Gene regulation in embryonic development

Berg, P.R. van den

3 KINETIC MODELING OF MULTI-OMICS DATA REVEALS MICRORNA-MEDIATED TRANSLATIONAL REGULATION IN STEM CELL DIFFERENTIATION

THIS CHAPTER IS BASED ON:
Patrick van den Berg, Noémie Bérenger-Currias, Marleen Feliksik, Esmée Adegeest, Mazène Hochane, Maria Mircea, Bogdan Budnik, Nikolai Slavov, Stefan Semrau. "Kinetic modeling of multi-omics data reveals microRNA-mediated translational regulation in stem cell differentiation". In: *Unpublished* (2020)[1]

Abstract

Stem cell differentiation is a highly dynamical process involving intricate gene regulatory mechanisms at multiple levels. A lack of detailed understanding of these mechanisms makes it challenging to improve existing differentiation protocols, which are gleaned from in vivo development and are typically slow and inefficient. The large majority of existing studies on differentiation has focused on transcriptional regulation, while the extent and mechanisms of translational regulation are much less explored. Here, we present a time-resolved, multi-omics study of retinoic-acid driven differentiation of mouse embryonic stem cells, comprising mass spectrometry, mRNA-sequencing of cytoplasmic and nuclear fractions, as well as micro-RNA sequencing. We develop a hierarchical kinetic rate model that allows us to integrate these datasets and explore the factors that determine protein levels. While the cytoplasmic-to-nuclear ratio of mRNA only has a minor effect, our model reveals microRNAs that have a significant influence on the translation of their putative targets. Multi-omics factor analysis finally identifies the major biological factors involved in the differentiation process. All in all, our study shows how a refined kinetic model, in conjunction with stringent model selection, can be used to discover regulatory mechanisms in a high-throughput manner, without the need for perturbations.

[1] S.S. and N.S. conceived the project. S.S. acquired funding. N.S. and B.B. supervised and performed the proteomics experiments. S.S., P.v.d.B, N.N. and M.F. performed all other experiments with support from E.A. and M.H.. P.v.d.B. analyzed, interpreted and modeled the data with assistance from M.M.. P.v.d.B. and S.S. wrote the manuscript.

3.1 Introduction

Much of the medical potential of pluripotent stem cells is due to their ability to differentiate into all tissue types of the adult body [2]. 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 [3, 4, 5, 6, 7]. Gene regulation occurring at the level of translation is much less explored. Most transcriptomics studies make the implicit assumption that mRNA levels are a good proxy for protein levels. It has been shown that in steady state, roughly 40% of protein variability across the proteome, can be explained by differences in mRNA abundance ([8]). Models of the steady-state protein to mRNA ratio (PTR) can explain up to two-thirds of the variability when taking transcript sequence features -such as coding sequence length or amino acid frequencies- into account [9]. In highly dynamical systems, such as differentiating stem cells, protein abundance is typically modeled with differential equations. These models are different from steady-state models in that they cannot explain absolute protein levels, but they can be used to infer kinetic rates for protein synthesis and degradation [10, 11, 12]. Here, we show that such models can also be used to reveal regulatory mechanisms during stem cell differentiation in an unbiased, high-throughput manner. We collected a multi-omics dataset of retinoic acid (RA) driven differentiation of mouse embryonic stem cells. Samples taken over a period of 96h were subjected to: mass spectrometry, bulk RNA-sequencing of nuclear and cytoplasmic fractions, as well as small RNA sequencing to quantify micro-RNA (miR) abundance. To model protein dynamics we refined a birth-death model by considering explicitly the cytoplasmic-to-nuclear ratio of mRNA abundance and the influence of certain technical artifacts related to mass spectrometry. By modeling the influence of miRs on protein synthesis, we identified several miR that likely have a significant influence on protein regulation. Finally, we used multi-omics factor analysis (MOFA) to reveal the overall relevance of translational regulation for in vitro differentiation.

3.2 Results

3.2.1 Pervasive discordance between RNA and protein in retinoic acid driven mESC differentiation

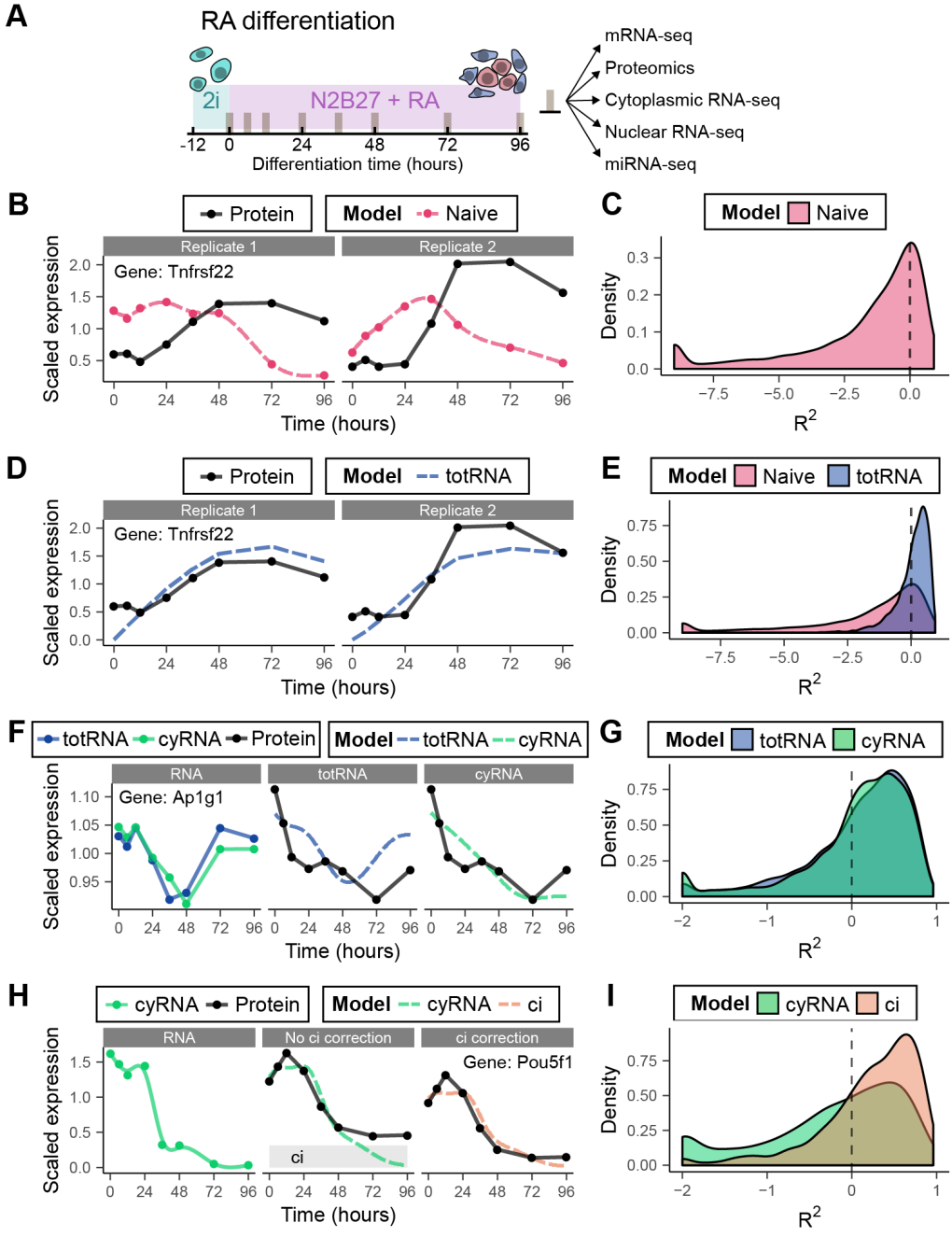
We used RA differentiation of mouse emrbyonic stem cells (ESCs) as a generic model for in vitro differentiation. Previously, we characterized this differentiation assay in detail at the transcriptional level by single-cell RNA-seq [3] In particular, we have shown that within 96 h of RA exposure, mouse ESCs bifurcate into an extraembryonic endoderm-like and an ectoderm-like cell type (XEN and ECT respectively). Here we collected samples during an

RA differentiation time course (Fig 1A). For each time point we quantified total 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 of 6271 of genes (Fig S1A-E) for 8 time points in duplicate. After correction for batch effects due to different runs and sequencing methods (Fig 1H), we achieved highly similar results for the two biological replicates. To investigate, in how far protein expression can be predicted from RNA expression, we started with the simplest conceivable model (termed *naive* here), which assumes that protein expression is identical to RNA scaled with a constant factor. This model is justified if RNA expression changes slowly on the time scale of protein degradation, resulting in a quasi-steady state. Consequently, the PTR would be approximately constant over the time course. To test this model, we scaled both protein and RNA to their respective means, which should result in a constant PTR of 1, if the naive model is valid. We observed that for a large fraction of genes the naive model is inaccurate, resulting in a low coefficient of determination (R^2) and low correlation coefficient (Fig 1C, Fig 2A). For particular genes we could even observe significant anti-correlation between RNA and protein (Fig 1B). This result shows that the assumptions of the naive model are likely wrong for the majority of genes and a more sophisticated model is necessary to explain the relationship between RNA and protein.

3.2.2 Protein turnover model explains RNA-protein discordance for most genes

To relax the assumption that expression is in steady state, we next considered a kinetic model that implements a birth-death process for protein turnover (Eq 3.1). Similar models have been used previously to describe protein turnover during the stress response in yeast [10], as well as embryonic development of *Xenopus* [11] and *Drosophila* [12]. The birth-death model assumes constant rates for protein synthesis (k_s) and degradation (k_d). All processes related to protein production (translation, initiation, elongation, etc) are lumped into k_s, while k_d represents all processes leading to a reduction in protein levels (dilution due to cell division, active degradation, etc.). We do not consider simpler, degenerate models (without k_s and/or

Figure 1 *(following page)*. Protein turnover models outperform the naive model in predicting protein temporal profiles. (A) Schematic overview of RA differentiation time course and subsequent omics measurements. (B) Example fit of the naive model. The naive model is a smoothing spline fit of RNA scaled to match the mean protein expression. (C) R^2 distribution of the naive model. (D) Example fit of the totRNA model. (E) R^2 distribution of the totRNA model model. (F) Example fit of the totRNA and fullRNA model, replicate 1. (G) R^2 distribution of the cyRNA model. (H) Example fit of the ci model, replicate 1. The height of the grey bar indicates the fitted ci parameter. (I) R^2 distributions of the ci model. Only genes that are improved by the ci model are shown. The full distribution of genes is shown in Fig S2E. Some genes with extremely low R^2 values are set to the minimum value of the plot for clarity. Corresponding Pearson's r distributions are plotted in Fig S2.


A
RA differentiation
2i
N2B27 + RA
-12
0
24
48
72
96
Differentiation time (hours)
mRNA-seq
Proteomics
Cytoplasmic RNA-seq
Nuclear RNA-seq
miRNA-seq
B
Protein
Model
Naive
Replicate 1
Replicate 2
Gene: Tnfrsf22
Scaled expression
Time (hours)
C
Model
Naive
Density
R2
D
Protein
Model
totRNA
Replicate 1
Replicate 2
Gene: Tnfrsf22
Scaled expression
Time (hours)
E
Model
Naive
totRNA
Density
R2
F
totRNA
cyRNA
Protein
Model
totRNA
cyRNA
RNA
totRNA
cyRNA
Gene: Ap1g1
Scaled expression
Time (hours)
G
Model
totRNA
cyRNA
Density
R2
H
cyRNA
Protein
Model
cyRNA
ci
RNA
No ci correction
ci correction
Gene: Pou5f1
ci
Scaled expression
Time (hours)
I
Model
cyRNA
ci
Density
R2

k_d [11]), because these models are not biologically meaningful in our opinion. It seems reasonable to assume that synthesis and degradation always occur to some degree. To reduce the influence of uninformative small fluctuations, we applied a smoothing spline to the expression data prior to inferring model parameters by non-linear least-squares fitting. Compared to the naive model R^2 and correlation improved markedly (Fig 1DE, Fig 2B), which might be expected given the increase in model flexibility. To correct for a difference in the number of fit parameters and thus compare model performance fairly, we used the Bayesian information criterion (BIC) (see Methods). According to the BIC, 3551 out of 4580 genes were better fit by the kinetic model. These genes are thus likely out of steady state for the duration of the experiment as a result of the differentiation cue.

$$\begin{aligned} P^g(t) &= k_s^g \cdot R^g(t) - k_d^g \cdot P^g(t) \\ k_s^g &\geq 0, k_d^g \geq 0 \end{aligned} \tag{3.1}$$

In summary, these results showed that a simple birth-death model outperforms the naive model of protein turnover.

Despite the overall improvement observed with the kinetic model, many genes were still not properly fit. We would like to interpret the remaining discrepancies as signs of biologically interesting, dynamic regulation. To be able to do so, we had to exclude technical limitations of our measurements as possible explanations. We first considered the subcellular localization of mRNA. In our first experiment we measured total RNA, whereas only cytoplasmic mRNA is available for translation. Nuclear retention of mRNA was found to reduce variability in cytoplasmic mRNA concentration and thereby protein synthesis. Moreover, specific genes are retained in the nucleus as a form of translational regulation [13, 14, 15]. To measure the cytoplasmic mRNA fraction of each gene, we repeated the differentiation experiment in triplicate and separated cell lysates into a nuclear and cytoplasmic fraction before performing RNA-seq. To obtain a global scaling factor between cytoplasmic and nuclear expression, we regressed totRNA reads, measured previously, on nuclear RNA (nuRNA) and cyRNA reads across all genes (see methods) . Then the cytoplasmic fraction C was calculated for each gene and each time point. To our surprise, C did not vary substantially between genes (Mean= 0.817, Std=0.0161, subset of 3,563 genes without any missing values) (Fig S1F). In addition, C also did not fluctuate much in time for individual genes (Fig S1G). Despite the low variability of C, we incorporated this parameter into our model (Eq 3.2). As expected, adding C brought overall only a very subtle improvement (Fig 1G), although for individual cases, the improvement can be quite significant (Fig 1F). We opted to fit further models including the cytoplasmic fraction due to the overall slightly better performance.

$$\begin{aligned} P^g(t) &= k_s^g \cdot C^g(t) \cdot R^g(t) - k_d^g \cdot P^g(t) \\ 0 &\leq C^g(t) \leq 1 \end{aligned} \tag{3.2}$$

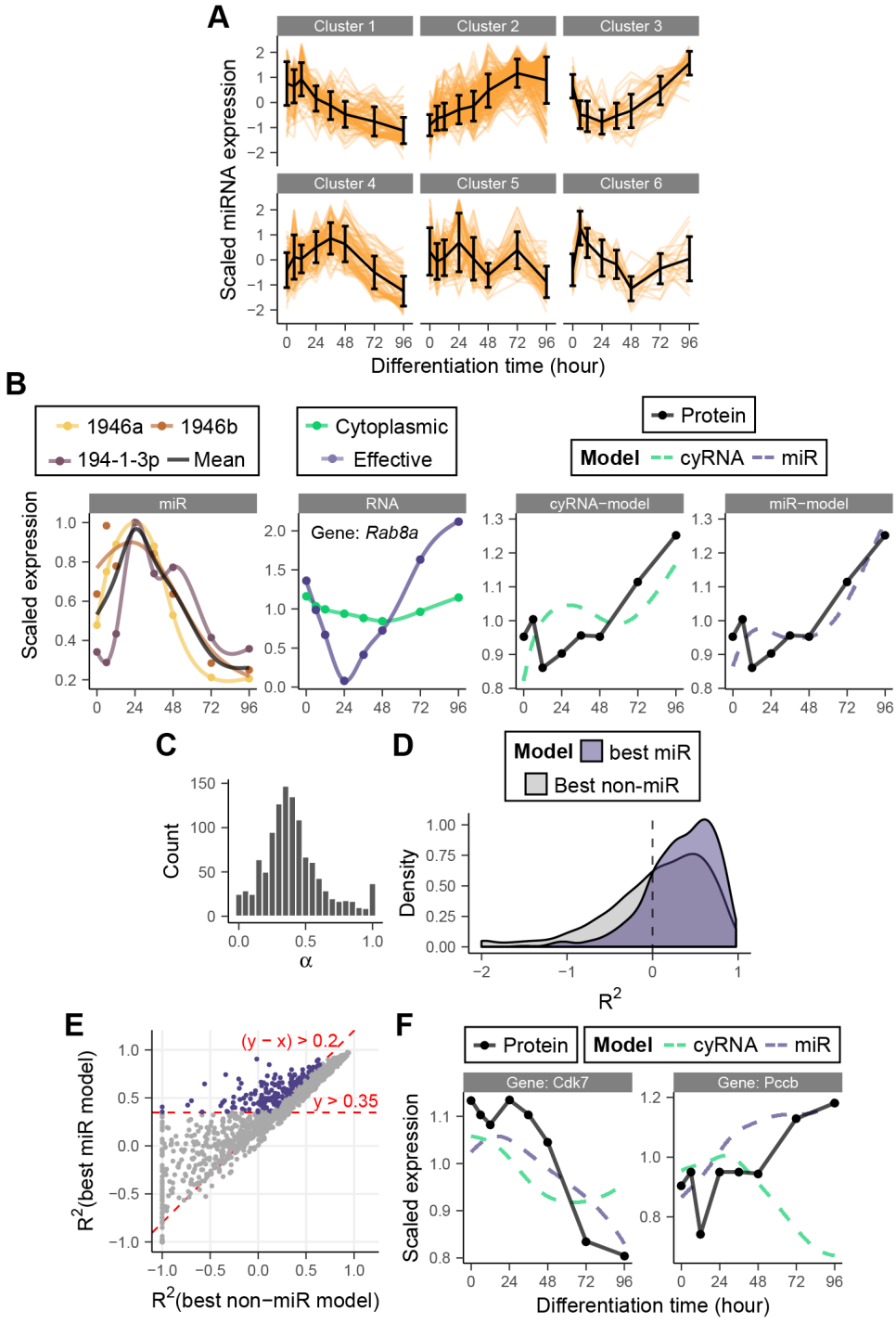
Another important technical limitation was inherent to the proteomics method we employed. TMT-based proteomics suffers from coisolation interference, a process in which two peptides are co-isolated in the second MS step. The contaminating peptide interferes with the quantification of the peptide of interest. To model this contamination, we added an additional parameter (ci) to the model (Eq 3.3), which we assume to be constant for all TMT tags (i.e. time points). Effectively, including this parameter allows protein expression to have a bigger dynamic range, which can improve the fit for certain genes significantly (Fig 1H-I, Fig 1DE). Judged by the BIC, 598 genes were fit better including ci. All in all, this result shows that it is very important to consider co-isolation interference.

$$\begin{aligned} P^g(t) + ci^g = k_s^g \cdot C^g(t) \cdot R^g(t) - k_d^g \cdot (P^g(t) + ci^g) \\ 0 \leq ci^g \leq \min\{P^g(t)\} \end{aligned} \tag{3.3}$$

3.2.3 Including miRs improves model performance and identifies miR-gene interactions

Having ruled out major systematic errors, we were in a position to explore biological explanations for the remaining discordance between RNA and protein. We chose to explore miRs since they are known to play an important role in gene regulation during differentiation. Specifically, we wanted to study the influence of miRNAs on protein translation initiation. In order to study the role of miRs in our system we repeated the differentiation assay and measured the miRnome by small RNA-seq in quadruplicate. We quantified around 1000 mature miRs per time point (Fig 1ACE). To identify possible miR interactions, we used the list of predicted targets provided by TargetScanMouse [16]. We further limited the number of miRs per mRNA by filtering the miR-gene interactions leniently using the context score (Fig 1D). In the end we retained 4527 genes with 560 unique mature miRs and 45,882 potential interactions

Figure 2 *(following page).* The addition of miRs further improves the protein turnover model for a subset of genes and reveals novel candidate miR-gene interactions. (A) Expression profiles of 560 miRs in six clusters. (B) Example fit of miR model for the gene *Rab8a*, replicate 1. First panel: expression of the assigned miRs of a single cluster. Colored lines are individual smoothing spline fits. Second panel: Cytoplasmic RNA expression and the effective RNA concentration available for translation (see Materials and methods). Solid lines represent smoothing splines. Third/fourth panel: cyRNA and miR model fits. (C) Distribution of inferred α for genes that benefit from miR model. (D) R^2 distribution of the miR model and the next best model (either naive, totRNA, cyRNA or ci). Only genes that benefit from the miR model are shown. Some genes with extremely low R^2 values are set to the minimum value of the plot for clarity. Corresponding Pearson's r distributions are shown in Fig S2F. (E) R^2 distributions of (D) compared in a scatter plot. Colored dots are defined by the cutoffs indicated in red and represent a subset of genes with a miR-gene interaction of higher confidence. Some genes with extremely low R^2 values are set to the minimum value of the plot for clarity. (F) miR model fits of two genes (*Cdk7, Pccb*) from the subset highlighted in (E).


A
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Scaled miRNA expression
Differentiation time (hour)
B
1946a
1946b
194-1-3p
Mean
Cytoplasmic
Effective
Protein
Model
cyRNA
miR
miR
RNA
cyRNA−model
miR−model
Scaled expression
Gene: Rab8a
C
Count
α
D
Model
best miR
Best non-miR
Density
R2
E
(y − x) > 0.2
y > 0.35
R2(best miR model)
R2(best non−miR model)
F
Protein
Model
cyRNA
miR
Gene: Cdk7
Gene: Pccb
Scaled expression
Differentiation time (hour)

(Fig 1DE).

If multiple miRs with similar temporal profiles target the same gene, we considered them to be indistinguishable. Therefore we globally clustered miRs into six clusters with similar temporal profiles (Fig 2A) and averaged miRs targeting the same gene per cluster (Fig 2B). To keep the model simple we assumed that the inhibitory effect of miRs on protein translation grows linearly with miR abundance (Eq 3.4). We fit one of the six miR clusters at a time and identified the improvement in model performance for each cluster and each gene.

$$\begin{aligned} P_m^g(t) = k_s^g \cdot (1 - \alpha_m^g \cdot M_m^g) \cdot C^g(t) \cdot R^g(t) - k_d^g \cdot P^g(t) \\ 0 < \alpha_m^g \leq 1 \end{aligned} \tag{3.4}$$

Including miRs greatly improved the fits for some genes, especially when there is a transient discordance between RNA and protein expression (Fig 2B). Typically, the "effective" mRNA abundance (cytoplasmic mRNA corrected for miR effects) was more dynamic than nominal mRNA abundance. For many of the genes that benefit from the addition of miRs, their influence is typically large. For these genes, 50% of translation is blocked on average at peak miR expression (Fig 2C). Overall, the addition of miRs significantly improved the coefficient of determination for a quarter of the genes (Fig 2D).

To use the model for identifying novel miR-gene interactions, we ranked the genes by the quality of the model fit and model performance improvement compared to the simpler models without miRs (Fig 2E, Suppl Table 2). Among this list of candidate genes we selected seven genes and their putative miRs (Fig 2BF, Fig S3), whose interaction we intend to validate in a follow-up study. (*Rab8a*, *Cdk7*, *Pccb*, *Acad8*, *Mfge8*, *Eif4h* and *Srgap2*).

3.2.4 The best protein turnover model explains 45% of total protein variance

While each model refinement introduced above improved model performance overall, each discussed model was optimal for a subset of genes, judged by the BIC (Fig 3A). In about 16% of cases the naive mode was optimal, meaning that for these genes none of the protein turnover models improve prediction by a significant amount. 25% and 26% of genes are best predicted with the kinetic model without or with considering mRNA localization, respectively. So for 51% of genes, protein expression is out of steady state, but explainable by a simple model with fixed synthesis and degradation rates. For 8% of genes the model including co-isolation interference was optimal. The increased relative dynamic range due to subtracting a constant increased the fit for these genes significantly. Finally, 25% of genes were fit optimally with one of the miR clusters, meaning that translational regulation plays a significant role for these genes. All things considered, 84% of genes were insufficiently described by RNA alone, leading to a very significant lack of variance explained (Fig 3BC). Therefore, it seems in general not advisable to consider mRNA abundance a good proxy for protein levels in a highly dynamical setting.

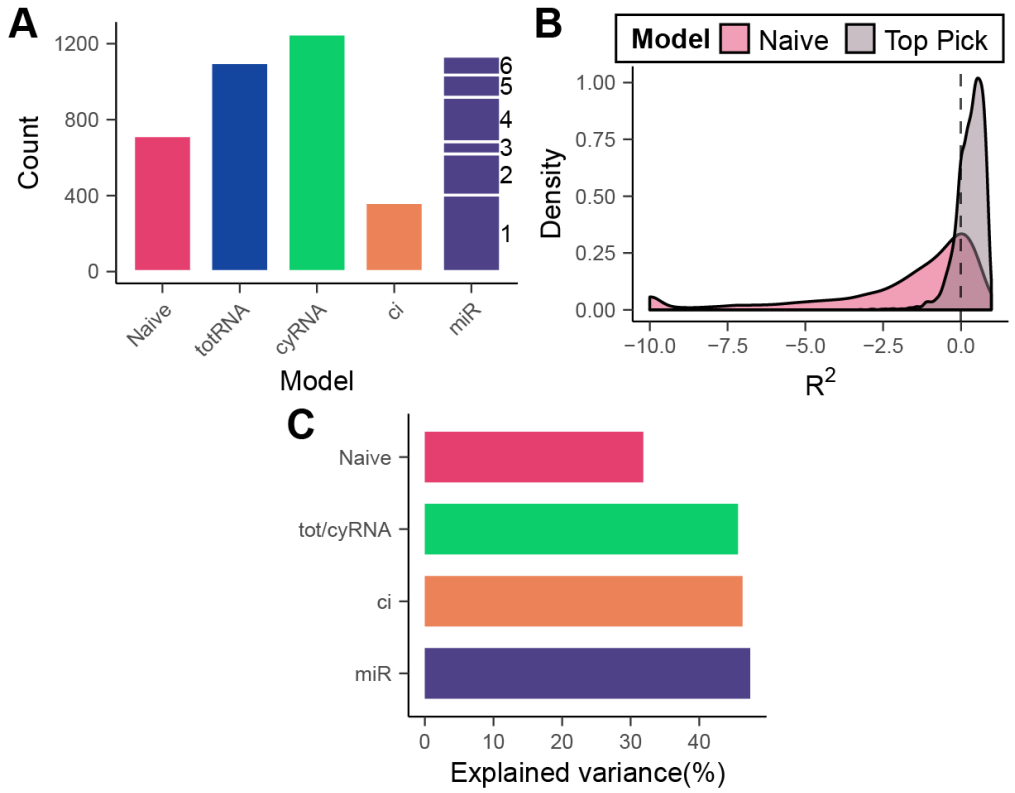


Figure 3. Selecting the optimal model on a gene-by-gene basis increases the total explained variance of protein expression from 30% to 50%. (A) Assignment of the optimal model for each gene based on BIC. The number next to the miR bar indicates the miR cluster giving the best fit. (B) R^2 distribution of the optimal fits from (A) and their naive model counterpart. Some genes with extremely low R^2 values are set to the minimum value of the plot for clarity. (C) Median percentage of protein variance explained by each model, picking among the best models progressively. Fits with negative R^2 were ignored.

3.2.5 Multi-omics factor analysis reveals global factors driving translational regulation in mESC differentiation

The above analyses focused on how individual protein turnover is regulated. In order to describe translational regulation in our system in a more comprehensive manner we performed MOFA. MOFA is an extension of factor analysis which integrates omics data from different sources, like the genome, transcriptome or metabolome. MOFA extracts low dimensional, latent factors that represent the biological processes underlying the variation observed in the data. In order to identify factors involved in translation only, we used two datasets: PTR and miR abundance (Fig 4A). MOFA is typically run directly on gene expression, but here we used the PTR, because we were most interested in explaining how post-transcriptional

regulation can drive fluctuations away from steady state. From this analysis we obtained six driving factors of translational regulation (Fig 4B).

The six factors had distinct temporal trends (Fig 4C), which we used to interpret their biological identity. Since MOFA sorts the factors in order of variance explained and the first three factors had the simplest trends we restricted our interpretation to these three factors (Fig 4D). The genes contributing most to factor 1 are enriched for gene ontology (GO) terms related to development, differentiation, cell cycle and nucleotide synthesis (Fig 4E). Thus, we interpret this factor as the main differentiation program which includes changes in metabolism. This interpretation is supported by the miR that contribute most to factor 1 (Table 3.1): Several of these miRs are known to be involved in differentiation. The let-7 family in particular is known to play an important role in embryonic stem cell pluripotency and self-renewal. Factor 2 is enriched for GO terms in morphogenesis, cell adhesion and signaling. It appears to be the factor driving the specification of the cell type as a response to the external signals. Top miRs of this factor are involved in multiple differentiation paths (osteocytes, adipocytes, trophectoderm, neurons, see Table 3.1). Finally, Factor 3 appears to be related to epithelial-mesenchymal transition (EMT) and mesenchymal-epithelial transition (MET) as many of the top miRs are involved in EMT various types of cancers. Moreover, one of the few enriched GO terms is "epithelial tube morphogenesis". All in all, this analysis shows that meaningful biological factors can be revealed by considering protein synthesis and its regulation.

Factor	Rank	miR symbol	Short description
1	$\bigtriangledown$ 1	miR-1843a-5p	Differentially expressed in traumatic brain injury [17].
1	$\bigtriangledown$ 2	miR-27b-5p	Clustered with miR-23b. Upregulated in *Smad4* knockout cardiomyocytes, involved in cardiac hypertrophy [18]. Induces EMT in gastric cancer [19].
1	$\bigtriangledown$ 3	miR-23b-5p	Clustered with 27b. Regulates osteoclast differentiation [20]. Attenuates glucose-mediated EMT in diabetic nephropathy [21].
1	$\bigtriangledown$ 4	let-7d-3p	let-7 family is involved in pluripotency and self-renewal, and is differentially expressed between different ESC states [22]. 7d: inhibitions of leads to EMT in idiopathic pulmonary fibrosis [23].
1	$\bigtriangledown$ 5	let-7f-5p	
1	$\bigtriangleup$ 1	miR-34c-5p	Downregulation of this miR promotes EMT in breast cancer.[24]
1	$\bigtriangleup$ 2	miR-34c-3p	
1	$\bigtriangleup$ 3	miR-10a-5p	Critical for smooth muscle cell differentiation from mESC [25].
1	$\bigtriangleup$ 4	miR-9-3p	Involved in neurogenesis [26]. Suppressor of EMT in nasopharyngeal carcinoma. [27]
1	$\bigtriangleup$ 5	miR-34b-5p	See miR-34c.
2	$\bigtriangledown$ 1	miR-7b-5p	Represses self-renewal [28]. Reverses EMT in breast cancer through *STAT3* [29].

Continued on next page

Continued from previous page

Factor	Rank	miR symbol	Short description
2	▽ 2	miR-3058-3p	
2	▽ 3	miR-195a-3p	Inhibits adipocyte differentiation [30]. Inhibits EMT in Prostate cancer through *FGF2* [31]. Inhibits EMT in colorectal cancer through *NOTCH2* [32].
2	▽ 4	miR-3095-5p	
2	▽ 5	miR-187-3p	Inhibits osteogenic differentiation [33]. Inhibits EMT in hepatocellular carcinoma [34].
2	△ 1	miR-297c-3p	
2	△ 2	miR-297a-3p	Involved in trophectoderm specification in mouse [35].
2	△ 3	miR-297b-3p	
2	△ 4	miR-466f-3p	In cluster with each other. Inhibits *NeuroD1*, which is required for neuron differentiation [36].
2	△ 5	miR-669f-3p	
3	▽ 1	miR-770-3p	Inhibits EMT in non-small cell lung cancer [37].
3	▽ 2	miR-760-3p	Inhibits EMT in breast cancer [38].
3	▽ 3	miR-1306-5p	Involved in hepatocellular carcinoma, regulates *Snail*-mediated metastasis [39].
3	▽ 4	miR-301b-5p	Promotes proliferation, mobility and EMT in bladder cancer by targeting *EGR1* [40].
3	▽ 5	miR-369-5p	The -3p variant targets *Sox4* [41].
3	△ 1	miR-452-3p	Inhibits EMT in hepatocellular carcinoma through TGF-β1 [42].
3	△ 2	miR-340-5p	Targets *Bcl-w* and *Sox2* and inhibitions of miR promotes cancer progression [43].
3	△ 3	miR-186-3p	Affects EMT through Cdc42 in lung cancer [44].
3	△ 4	miR-700-5p	
3	△ 5	miR-106a-5p	Downregulates Twist1 which causes EMT [45].

Table 3.1. Top 5 negative and positive miRs for the first three MOFA factors.

3.3 Discussion

Widespread discordance between steady-state protein and mRNA levels has been observed in several mammalian systems [46, 47, 48]. Importantly, low correlation does not immediately imply a significant role of gene-specific regulation, as technical noise tends to reduce the observed correlation and conventional correction schemes typically ignore the effect of systematic, correlated errors [49] Edfors et al. [48] showed recently that the PTR for a specific gene is constant across several tissues [48]. While the PTR might allow the prediction of absolute protein levels, it is unable to capture relative changes over time or relative differences

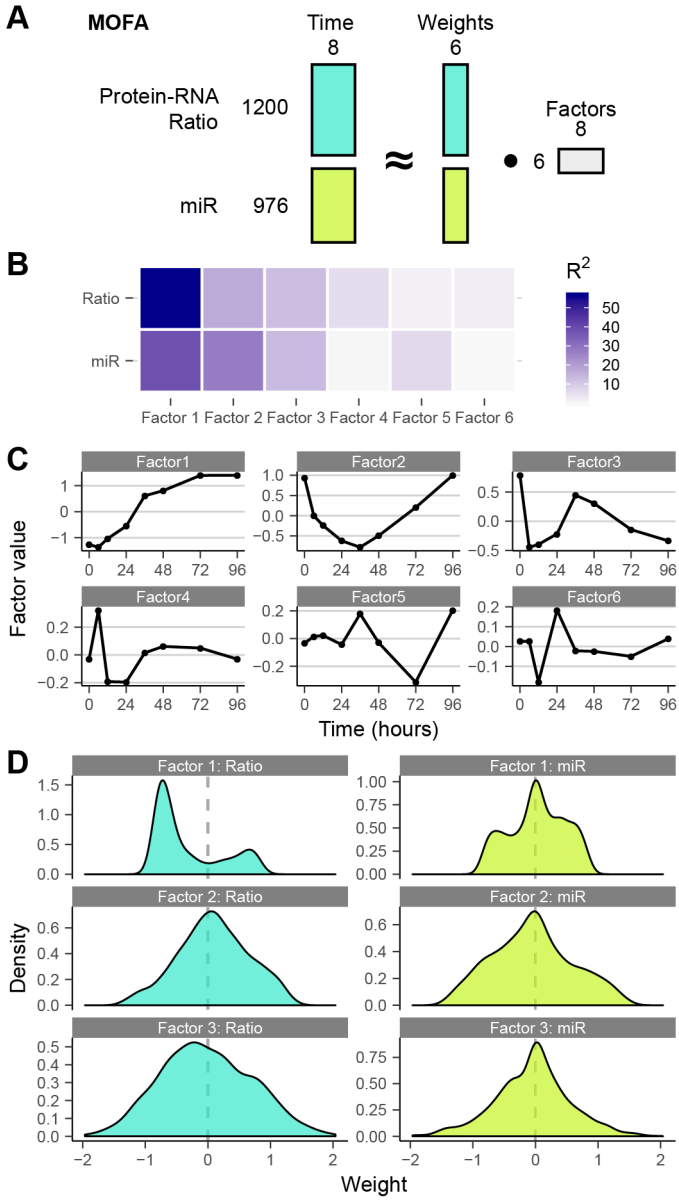


Figure 4. Multi-omics factor analysis uncovers underlying factors that drive translational regulation in mESC differentiation. **(Continued on next page)**

Figure 4. Multi-omics factor analysis uncovers underlying factors that drive translational regulation in mESC differentiation. (Continued from last page) (A) Schematic of the matrices used in MOFA analysis. (B) Variance explained (R^2) by each of the six factors. (C) Temporal profile of each factor. (D) Distributions of the factor weights for the first three factors. (E) GO term over and under representation by the top (+) and bottom (-) 120 genes of the first three factors. The top 10 enriched GO terms (with p<0.1) is shown for each factor and direction. Test performed is elim with KS statistic, see Materials and methods).

between tissues [50].

To identify dynamic translational regulation during stem cell differentiation, we therefore collected a time-resolved gene expression dataset and found overall low correlation between mRNA and protein abundance across time. Such low correlation has been observed recently in several other systems, in particular: *Xenopus* development [11], *C. elegans* development [51], macrophage differentiation [52] and mouse ESC differentiation [53]. While the lack of strong correlation is typically interpreted as a sign of (post) translational regulation [51, 53], theoretical work showed that a simple delay between mRNA and protein production can lead to a reduction in gene-wise correlation [54, 55] A simple model with constant kinetic rates explained the protein dynamics of a third of all genes during stress response in yeast [10] and of 75% of all genes in *Xenopus* development [11]. In our system, 3551 out of 4580 genes were explained better by this model, compared to a naive model which assumes a constant PTR.

To explain the remaining discordance we explored the cytoplasmic-to-nuclear ratio of mRNA abundance, but did not find a strong effect. On the other hand, including a parameter modeling co-isolation interference markedly improved the fit for some genes. We therefore posit that co-isolation interference should be included in any kinetic model when TMT labeling is used for multiplexing the mass spectrometry measurements.

miRs have been identified as a key regulator of stem cell pluripotency and differentiation [56]. For example, members of the the let-7 and miR-290 families have been implied as drivers for differentiation of ESCs as well as in the maintenance of pluripotency [57, 22, 56, 58] To find putative targets of miRs, various computational methods, typically based on sequence complementarity and conservation, have been developed [16, 59, 60]. These methods predict hundreds of thousands of interactions, among which are likely many false positives. The gold standard for validation, the luciferase assay, is time consuming, which means that the majority of potential interactions have not been verified. To our knowledge, there is currently no high-throughput experimental method to identify miR-mediated translational regulation in a genome- and miRnome-wide manner. We believe that our modeling approach is able to reduce the number of potential interactions to a much smaller set, which can be easily validated by conventional methods.

In addition to the possibility to infer regulatory interactions between different molecular players, multi-omics data sets are also useful to identify major driving factors of biological processes in development and disease. A living cell is typically considered to be a highly complex dynamical system that defies many traditional modeling approaches due to the large amount of unobserved or indeterminable parameters. There is, however, the hope that many biological processes in fact occur on low-dimensional manifolds within the high-dimensional space needed to describe the state of a cell. Multi-omics measurements will allow us to ascertain if there are in fact such manifolds, which would significantly simplify a complete quantitative understanding of biological dynamics. Our study indicates the presence of at least 3 factors that co-regulate miR and protein abundance during differentiation. To unravel how co-regulation is achieved molecularly and how the factors can be perturbed

to in vitro differentiation are fascinating challenges for the future.

3.4 Materials and methods

3.4.1 Cell culture

E14 mouse embryonic stem cells were cultured as previously described [3]. Briefly, cells were grown in modified 2i medium [61]: 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).

3.4.2 Retinoic acid differentiation and sample collection

Retinoic acid induced differentiation was carried out exactly as described before [3]. Prior to differentiation cells were grown in 2i medium for at least 2 passages. Cells were seeded at 2.5e5 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 and spun down. Full RNA and cytoplastmic/nuclear RNA were always immediately extracted (RNeasy, Qiagen and SurePrep, Fisher Scientific, resp.) and the purified RNA was stored at -80C until RNA-sequencing was performed. For proteomics and miR-sequencing, pellets were flash frozen in liquid nitrogen and stored at -80C until further processing.

3.4.3 RNA and miR sequencing

The libraries for RNA sequencing were prepared under standard conditions using Illumina's TruSeq stranded mRNA sample preparation kit. The stranded single end libraries were sequenced using Illumina HiSeq at 40bp with an average read depth of 40 million reads per sample. Paired-end libraries for RNA sequencing were sequenced on an Illumina NextSeq 500 at 150bp per strand at a read depth of 10 million reads per sample. miRs were extracted from frozen pellets using miRNeasy (Qiagen) kit. Libraries for small RNA sequencing (miR sequencing) were prepared using NEBNext Small RNA Library Prep Set for Illumina (New England Biolabs) and were sequenced on an Illumina NovaSeq 600 at 150bp paired-end with a range of 4 to 15 million samples. Specification table for the sequencing strategy is available upon request. All sequencing data is available through GEO.

3.4.4 Mass spectrometry

Pelleted cells were lysed in 400 µl RIPA buffer, except for the sorted cells. Volumes of cell lysate corresponding to 100 µg protein per sample were digested with trypsin using a modified FASP protocol [62]. Subsequently each sample was labeled with TMT 10-plex, 6-plex or 11-plex reagent (Thermo Fisher)) according to the manufacturer's protocol. All labeled samples were combined into a set-sample. Which labels were assigned to each sample is specified in the specification table, which is available upon request. The labeled set-sample was fractionated by electrostatic repulsion-hydrophilic interaction 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.

3.4.5 RNA-seq processing

We used genome assembly mm10 release 93 from Ensembl. First an RSEM (v1.3.1) reference was created with default settings. Next we performed adapter and quality trimming with Trimmomatic (v0.38). Finally all reads were aligned with RSEM with STAR (v2.6.1a) with the option enabled for stranded libraries. Expected counts from RSEM were used as input for DESeq2 (v1.26) to obtain regularized $\log_2$ counts with stabilized variance to make comparisons between samples more reliable. From these values we obtained regularized counts and used these for all further analyses and as input for batch correction.

3.4.6 Proteomics processing

Peptide search was performed on peptides identified in full RNA seq data to increase specificity of the protein quantification with MaxQuant. Proteins were quantified from the peptide measurements in the evidence.txtoutputs. Reversed peptides and contaminants were removed. Each column in the file was then normalized to the mean. Some peptides for some samples were quantified multiple times, due to multiple mass-spectrometry runs or multiple tmt tags in the same sample mix. These values were averaged. Multiple peptides assigned

to a single ensembl gene ID were also averaged to obtain normalized protein expression that we used for batch correction.

3.4.7 Batch correction

We observed global expression differences in protein and rna expression depending on the seeding day which we decided to correct using the RemoveBatchEffect function from limma (v3.42.2). We applied it on the protein and totalRNA datasets separately with three different levels of batches, one each for the first replicate, the second replicate and for two samples we used to replace failed samples of the first replicate (Specification table available upon request). The resulting batch-corrected values were used as input for most further analysis (Fig S1H).

3.4.8 miR-seq processing

For alignment of miR-seq data we used the same genome release as above with miRnome release 22.1 from miRBase using the mature miR sequences. To prepare the reads we performed adapter and quality trimming with Trimmomatic and obtained a consensus forward sequence using both the forward and reverse read and PEAR (v0.9.6). We next ran bowtie-prepare from bowtie (v1.0.0-1). Finally we quantified each sample with the mapper.pl and quantifier.pl scripts from mirdeep2 (v2.0.1.2). The obtained counts were processed the same way as the RNA-seq data, but separately.

3.4.9 miR-gene interactions

We looked for putative miR-gene interaction using TargetScanMouse release 7.1. We filtered the "miR family" table for expressed miRs and expressed RNAs. We next did a lenient filter on the miR-gene interactions; we filtered out all interactions with a cumweightscore lower than -0.3. Finally, we wanted to keep only miRs with high dynamics over the time course and high reproducibility. To achieve this we calculated for each miR the Coefficient of variation (CV) across the mean miR expression of each time point and the mean of the CV's across the biological replicates. We fit a gaussian mixture model to these to values using mclust (v5.4.6) where each distribution has an equal diagonal shape, but with varying volumes ("VEE" modelNames option). We filtered the miR-gene interactions list for miRs from cluster 1 because they fit our high variance and high reproducibility criteria (Fig S1I). The final putative list of miR-gene interactions comprised 560 miRs and was used in the miR clustering and model fit (see below).

3.4.10 miR clustering

To cluster miRs into sets of similar temporal profiles, miR expression of the miRs in the interaction table was first averaged per time point. A miR to miR distance matrix was created with

1–Pearson correlation on $\log_2$-transformed values. This matrix was then used to perform hierarchical clustering with complete linkage (base R) and the resulting dendrogram was cut into 6 clusters.

3.4.11 C-fraction calculation

To obtain a per gene C-fraction we needed to map how cytoplasmic and nuclear sequencing reads relate to each other. We took total, cytoplasmic and nuclear RNA and removed genes that had any raw count lower than 10. We then took the top 500 genes with the lowest variance and, under the assumption that for these genes that C-fraction was the most stable, fit the following linear model: $R_{tot} = \beta_c \cdot R_c + \beta_n \cdot R_n$, where R_{tot} is total RNA, R_c is cytoplasmic RNA and R_n is nuclear RNA. T For each we applied the regularized $\log_2$ counts. The fit beta parameters map cytoplasmic and nuclear values to total RNA values and were 0.815 and 0.183 respectively. Then for each gene g we calculated the C-fraction: $C^g = \frac{\beta_c \cdot R_c^g}{\beta_c \cdot R_c^g + \beta_n \cdot R_n^g}$

3.4.12 Rate model fitting

We fit several rate models for every gene in our clean set of genes (Fig S1E). We first scaled totRNA and protein expression by dividing each by its replicate mean. We next with a smoothing spline (smooth.spline function, base R) to the totRNA data for each replicate with 7 degrees of freedom (DF). We opted for manually setting the DF over letting the function determine it because we observed that what seems like noise at the RNA level is sometimes replicated at the protein level. smooth.spline would sometimes oversimplify the dynamics of RNA and this would lead to bad fits at the protein level if the protein has more dynamics than the resulting spline. Perhaps counterintuitively, we deemed the fit more conservative if we forced high dynamics at the RNA level at the cost of introducing some noise. We let smooth.spline determine the DF for all other smoothing spline fits using leave-one-out-cross-validation. We fit smoothing splines to the C-fraction and multiplied this with the smooth totRNA to get smooth cyRNA. miRs that were assigned to each gene were first averaged over replicates and then divided by the miRs maximum value. Smooth splines were fit to each miR and then the smooth miRs for each miR cluster were averaged. We solved the differential equation using deSolve (v1.28), given a rate model, parameters, totRNA or cyRNA and a miR cluster. Parameters put into the solvers were $\log_2(k_{prod})$, $\log_2(k_{div})$, P_0, ci and α, depending on the differential equation being solved. $k_{prod} = k_s \cdot k_d$ and $k_{div} = k_s/k_d$, which are perpendicular to the original values on purpose because the optimization algorithms sometimes had difficulty finding an optimal fit, because covarying k_s and k_d may result in a very similar fit. We $\log_2$ transformed these values to give the optimization function more control over fitting it. P_0 is the protein concentration at t=0h, we opted for adding this as a model parameter instead of setting it to the observed concentration at t=0h and losing that value for fitting. ci is set to be between 0 and the minimum

observed protein expression, because the consolation interference cannot exceed what was actually observed. α is between 0 and 1 and describes the miR clusters' influence on translation rate. Since both α and the miR clusters' expression cannot exceed 1, only at peak miR cluster expression can translation be completely turned off in the model. To find the optimal parameters we fit using the optim function (base R). Sum of squared residuals (SSR) were minimized using the "L-BFGS-B" method of the optim function. However for both Eq 3.2 and 3.3 we first minimized: $SSR + 10 \cdot \log_2(k_{div}) \cdot 2$. Due to our scaling of RNA and protein $\log_2(k_{div})$ is expected to be close to 0 so we penalized divergence first to get a decent estimate for $\log_2(k_{prod})$ first, since we observed that otherwise the fits sometimes had extreme values. The resulting parameters were used as starting values for the unpenalized fits. We used BIC to compare models with different numbers of parameters: $BIC = k \cdot \ln(n) - 2 \cdot SLL$, where k is the number of parameters, n is the number of samples, and SLL the sum of log-likelihood. k is 0 for the naive model, 3 for Eq 3.1 and 4 for Eq 3.2,3.3. n=8, the number of time points. The error of the fits was assumed to be normally distributed in order to calculate the LSL. When comparing models, the model with the lower BIC was considered superior.

3.4.13 MOFA analysis

We performed MOFA analysis to identify factors that drive translation. In contrast to what MOFA is intended for we applied it on the PTR ratios, to only look at the shifts out of steady state and not the actual values. Since we did not have matching protein-rna-miR samples we averaged ratios and miR to each time point. We filtered miRs for a minimum raw count of 10 and obtained 976 miRs. We then selected the top 1200 genes with the highest PTR ratio CV to obtain a similar number as recommended by the authors. MOFA was run in R with the default settings (MOFA2, v1.0).

3.4.14 GO term enrichment

We performed GO term enrichment analysis on the genes with extreme weights in the MOFA analysis. The top 120 positive and negative genes of each factor were used as ordered input for the topGO package (v2.38.1) We used a minimum term size of 20, the "ks" statistic and the "elim" algorithm. The elim algorithm takes the neighboring GO terms into account when calculating p-values. As a result, these p-values are not independent and therefore the authors do not recommend correcting for multiple hypotheses [63].

3.5 Supplementary information

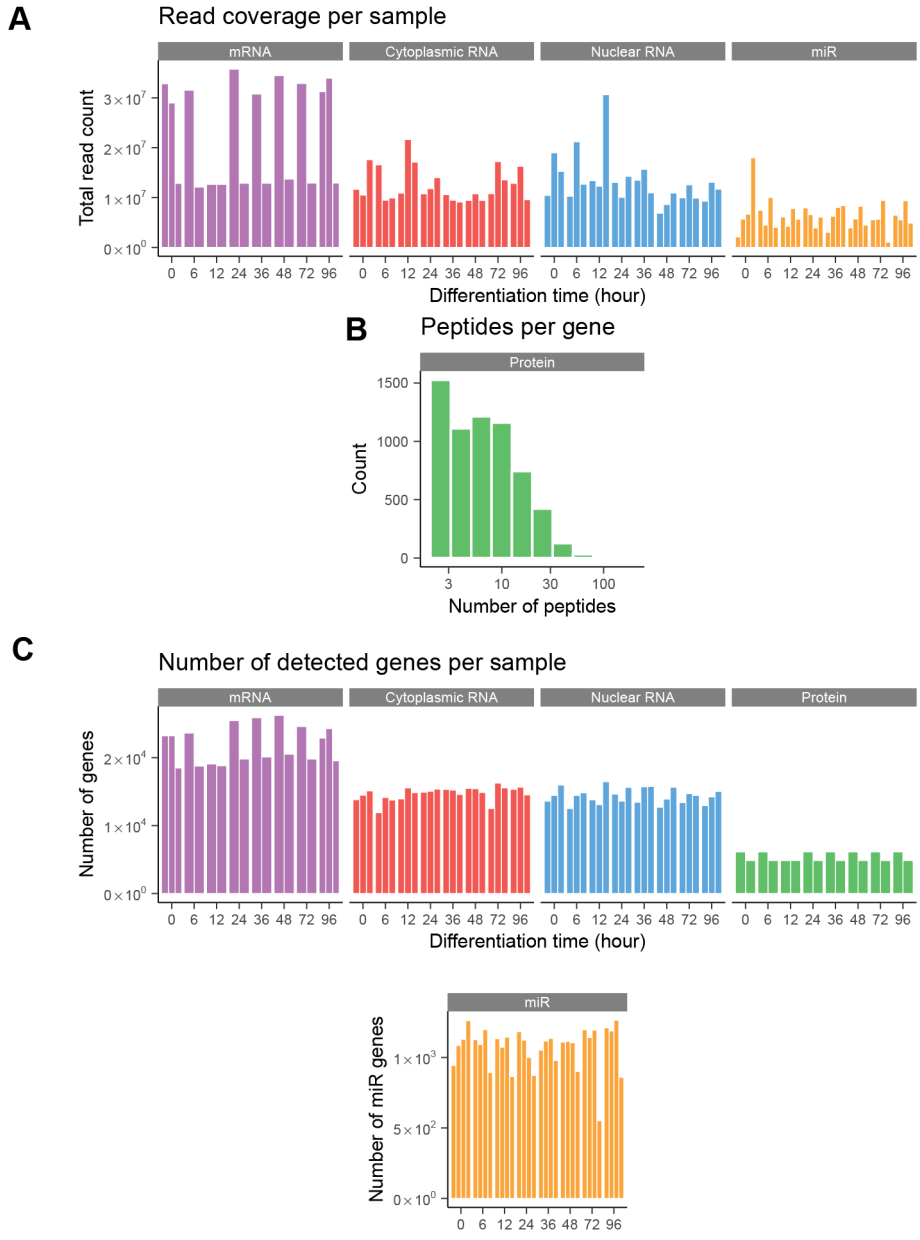


Figure S1. Quality control of full, cytoplasmic, nuclear RNA-seq, miRNA sequencing and proteomics. **(Continued on next page)**

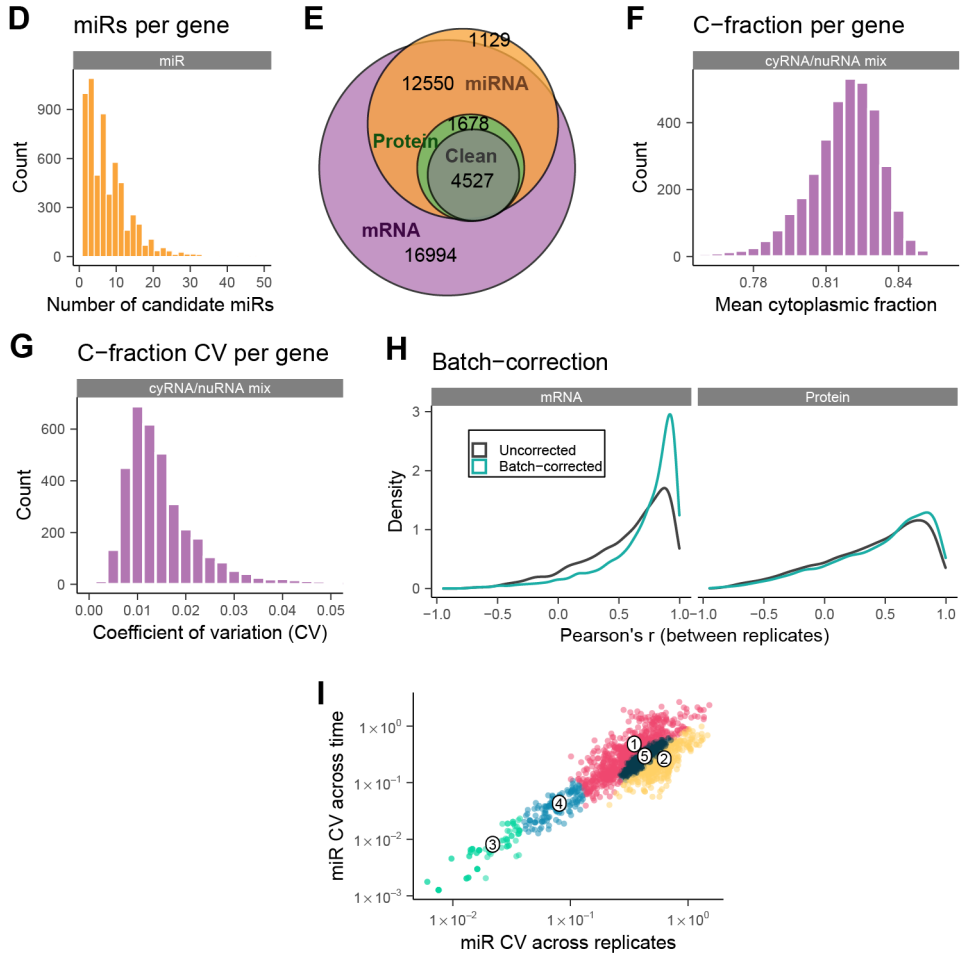


Figure S1. Quality control of full, cytoplasmic, nuclear RNA-seq, miRNA sequencing and proteomics. (Continued from last page) (A) Total number of reads for all sequencing samples. (B) Distribution of the number of peptides used for the quantification of each protein. (C) Number of detected genes or miRs in each sample. Individual replicates are plotted as separate bars in (A, C). (D) Distribution of miR-gene interactions per gene. (E) Euler diagram of all gene sets. The "miRNA" set indicates genes with predicted miR interaction and the "clean" set is a subset of genes without missing values in either RNA or protein. 53 genes are in the set: RNA&Protein&Clean (no miR-gene interactions), 13 genes are in the set: RNA&Protein (no miR-gene interactions, and some genes have missing values). (F) Distribution of the mean cytoplasmic fraction (C-fraction) per gene. (G) Coefficient of variation of C-fraction per gene. (H) Distribution of RNA-RNA and protein-protein Pearson's r with and without batch correction. (I) Gaussian mixture model based clustering of miRs to select a cluster with low noise and high variance (cluster 1), see Materials and methods.

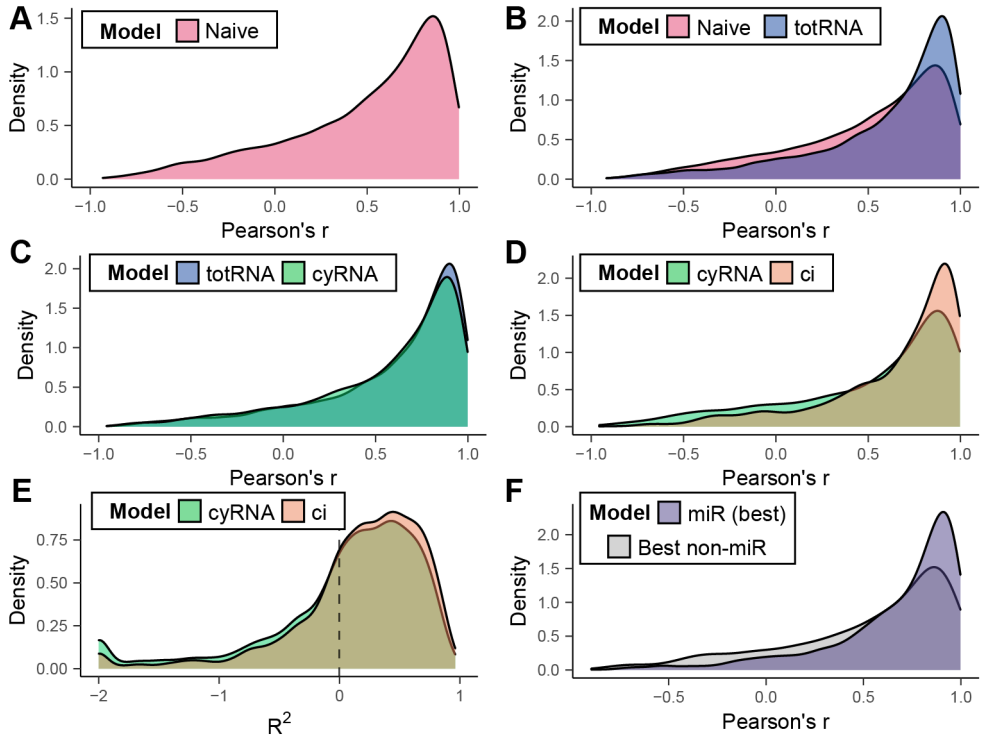
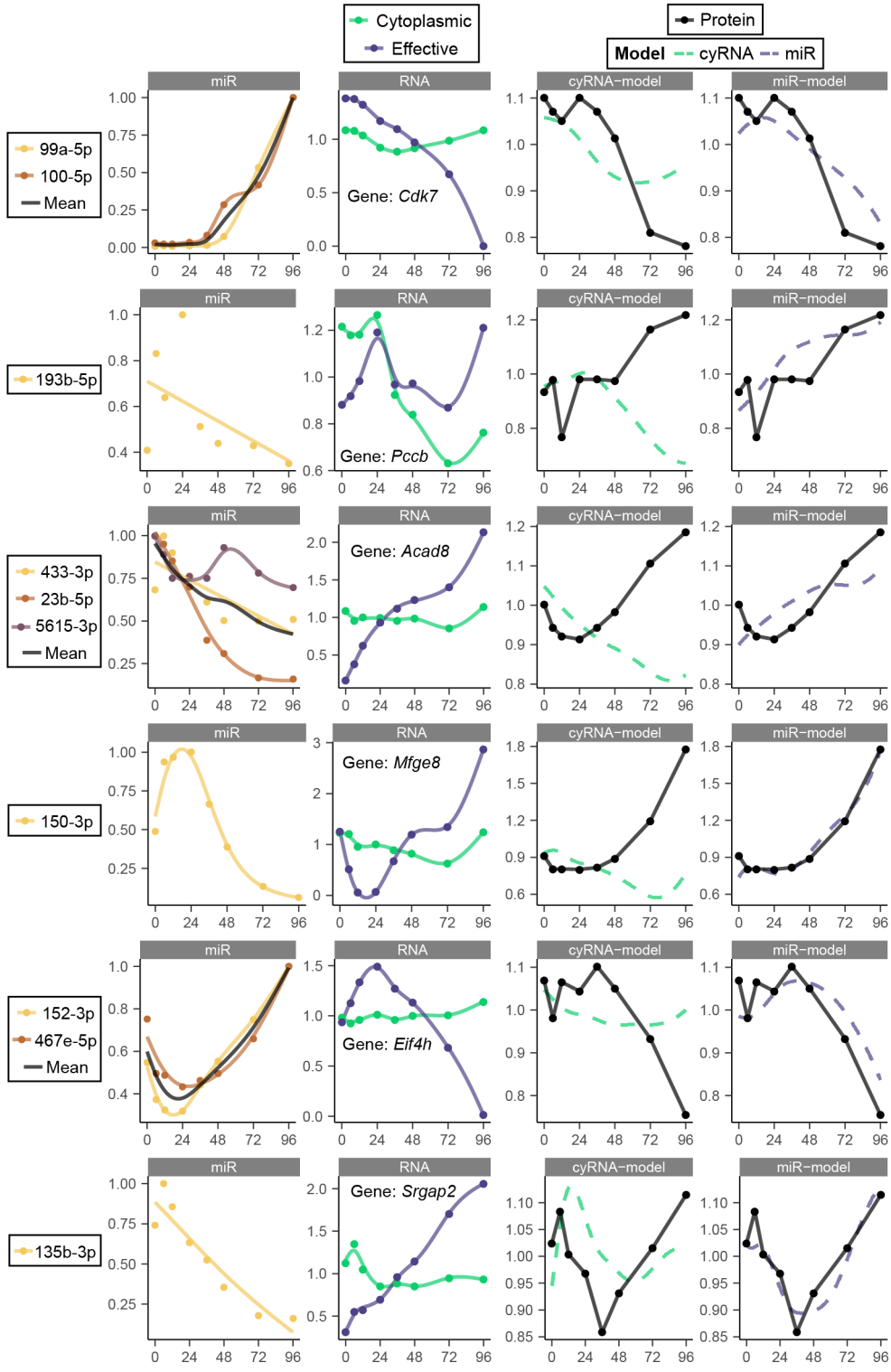


Figure S2. Model performance comparison with Pearson's r. (A-D) Pearson's r distribution of various protein models. Corresponding R^2 distributions are shown in Fig 1. (E) R^2 distribution of cyRNA and ci model for all genes. The R^2 distribution of the subset of genes that are best fit by the ci model is shown in Fig 1I. (F) Pearson's r distribution of the miR model and the next best model (either naive, totRNA, cyRNA or ci). Only genes that are best fit by the miR model are shown. Corresponding R^2 distributions are plotted in Fig 2D.

Figure S3 *(following page).* Six candidate miR-gene interactions. Example fit of the miR model for genes *Cdk7*, *Pccb*, *Acad8*, *Mfge8*, *Eif4h* and *Srgap2* (rows). First column: expression of the assigned miRs of a single cluster. Colored lines are individual smoothing spline fits. Second column: Cytoplasmic RNA expression and the effective RNA concentration available for translation (see Materials and methods). Solid lines represent smoothing splines. Third/fourth column: cyRNA and miR model fits.


Cytoplasmic
Effective
Protein
Model cyRNA miR
miR
RNA
cyRNA−model
miR−model
99a-5p
100-5p
Mean
Gene: Cdk7
193b-5p
Gene: Pccb
433-3p
23b-5p
5615-3p
Mean
Gene: Acad8
150-3p
Gene: Mfge8
152-3p
467e-5p
Mean
Gene: Eif4h
135b-3p
Gene: Srgap2

Acronyms

BIC Bayesian information criterion

CV Coefficient of variation

DF degrees of freedom

EMT epithelial-mesenchymal transition

ERLIC electrostatic repulsion-hydrophilic interaction chromatography

ESC emrbyonic stem cell

GO gene ontology

HF high field

MET mesenchymal-epithelial transition

miR micro-RNA

MOFA multi-omics factor analysis

MS/MS tandem mass spectrometry

nuRNA nuclear RNA

PTR protein to mRNA ratio

RA retinoic acid

TMT tandem mass tag

3.6 References

[1] Patrick van den Berg et al. “Kinetic modeling of multi-omics data reveals microRNA-mediated translational regulation in stem cell differentiation”. In: *Unpublished* (2020).

[2] F Soldner et al. “Medicine. iPSC disease modeling”. In: *Science* 338.6111 (2012), pp. 1155–1156. DOI: 10.1126/science.1227682.

[3] Stefan Semrau et al. *Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells*. Tech. rep. 1. 2016, pp. 1–16. DOI: 10.1038/s41467-017-01076-4.

[4] Kyle M Loh et al. “Mapping the Pairwise Choices Leading from Pluripotency to Human Bone, Heart, and Other Mesoderm Cell Types”. In: *CELL* 166.2 (2016), pp. 451–467. DOI: 10.1016/j.cell.2016.06.011.

[5] Allon M Klein et al. “Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells”. In: *CELL* 161.5 (2015), pp. 1187–1201. DOI: 10.1016/j.cell.2015.04.044.

[6] Anna S E Cuomo et al. “Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression”. In: *Nature Communications* 11.1 (2020), pp. 1–14. DOI: 10.1038/s41467-020-14457-z.

[7] Pavithra Kumar et al. “Understanding development and stem cells using single cell-based analyses of gene expression”. In: *Development (Cambridge, England)* 144.1 (2017), pp. 17–32. DOI: 10.1242/dev.133058.

[8] Christine Vogel et al. “Insights into the regulation of protein abundance from proteomic and transcriptomic analyses”. In: *Nature Publishing Group* 13.4 (2012), pp. 227–232. DOI: 10.1038/nrg3185.

[9] C Vogel et al. “Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line”. In: *Mol Syst Biol* 6.1 (2010), p. 400. DOI: 10.1038/msb.2010.59.

[10] Konstantine Tchourine et al. “One third of dynamic protein expression profiles can be predicted by a simple rate equation”. In: *Molecular BioSystems* 10.11 (2014), pp. 2850–2862. DOI: 10.1039/C4MB00358F.

[11] L Peshkin et al. “On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development”. In: *Developmental Cell* 35.3 (), pp. 383–394. DOI: 10.1016/j.devcel.2015.10.010.

[12] Kolja Becker et al. “Quantifying post-transcriptional regulation in the development of Drosophila melanogaster”. In: *Nature Communications* 9.1 (2018), p. 4970. DOI: 10.1038/s41467-018-07455-9.

[13] Kannanganattu V Prasanth et al. “Regulating Gene Expression through RNA Nuclear Retention”. In: *CELL* 123.2 (2005), pp. 249–263. DOI: 10.1016/j.cell.2005.08.033.

[14] Carole Iampietro et al. “Developmentally Regulated Elimination of Damaged Nuclei Involves a Chk2-Dependent Mechanism of mRNA Nuclear Retention”. In: *Developmental Cell* 29.4 (2014), pp. 468–481. DOI: 10.1016/j.devcel.2014.03.025.

[15] Antoine Graindorge et al. “Sex-lethal promotes nuclear retention of msl2 mRNA via interactions with the STAR protein HOW.” In: *Genes & Development* 27.12 (2013), pp. 1421–1433. DOI: 10.1101/gad.214999.113.

[16] Vikram Agarwal et al. “Predicting effective microRNA target sites in mammalian mRNAs”. In: *eLife* 4 (2015). DOI: 10.7554/eLife.05005.

[17] Niketa A Patel et al. “Long noncoding RNA MALAT1 in exosomes drives regenerative function and modulates inflammation-linked networks following traumatic brain injury”. In: *Journal of Neuroinflammation* 15.1 (2018), pp. 1–23. DOI: 10.1186/s12974-018-1240-3.

[18] Jian Wang et al. “Cardiomyocyte overexpression of miR-27b induces cardiac hypertrophy and dysfunction in mice”. In: *Cell Research* 22.3 (2012), pp. 516–527. DOI: 10.1038/cr.2011.132.

[19] Ziping Zhang et al. “miR-27 promotes human gastric cancer cell metastasis by inducing epithelial-to-mesenchymal transition”. In: *Cancer Genetics* 204.9 (2011), pp. 486–491. DOI: 10.1016/j.cancergen.2011.07.004.

[20] J Chai et al. “miR-23b-3p regulates differentiation of osteoclasts by targeting PTEN via the PI3k/AKT pathway”. In: *Archives of Medical Science* (2019). DOI: 10.5114/aoms.2019.87520.

[21] Jing Zhang et al. “MiR-23b-3p induces the proliferation and metastasis of esophageal squamous cell carcinomas cells through the inhibition of EBF3”. In: *Acta Biochimica et Biophysica Sinica* 50.6 (2018), pp. 605–614. DOI: 10.1093/abbs/gmy049.

[22] Roshan M Kumar et al. “Deconstructing transcriptional heterogeneity in pluripotent stem cells”. In: *Nature* 516.7529 (2014), pp. 56–61. DOI: 10.1038/nature13920.

[23] Haihai Liang et al. “miR-26a suppresses EMT by disrupting the Lin28B/let-7d axis: potential cross-talks among miRNAs in IPF”. In: *Journal of Molecular Medicine* 94.6 (2016), pp. 655–665. DOI: 10.1007/s00109-016-1381-8.

[24] Fengyan Yu et al. “MicroRNA 34c gene down-regulation via DNA methylation promotes self-renewal and epithelial-mesenchymal transition in breast tumor-initiating cells.” In: *Journal of Biological Chemistry* 287.1 (2012), pp. 465–473. DOI: 10.1074/jbc.M111.280768.

[25] Huarong Huang et al. “miR-10a contributes to retinoid acid-induced smooth muscle cell differentiation.” In: *Journal of Biological Chemistry* 285.13 (2010), pp. 9383–9389. DOI: 10.1074/jbc.M109.095612.

[26] Marion Coolen et al. “miR-9: a versatile regulator of neurogenesis”. In: *Frontiers in Cellular Neuroscience* 7 (2013), p. 220. DOI: 10.3389/fncel.2013.00220.

[27] Yu Ding et al. “Elevation of MiR-9–3p suppresses the epithelial-mesenchymal transition of nasopharyngeal carcinoma cells via down-regulating FN1, ITGB1 and ITGAV”. In: *Cancer Biology & Therapy* 18.6 (2017), pp. 414–424. DOI: 10.1080/15384047.2017.1323585.

[28] Qimin Wang et al. “Sevoflurane represses the self-renewal ability by regulating miR-7a,7b/Klf4 signalling pathway in mouse embryonic stem cells”. In: *Cell Proliferation* 49.5 (2016), pp. 609–617. DOI: 10.1111/cpr.12283.

[29] Hongyi Zhang et al. “MiR-7, Inhibited Indirectly by LincRNA HOTAIR, Directly Inhibits SETDB1 and Reverses the EMT of Breast Cancer Stem Cells by Downregulating the STAT3 Pathway”. In: *STEM CELLS* 32.11 (2014), pp. 2858–2868. DOI: 10.1002/stem.1795.

[30] Ui Jeong Yun et al. “miR-195a Inhibits Adipocyte Differentiation by Targeting the Preadipogenic Determinator Zfp423”. In: *Journal of Cellular Biochemistry* 116.11 (2015), pp. 2589–2597. DOI: 10.1002/jcb.25204.

[31] Chunhui Liu et al. “miR-195 Inhibits EMT by Targeting FGF2 in Prostate Cancer Cells”. In: *PLOS ONE* 10.12 (2015), e0144073. DOI: 10.1371/journal.pone.0144073.

[32] Xiaobin Lin et al. “miR-195-5p/NOTCH2-mediated EMT modulates IL-4 secretion in colorectal cancer to affect M2-like TAM polarization”. In: *Journal of Hematology & Oncology* 12.1 (2019), pp. 1–14. DOI: 10.1186/s13045-019-0708-7.

[33] Aihua Xu et al. “Inhibiting effect of microRNA-187-3p on osteogenic differentiation of osteoblast precursor cells by suppressing cannabinoid receptor type 2”. In: *Differentiation* 109 (2019), pp. 9–15. DOI: 10.1016/j.diff.2019.07.002.

[34] Changwei Dou et al. “miR-187-3p inhibits the metastasis and epithelial–mesenchymal transition of hepatocellular carcinoma by targeting S100A4”. In: *Cancer Letters* 381.2 (2016), pp. 380–390. DOI: 10.1016/j.canlet.2016.08.011.

[35] Srinivas R Viswanathan et al. “microRNA Expression during Trophectoderm Specification”. In: *PLOS ONE* 4.7 (2009), e6143. DOI: 10.1371/journal.pone.0006143.

[36] Jui-Tung Liu et al. “Arsenic Induces Members of the mmu-miR-466-669 Cluster Which Reduces NeuroD1 Expression”. In: *Toxicological Sciences* 162.1 (2017), pp. 64–78. DOI: 10.1093/toxsci/kfx241.

[37] Zhang Z et al. “MiR-770 inhibits tumorigenesis and EMT by targeting JMJD6 and regulating WNT/β-catenin pathway in non-small cell lung cancer”. In: *Life Sciences* 188 (2017), pp. 163–171. DOI: 10.1016/j.lfs.2017.09.002.

[38] S-H Hu et al. “miR-760 mediates chemoresistance through inhibition of epithelial mesenchymal transition in breast cancer cells.” In: *European review for medical and pharmacological sciences* 20.23 (2016), pp. 5002–5008.

[39] Zhi-Jiang He et al. “miR-1306–3p targets FBXL5 to promote metastasis of hepatocellular carcinoma through suppressing snail degradation”. In: *Biochemical and Biophysical Research Communications* 504.4 (2018), pp. 820–826. DOI: 10.1016/j.bbrc.2018.09.059.

[40] Lei Yan et al. “MiR-301b promotes the proliferation, mobility, and epithelial-to-mesenchymal transition of bladder cancer cells by targeting EGR1”. In: *Biochemistry and Cell Biology* 95.5 (2017), pp. 571–577. DOI: 10.1139/bcb-2016-0232.

[41] Ana Rita Lourenço et al. “SOX4: Joining the Master Regulators of Epithelial-to-Mesenchymal Transition?” In: *Trends in Cancer* 3.8 (2017), pp. 571–582. DOI: 10.1016/j.trecan.2017.06.002.

[42] Tong Zhang et al. “Downregulation of miR-542-3p promotes cancer metastasis through activating TGF-β/Smad signaling in hepatocellular carcinoma”. In: *OncoTargets and Therapy* 11 (2018), pp. 1929–1939. DOI: 10.2147/OTT.S154416.

[43] Sanghwa Kim et al. “miR-340-5p Suppresses Aggressiveness in Glioblastoma Multiforme by Targeting Bcl-w and Sox2”. In: *Molecular Therapy - Nucleic Acids* 17 (2019), pp. 245–255. DOI: 10.1016/j.omtn.2019.05.022.

[44] Ying Dong et al. “MiR-186 Inhibited Migration of NSCLC via Targeting cdc42 and Effecting EMT Process”. In: *Molecules and Cells* 40.3 (2017), pp. 195–201. DOI: 10.14348/molcells.2017.2291.

[45] Rui Wang et al. “The PDGF-D/miR-106a/Twist1 pathway orchestrates epithelial-mesenchymal transition in gemcitabine resistance hepatoma cells”. In: *Oncotarget* 6.9 (2015), pp. 7000–7010. DOI: 10.18632/oncotarget.3193.

[46] B Schwanhäusser et al. “Global quantification of mammalian gene expression control”. In: *Nature* 473.7347 (2011), pp. 337–342. DOI: 10.1038/nature10098.

[47] Mathias Wilhelm et al. “Mass-spectrometry-based draft of the human proteome”. In: *Nature* 509.7502 (2014), pp. 582–587. DOI: 10.1038/nature13319.

[48] Fredrik Edfors et al. “Gene-specific correlation of RNA and protein levels in human cells and tissues”. In: *Molecular Systems Biology* 12.10 (2016), pp. 883–2. DOI: 10.15252/msb.20167144.

[49] G Csárdi et al. “Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast”. In: *PLoS Genet* 11.5 (2015), e1005206. DOI: 10.1371/journal.pgen.1005206.

[50] Alexander Franks et al. “Post-transcriptional regulation across human tissues”. In: *PLoS Computational Biology* 13.5 (2017), e1005535. DOI: 10.1371/journal.pcbi.1005535.

[51] Dominic Grün et al. “Conservation of mRNA and Protein Expression during Development of C. elegans”. In: *CellReports* 6.3 (2014), pp. 565–577. DOI: 10.1016/j.celrep.2014.01.001.

[52] Anders R Kristensen et al. “Protein synthesis rate is the predominant regulator of protein expression during differentiation”. In: *Molecular Systems Biology* 9 (2013), p. 689. DOI: 10.1038/msb.2013.47.

[53] Rong Lu et al. “Systems-level dynamic analyses of fate change in murine embryonic stem cells”. In: *Nature* 462.7271 (2009), pp. 358–362. DOI: 10.1038/nature08575.

[54] Tomáš Gedeon et al. “Delayed Protein Synthesis Reduces the Correlation between mRNA and Protein Fluctuations”. In: *Biophysical Journal* 103.3 (2012), pp. 377–385. DOI: 10.1016/j.bpj.2012.06.025.

[55] B Munsky et al. “From analog to digital models of gene regulation”. In: *Phys. Biol* 12.4 (2015), p. 045004. DOI: 10.1088/1478-3975/12/4/045004.

[56] Nelly Rahkonen et al. “Mature Let-7 miRNAs fine tune expression of LIN28B in pluripotent human embryonic stem cells”. In: *Stem Cell Research* 17.3 (2016), pp. 498–503. DOI: 10.1016/j.scr.2016.09.025.

[57] Lin He Meng Amy Li. “microRNAs as novel regulators of stem cell pluripotency and somatic cell reprogramming”. In: *BioEssays : news and reviews in molecular, cellular and developmental biology* 34.8 (2012), pp. 670–680. DOI: 10.1002/bies.201200019.

[58] Zsuzsanna Lichner et al. “The miR-290-295 cluster promotes pluripotency maintenance by regulating cell cycle phase distribution in mouse embryonic stem cells”. In: *Differentiation* 81.1 (2011), pp. 11–24. DOI: 10.1016/j.diff.2010.08.002.

[59] Harsh Dweep et al. “miRWalk – Database: Prediction of possible miRNA binding sites by “walking” the genes of three genomes”. In: *Journal of Biomedical Informatics* 44.5 (2011), pp. 839–847. DOI: 10.1016/j.jbi.2011.05.002.

[60] S Griffiths-Jones et al. “miRBase: tools for microRNA genomics”. In: *Nucleic Acids Research* 36.Database (2007), pp. D154–D158. DOI: 10.1093/nar/gkm952.

[61] Qi-Long Ying et al. “The ground state of embryonic stem cell self-renewal”. In: *Nature* 453.7194 (2008), pp. 519–523. DOI: 10.1038/nature06968.

[62] Jacek R Wisniewski et al. “Universal sample preparation method for proteome analysis”. In: *Nature Methods* 6.5 (2009), pp. 359–362. DOI: 10.1038/nmeth.1322.

[63] Adrian Alexa et al. *topGO: Enrichment Analysis for Gene Ontology*. 2019.