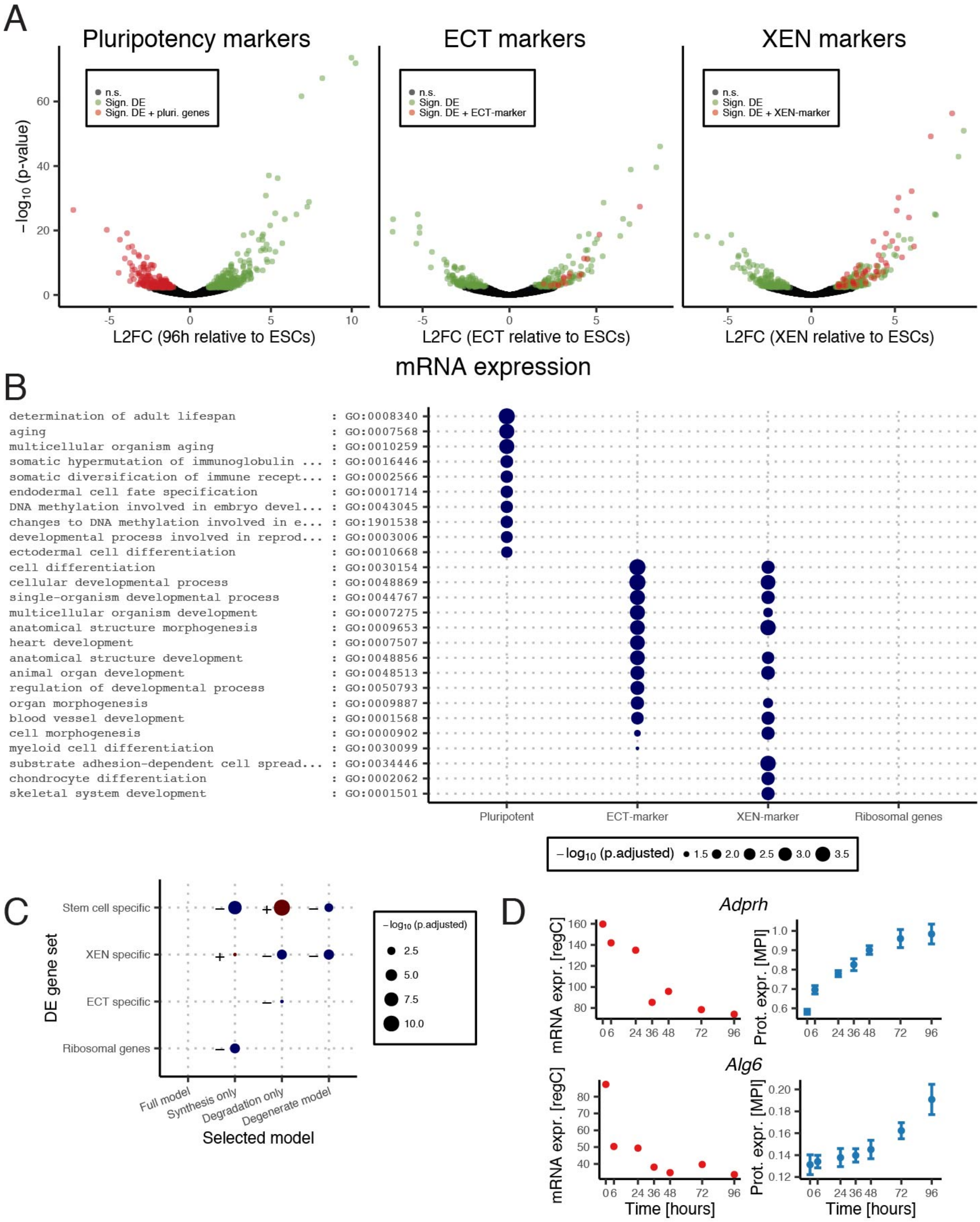
(F) Distribution of the parameter β of the extended model, which sets the strength of the influence of the CDS-3'UTR ratio on the synthesis rate. Shown are the values of β for the 492 genes that are improved by the extended kinetic model (according to the BIC).



Supplementary Figure 3. Related to figure 3. Genes in the MAPK signaling pathway are regulated dynamically at the protein level during differentiation.

(A) Pearson correlation between measured protein and mRNA (r_{meas}) versus Pearson correlation between measured and predicted protein (r_{pred}). Background coloring indicates: concordant genes with (high r_{meas}, blue), discordant genes that are not well-fit (low r_{meas} , low r_{pred} , red) and discordant genes that are well-fit (low r_{meas} , high r_{pred} , green). Here we consider genes with r_{meas} < 0.7 to be discordant (see Methods). To assure the the model prediction correlates substantially better with the measured protein than the measured mRNA we require r_{pred} >= 0.8 for a gene to be considered well-fit.
(B) Dominant eigengene classification of all 7459 genes. The color of a tile indicates the mean fraction of variance explained (mean r^2) by the best-fitting kinetic model for genes with a particular combination of dominant mRNA and protein eigengene.
(C) Dominant eigengene classification of the 368 genes that are not well-fit by the basic kinetic model (red area of A) and exhibit a bigger than median fold-change between the contribution of the dominant eigengene and the second most contributing eigengene. The color of a tile indicates the number of genes with a particular combination of dominant mRNA and protein eigengene. Enrichment analysis revealed an enrichment of MAPK signaling pathway genes in the tile highlighted in green (q-value = 1.8e-3).

(D) mRNA and protein expression profiles of two genes from the tile highlighted in C. Error bars: SEM. regC = regularized counts; MPI = mean peptide intensity.



Supplementary Figure 4. Related to figure 4. The different subtypes of the kinetic model are enriched in gene sets defined by the differentiation process

(A) Volcano plots (mRNA relative expression versus p-value for differential expression) for the 96 h sample, the ECT sample and the XEN sample. mRNA expression is always relative to the 0 h sample (ESCs). Genes colored in both red or green are significantly differentially expressed with a false discovery rate (FDR) of 10%. Only genes colored red are considered marker genes: pluripotency markers are down regulated in the 96 h sample, ECT and XEN markers are upregulated and have a minimum fold change of 2 compared with the other purified sample (see Methods).

(B) Union of the top 10 significantly enriched *cellular differentiation* GO terms for genes in each of the three DE gene sets and the ribosomal genes. FDR = 10 %.

(C) Overrepresentation (+ / blue) and underrepresentation (– / red) of the various subtypes of the basic kinetic model in the gene sets from B. (D) Genes in pluripotency gene set with upregulated protein expression. regC = regularized counts; MPI = mean peptide intensity.