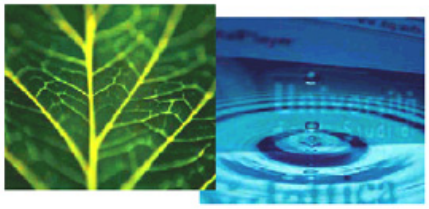
PhD Dissertation



International Doctorate School in Information and Communication Technologies

DISI - University of Trento

Computational and Experimental Detection of Uncoupling Between Transcriptome and Translatome Changes of Gene Expression

Toma Tebaldi

Advisor:

Prof. Alessandro Quattrone

University of Trento

CIBIO

March 2010

Abstract

Transcriptome analysis by total mRNA profiling provides a measurement of the degree of variation for the amount of each single mRNA species after a physiological or pathological transition of cell state. It has become a general notion that variations in protein levels do not necessarily correlate with variations in total mRNA levels, for the presence of post-transcriptional controls which influence the fate of cytoplasmic mRNAs and affect their translational fitness. Nevertheless, the extent of this phenomenon and the rules, if any, governing it are still generally unknown. To address this issue we took advantage of a number of studies performed using polysomal mRNA profiling in combination with classical total mRNA profiling in different mammalian and yeast systems. A normalization of the raw data coming from these datasets and a statistical meta-analysis aimed at maximizing uniformity in data processing have been performed. From the comparison of the results an extensive uncoupling between transcriptome and translatome variations of mRNA levels emerges, measured by a significant difference between steady state and polysomal fold changes induced by a cellular physiological or pathological transition. It seems clear that virtually the majority of significant changes in cytoplasmic mRNA steady-state levels are subjected to a further elaboration by a post-transcriptional decision program, leading either to a widespread buffering of the cytoplasmic changes which transfers only a small fraction of them to translation, either to the creation of new changes which cannot be detected at the tran-

scriptional level, yet capable of heavily influencing protein synthesis rates. An explanatory model characterized by a cytoplasmic mRNA storage compartments is proposed and the involvement of P-bodies and the miRNA pathway in post-transcriptional reprogramming of gene expression has been experimentally tested in the biological model of EGF induction, in order to explain how a change in translational fitness can counteract or magnify a parallel change in citoplasmic mRNA availability. To investigate the role of specific cellular mechanism in generating uncoupling between transcriptome and translatome changes, the experimental model has been altered through silencing of three key genes involved in post-transcriptional regulation pathways: 4E-T, Xrn1 and Dicer.

Keywords

[Translational controls, post-transcriptional regulation, polysomal profiling, microarray analysis,]

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The central dogma of molecular biology states that genetic information flows from nucleic acids to proteins [1]. In order to survive, living organisms have to regulate the expression of thousands of genes in response to multiple cellular needs and environmental stimuli. Expression control systems have to respond quickly and precisely to specific signals, and tune the level of expression of genes to regulate cell growth, adaptation to stress, homeostasis, and differentiation. In the past years scientific research on gene expression was mainly oriented towards decoding the molecular mechanisms of transcriptional control. This bias has both historical and technical reasons, since transcriptional control is the most basic step of gene expression and is simple to study with well-established experimental methods, but now the paradigm has changed and post-transcriptional regulations of mRNAs, including pre-mRNA splicing, maturation and quality control, mRNA transport to the cytoplasm [2] [3], localization in space and time [4] [5], editing [6], stability and degradation [7] [8], silencing and interference [9], circularization [10], translation initiation [11], nonsense-mediated mRNA decay [12]. All these processes acting on mRNA molecules are increasingly recognized as fundamental and influential steps in the flow of genetic information. Post-transcriptional regulation is dependent on the

activity of trans-acting factors, mainly RNA binding proteins (RBPs) [13], and non coding RNAs (ncRNAs) [14] which bind to cis consensus elements present mainly in the 3' and 5' untranslated regions of mRNAs, as presented schematically in figure 1.1

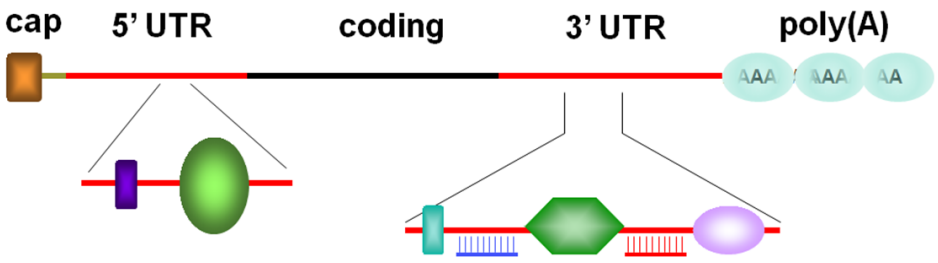


Figure 1.1: Linear structure of an eukaryotic mRNA, showing untranslated regions at both ends, with the main trans-acting factors (RNA binding proteins interacting mainly with 5'UTR and 3'UTR, microRNAs interacting mainly with 3'UTRs) involved in post-transcriptional controls of gene expression.

The polysome profiling technique involves separation of mRNA/ribosomal complexes by sucrose gradient centrifugation into inefficiently translated fractions (corresponding to monosomes or free mRNPs) and efficiently translated fractions (corresponding to polysomes) [15]. Microarray technology is then used to quantify the levels specific transcripts and detect which ones are redistributed between the different fractions in response to some stimulation, allowing recognition of translational up- or down-regulation at a gene-specific level. DNA microarray analysis can be used to simultaneously monitor transcriptome and translational changes in a cell. The analysis of a transcriptome through total RNA profiling provides only information on the template that is available for a cell to undergo translation processes under certain physiological conditions. Anyway, proteins are the real effectors of cell phenotype, and their levels and activities do not necessarily correlate with total mRNA levels, because

post-transcriptional controls act in the middle. In fact, the synthesis of individual protein species is regulated not only by transcript level, but by cis and trans elements that confer unique translational properties (a specific translational fitness) on individual mRNA molecules, and determine their fate: translation, degradation or silencing. The progressive discovery of how much post-transcriptional controls are pervasive and weighty has led to the conclusion that the explicit analysis of these mechanisms is determinant and unavoidable if we want to study biological systems without incurring in deviant simplifications [16]. Several works in the past ten years have compared transcriptome mRNA levels to the corresponding protein levels using high-throughput techniques, and they all have shown that the correlation level between the two measures is globally limited. It seems that the differential expression of mRNAs (in both directions, up or down) can capture and predict at most 60% of the corresponding variations of protein expression [17] [18]. This result is indeed limited to the number of proteins for which a direct comparison between high-throughput transcriptomic and proteomic measures are available. In light of these points, it would be valuable to have information on mRNA expression patterns with estimates of translation efficiencies of individual transcripts. Polysomal mRNA profiling should be more informative in this direction, revealing every mRNA whose translation is uncoupled from its transcription. In the last few years several works have been published in many scientific journals on the comparison between total mRNA profiling, based on the extraction and microarray analysis of all the mRNA contained in the studied cells, and polysomal profiling, based on the extraction and analysis of polysomal mRNA, i.e. the fraction of mRNA which is actively translated at the moment of the extraction. For the analysis presented in chapter 3 all the works whose raw data were at disposal have been considered. Different datasets have been classified according to the different phenom-

ena which are expected to generate a phenotypic variation in the studied cells. Most of these works compare total RNA data with polysomal RNA data, a minority of them compare polysomal RNA data with subpolysomal RNA data, derived from the analysis of poorly translated RNA fractions on the sucrose gradient. Post-transcriptional regulation of gene expression is much more intricate than previously thought, and elucidating the basic mechanisms of post-transcriptional control will be essential to gain a full understanding of how gene expression is regulated at different levels, of the interplay between these mechanisms, and of the extensive involvement of post-transcriptional dysfunction in numerous genetic disorders and cancer.

1.1 Proposed Solution

Expression levels for total and polysomal RNAs were calculated from raw data and normalized using the Robust Multichip Average algorithm (RMA) implemented in the Affy package of Bioconductor [19]. Significant differentially expressed genes in the total and polysomal RNA fractions were identified using a statistical technique based on rank products and implemented in the RankProd package of Bioconductor [20]. In comparison with other techniques for detection of differentially expressed genes, this one has been proven to be particularly suited to meta-analysis of multiple microarray experiments based on different platforms [21]. Populations of differentially expressed genes detected from transcriptome profiling and translatome profiling were compared and overlapped in order to calculate a categorical measure of uncoupling based on gene identities.

In order to model uncoupling as a quantitative measure, principal component analysis was performed on total RNA and polysomal RNA fold changes values: the underlying assumption is that the first principal component pins down the ideal line on which polysomal and total fold changes

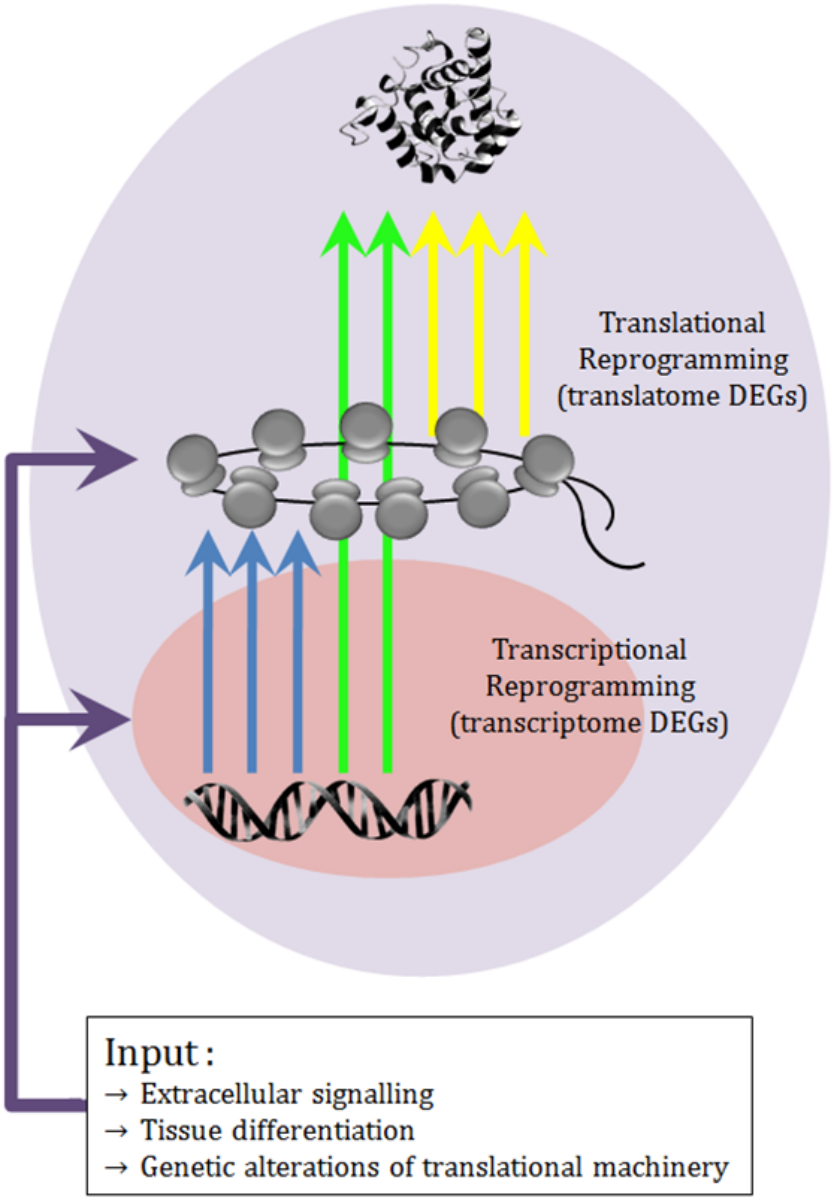


Figure 1.2: Diagram showing how external and internal perturbations reprogram gene expression regulation at a double level, transcriptional and translational, originating three types of differentially expressed genes (DEGs): those detectable only by transcriptome profiling, those detectable only by translatome profiling, those detected by both analyses.

are perfectly related. Uncoupling between transcriptome and translatome changes can be measured for all genes as their distance from the first prin-

cipal component. Since the collected datasets originated from different species, orthologous gene families among Homo sapiens, Mus musculus and Rattus norvegicus have been created using ENSEMBL orthology relations, in order to assess the recurrent presence of homologous genes in the populations of coupled or uncoupled genes.

1.2 Innovative Aspects

Computational meta-analysis: Comprehensive analysis of all published and high-quality microarray comparisons between transcriptome and translatome profiling data. Calculation of categorical uncoupling as the overlapping degree between lists of transcriptome and translatome differentially expressed genes. Calculation of quantitative uncoupling applying principal component analysis to transcriptome and translatome fold changes and considering the second principal component as the uncoupling dimension. Calculation of an ontological uncoupling, which measures the amount of alternative biological conclusions which can be drawn from the ontological analysis of these lists.

Experimental validation and alteration of uncoupling: Validation of uncoupling in a model of EGF induction and alteration of the model through silencing of key genes involved in post-transcriptional regulation pathways: 4E-T, Xrn1, Dicer.

Bayesian inference: Use of a bayesian inference approach to predict relationships between RNA binding proteins and target mRNAs based on changes in their translation efficiencies. Implementation and successful testing with synthetic data.

Untranslated sequences analysis: identification of hyper-conserved sequences in 5'UTR and 3'UTR of vertebrates, based on both sequence

identity and evolutionary coverage.

Ontology: Conceptualization of the "post-transcriptional regulation of gene expression" domain through design and implementation in OWL of the USER ontology.

Evolutionary approach: Sequence similarity based identification of a superfamily of RNA binding proteins with multiple RRM domains evolving from the PABP family and differentiating in binding specificities and performed molecular tasks.

1.3 Structure of the Thesis

Chapter 1: Introduction where the biological context, post-transcriptional regulation of gene expression, is briefly presented. The problem, lack of bioinformatic resources and information targeted to the discovery of post-transcriptional networks, is introduced. The solution proposed by this thesis and its innovative aspects are also outlined.

Chapter 2: State of the Art Where the biological information about post-transcriptional regulation and the currently available bioinformatic resources are described.

Chapter 3: Metanalysis results where transcriptome and translatome profiling data collected from literature are analyzed following the same pipeline, leading to a categorical and a quantitative measure of uncoupling.

Chapter 4: Experimental validation where uncoupling is verified in an experimental model of EGF induction in HeLa cells. The biological model is then altered through silencing of post-transcriptional key genes, in order to assess the involvement of p-bodies and the miRNA

pathway in the post-transcriptional reprogramming of gene expression.

Chapter 5: Conclusions where the results are summarized and future perspectives are outlined

Appendix A: Bayesian inference: where a bayesian inference model is designed to infer relationships between RNA binding proteins and mRNAs from translatome profiling experiments.

Appendix B: Ontology: where the USER ontology (Untranslated Sequence Elements for Regulation) is described.

Chapter 2

State of the Art

2.1 Gene expression

Gene expression is the process by which genome sequence is turned into proteins enabling our life. This process is divided into two main steps: transcription and translation. Transcription, also called RNA synthesis, is the step by which portions of DNA sequence are copied into molecules of messenger-RNA(mRNA). RNA polymerase and transcription factors are the main actors leading this process. The second step, translation, occurs on ribosomes, macromolecular complexes composed by proteins and RNA. At that moment, the mRNA sequence is read by the ribosome, codon by codon, in order to produce polypeptide chains. When translation is completed, the mature protein is released by the ribosome. Each mRNA includes two noncoding regions, called 5' and 3' UTR (Untranslated Region) at the beginning and at the end of the transcribed sequence. Translation of mRNAs is regulated also by means of these regions, thus making them particularly important. These regions both contain regulatory sequences, making them cis-regulatory elements(they contain sequences regulating the expression of the gene on the same strand), and are target of trans-factors(proteins used in the regulation of another target gene) like RNA-binding proteins.

2.1.1 mRNA maturation

Next to transcription, some further processing allows eukaryotic cells to produce mature and functional mRNAs from the newly transcribed RNA molecules(called pre-mRNA). First of all, the *5' cap* is added at the beginning of the transcript to avoid premature degradation during export from the nucleus to the cytoplasm. A similar process, driven by poly(A) polymerases and helped by PABPs(Poly-A Binding Proteins), adds a string of 100-250 adenine residues to the 3' end of the transcript; this structure, called *3' poly(A)-tail*, avoids the premature degradation of the transcript. Next, the *splicing* process deletes non coding-regions, called *introns*, from the RNA and joins the remaining regions, called *exons*, into a single sequence. An important process, called *alternative splicing*, allows to produce different mature mRNA transcripts by selecting different combinations of exons from the same pre-mRNA. Different proteins can be produced in this way from a single gene. Once all these processes are terminated, the structure of the produced functional mRNA, as illustrated in figure 1.1 starts with the 5' cap at the beginning of the transcript; then comes the 5'UTR, the coding sequence, and the 3'UTR; the mRNA is eventually closed by the 3' poly(A)-tail. Usually, the 3' UTR is much longer then the 5' UTR; the mean length of human UTRs is around 500 bases for 3' and 150 for 5'.

2.1.2 RNA export to the cytoplasm

In eukaryotes, translation of mRNA into functional proteins takes place in the cytoplasm, while transcription is a nuclear process. There is thus the need of transporting the mature mRNA outside of the nucleus to allow its translation: this can be done via structures nuclear pores that, localized on the nucleus membrane allow the export of RNA molecules. Nuclear

pores are composed by more than 100 proteins called nucleoporines, acting as selective pores allowing or prohibiting molecules passage. Diverse RNA-binding proteins(RBP) binds to the mRNA forming the mRNP complex; transport of this complex is further facilitated by mRNA-export units which interacts with specific proteins to allow transfer of the molecule to the cytoplasm. During this process, the mature mRNA is protected both by its cap and by the bound RBPs [2].

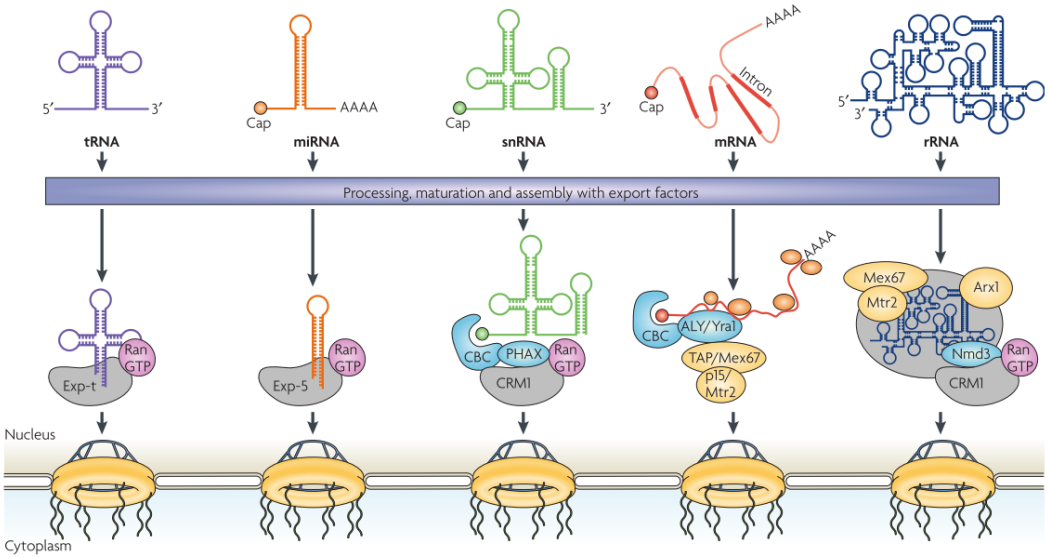


Figure 2.1: Figure taken from [2] showing the different RNA export routes for several RNA classes.

2.1.3 mRNA stability regulation

One important level of post-transcriptional regulation involves mRNA stability: proteins binding to control elements usually located in the 3'UTR can alter the decay rate of a transcript, thus favoring its quick degradation or slowing it down [8]. Elements in the 3'UTR that were observed to be associated with high decay rates are the *AU-rich elements* (*AREs*), which are regions composed by a great majority of A and U [22]; an example of trans-acting factor is the *PUF*(*Pumilo and feminizing mutation-3 mRNA-*

binding factor) RNA-binding protein family, which binds to the 3'UTR of the target transcripts and shorten their poly(A)-tail, thus reducing the stability of the mRNA. Shortening of the poly(A)-tail is indeed a very common way of targeting mRNAs for degradation: once the tail is too short for PABP binding, even the stabilization of 5' cap and initiation factors can no longer occur, thus favoring 5' decapping and consequent mRNA degradation by exonucleases.

Exosomes are important actors in mRNA degradation, being multiprotein complexes capable of degrading various kinds of RNA molecules. Instead of cleaving RNA molecules at a specific site, this complexes degrade RNA molecules by starting at the 3' end. Regulated at their turn by different proteins, exosomes are known to be involved in autoimmune diseases and cancer onset. Messenger RNAs are targeted to these complexes when they contain errors or as a part of their normal turnover; exosomes can also interact with RNA binding proteins interacting with AU-rich elements.

2.1.4 Transport between cytoplasmic granules

Polysomal RNA assumes a circularized conformation through interactions between poly(A)-binding protein 1 (PABP1) on 3'UTR and eukaryotic translation initiation factor 4G (eIF4G) on 5'UTR, which are stabilized by eIF3. In eukaryotic cells circularization is a necessary step bringing to the formetion of polyribosomes or polysomes: complexes of more ribosomes attached to the same mRNA molecule.Transformation of polysomes into linearized messenger ribonucleoproteins (mRNPs) seem to involve the transport to p-bodies, whereas circularized mRNPs are directed to stress granules. In the p-bodies pathway, the deadenylation complex CCR4NOT1 is recruited by destabilizing factors, such as tristetraprolin (TTP), or RNA-induced silencing complexes (RISCs), involving Argonaute proteins and microRNAs. Loss of circularization by loss of eIF3 or deadenylation-induced

loss of PABP1 produces a linear transcript. This linear mRNA recruits a decapping complex (which consists of decapping protein 1 (DCP1; DCP1A in humans), DCP2, enhancer of mRNA-decapping protein 3 (EDC3), RCK (also known as DDX6) and HEDLS) and a decapping activator complex (PAT1 bound to LSM17; PAT1 is not shown). Q/N-rich domains in LSM4 and EDC3 promote the aggregation of these mRNAs into PBs. In the 'circular' pathway (right), transiently stalled initiation complexes recruit TIA1 and TIAR (together shown as TIA) as elongating ribosomes run off the transcript, converting the polysome into a circular, adenylated mRNP. Aggregation of bound TIA1 and TIAR or G3BPUSP10 (G3BP is GTPase-activating protein SH3 domain-binding protein and USP10 is ubiquitin-specific processing protease 10) and/or modification of ribosomes with O-linked N-acetylglucosamine (GlcNAc) promote the assembly of these mRNAs into SGs. It is possible that mRNPs in PBs or SGs can be remodelled to nucleate the assembly of other types of RNA granules. Alternatively, selected mRNPs might move from one type of granule to another, thus creating transient tethers between different granules.[23]

2.1.5 Cap-dependent translation initiation

Translation initiation is the process of assembly of 80S ribosomes where the initiation codon is base-paired with the anticodon loop of initiator tRNA in the ribosomal P-site. It requires at least nine eukaryotic initiation factors (eIFs) and comprises two steps: the formation of 48S initiation complexes with established codonanticodon base-pairing in the P-site of the 40S ribosomal subunits, and the joining of 48S complexes with 60S subunits. On most mRNAs, 48S complexes form by a scanning mechanism, whereby a 43S preinitiation complex (comprising a 40S subunit, the eIF2GTPMet-tRNA^{Met} ternary complex, eIF1, eIF1A and probably eIF5) attaches to the capped 5' proximal region of mRNAs in a step that involves the unwinding

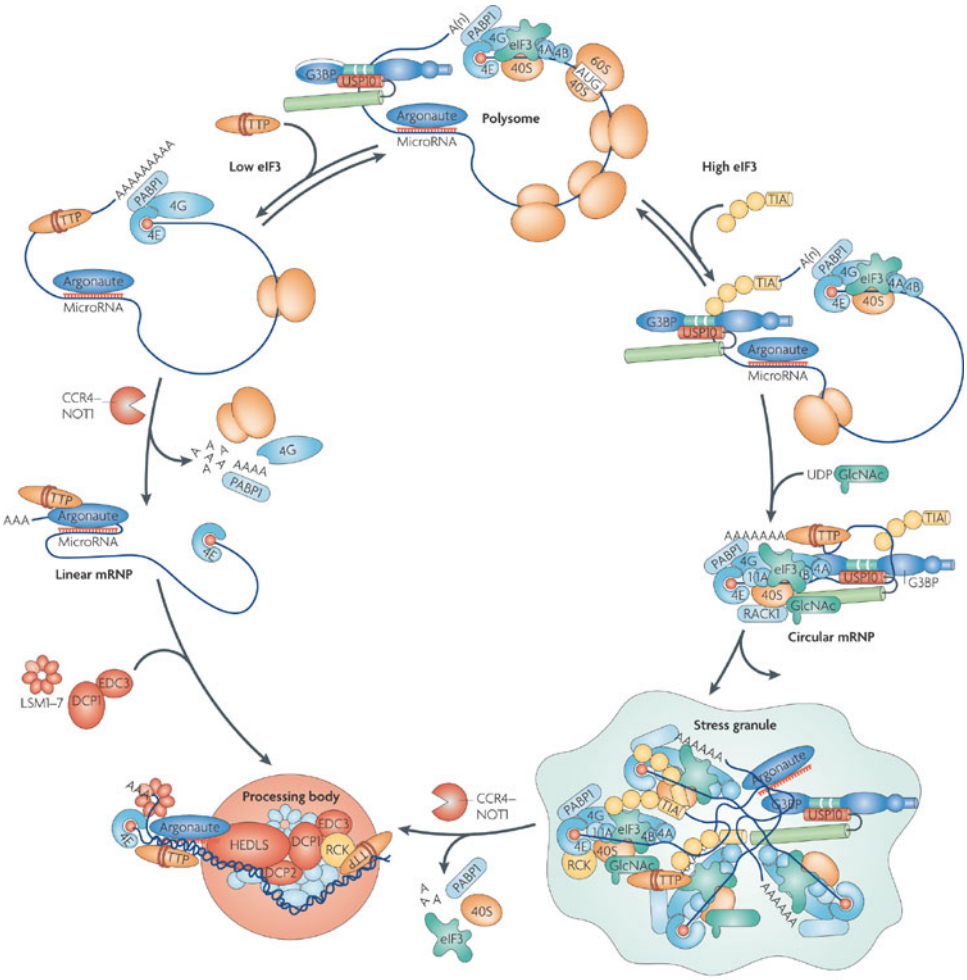


Figure 2.2: Taken from [23]. Molecular pathways connecting actively translating polysomes to distinct cytoplasmic storage and degradation granules: p-bodies and stress granules.

of the mRNAs 5' terminal secondary structure by eIF4A, eIF4B and eIF4F. The 43S complex then scans the 5' untranslated region (5' UTR) in the 5' to 3' direction to the initiation codon. After initiation codon recognition and 48S complex formation, eIF5 and eIF5B promote the hydrolysis of eIF2-bound GTP, the displacement of eIFs and the joining of a 60S subunit. Although most mRNAs use the scanning mechanism, initiation on a few mRNAs is mediated by internal ribosome entry sites.

Cap-dependent translation initiation entails the recruitment of the 40S

small ribosomal subunits (and associated factors) to the 5' end of the mRNAs. In this process, the mRNA 5'-cap structure, m^7GpppN (where N is any nucleotide), is recognized by eukaryotic initiation factor (eIF) 4E, one of the subunits of the eIF4F complex. The eIF4F complex also contains eIF4A, an ATP-dependent RNA helicase which is thought to unwind secondary structure present at the 5' end of the mRNA, and eIF4G, a large scaffolding protein that binds to eIF4E, eIF4A, PABP, and eIF3, consequently bridging the ribosome and the mRNA. eIF4E is the limiting factor in translation initiation under most circumstances and is an important effector of cellular proliferation, survival, and malignant transformation. The activity of eIF4E is regulated by a family of translational suppressors called the 4E-binding proteins (4E-BPs), which in mammals consists of three members: 4E-BP1, 4E-BP2, and 4E-BP3. 4E-BP1 and 4E-BP2 are expressed in most tissues, whereas 4E-BP3 exhibits a more restricted expression pattern. Binding of the 4E-BPs to eIF4E is controlled by the phosphorylation status of 4E-BPs. The hypophosphorylated forms of 4E-BP bind to eIF4E and prevent interaction of eIF4E with eIF4G, thus impairing cap-dependent translation. Conversely, in nutrient- or serum-stimulated cells, 4E-BPs become hyperphosphorylated, releasing eIF4E for interaction with eIF4G and assembly into the eIF4F complex, resulting in enhanced translation. The best-characterized 4E-BP is 4E-BP1, which contains six known proline-directed Ser/Thr phosphorylation sites, among which at least two sites are phosphorylated directly by mTOR (mammalian target of rapamycin). mTOR is a phylogenetically conserved Ser/Thr kinase that regulates cell growth and metabolism in response to diverse extracellular and intracellular cues. Growth factors and hormones (insulin/IGF), nutrients (amino acids/glucose), and high ATP/AMP ratio activate mTOR, resulting in hyperphosphorylation of 4E-BP1. Rapamycin, an inhibitor of mTOR, impairs the phosphorylation of 4E-BP1.

While eIF4E is predominantly cytoplasmic, in mammalian cells and in yeast, a significant fraction (12% 33% in mammalian) is localized to the nucleus at steady-state levels as determined by biochemical fractionation studies and immunofluorescence analysis using several antibodies. In the nucleus, eIF4E colocalizes with splicing factors in speckles. The nuclear import of eIF4E is mediated by 4E-T (eIF4E-transporter), which binds to eIF4E through a conserved binding motif shared with 4E-BPs and eIF4G, and simultaneously interacts with nuclear import receptors, importin $\alpha-\beta$ (Dostie et al. 2000a). While the role of eIF4E in the nucleus has not been as extensively studied as its cytoplasmic role, it is known to promote the nuclear export of a subset of mRNAs. How the steady-state pool of nuclear eIF4E is maintained and regulated is not clear.

2.2 Bioinformatic coverage of post-transcriptional controls

Actually there are many bioinformatic resources which cover particular facets of the post-transcriptional regulation field. Many of the available databases are manually curated by few people belonging to single laboratories: this can lead to several negative consequences:

- when curators move to other research groups or to other projects, they usually are not replaced and the databases, without the necessary updates, freeze and lose their usefulness
- databases are partially redundant in the data they display
- databases are isolated, since they may provide web links to each other but they lack integration at the data level

These are common problems in biological databases, partially due to the difficult task of organizing and collecting data which evolve rapidly follow-

ing changes in the experimental procedures and in the knowledge representations shared by the scientific community: this informational turnover is particularly marked in the dynamic and emerging post-transcriptional control field. The bioinformatic long term aim of the Laboratory of Translational Genomics at CiBio is to equip this fragmented and volcanic domain of poleis-like databases with a common platform through which the user could get a unified and meaningful view of post-transcriptional processes and make quantitative predictions on their combined effect on mRNA control and fate.

2.3 Proteome-Transcriptome comparisons

Several works in the past ten years have compared transcriptome mRNA levels to the corresponding protein levels using high-throughput techniques, and they all have shown that the correlation level between the two measures is globally limited. It seems that the differential expression of mRNAs (in both directions, up or down) can capture and predict at most 60% of the corresponding variations of protein expression [17] [18]. This result is indeed limited to the number of proteins for which a direct comparison between high-throughput transcriptomic and proteomic measures are available.

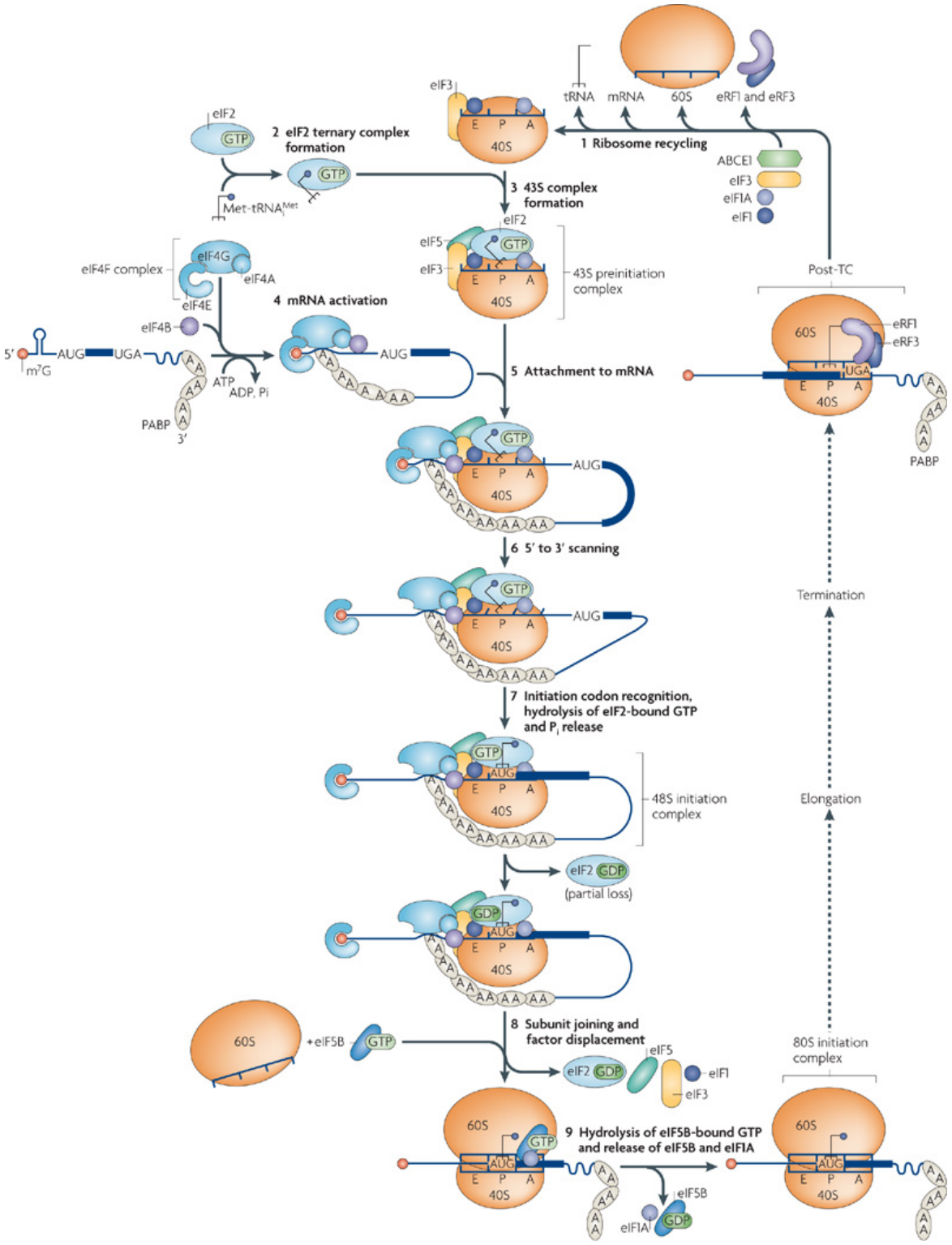


Figure 2.3: Picture taken from [11] showing the sequence of steps involved in the canonical pathway of cap-dependent translation initiation.

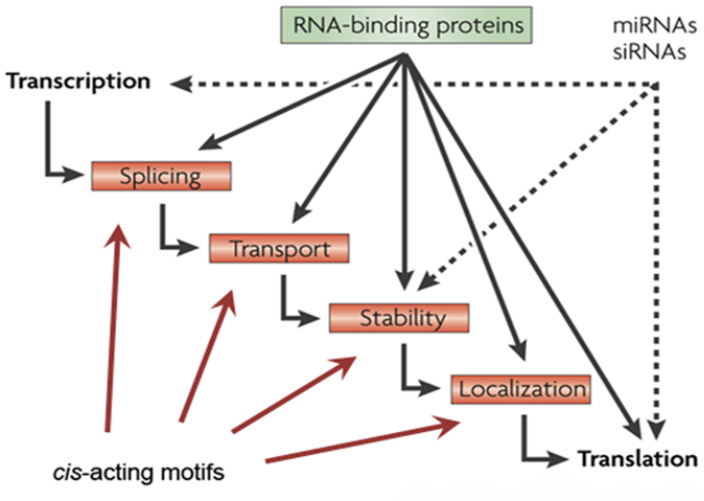


Figure 2.4: Multiple steps of regulation from transcription to translation occurring in eukaryotic cells and involving the binding of RBPs and small non-coding RNAs to cis-acting motifs on mRNAs. Adapted from [24].

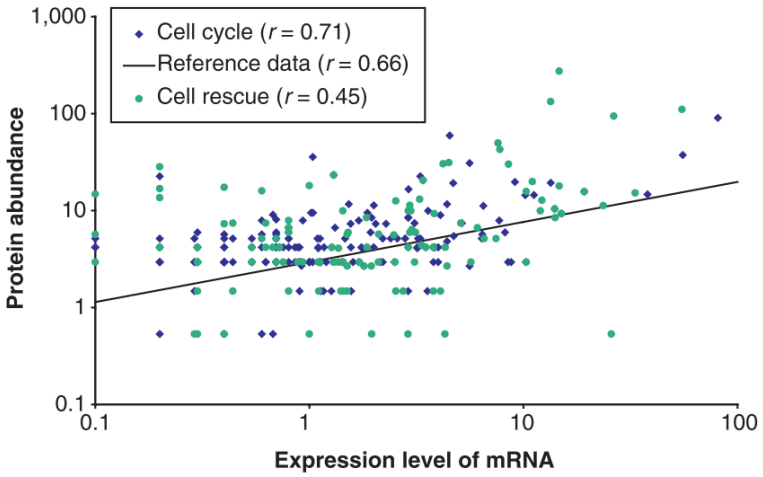


Figure 2.5: Comparison of mRNA expression and protein abundance taken from [18]. The mRNA axis is in copies per cell; the protein axis is in thousand copies per cell.

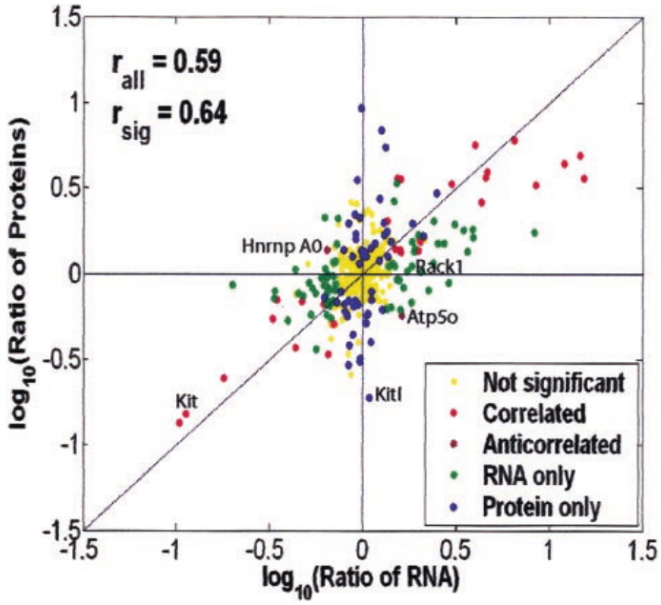


Figure 2.6: Scatter plot of mRNA versus cognate protein fold changes taken from [17]. The overall correlation coefficient for all the 425 genes in the analysis is 0.59.

Chapter 3

Computational detection of uncoupling: metanalysis

Genome-wide measurements of mRNA variations are widely proposed as truthful representations of changes in global protein abundance systematically neglecting the impact of post-transcriptional events. To estimate this impact we performed a normalized analysis of all technically comparable mammalian datasets for which coupled transcriptome and translatome [25] (mRNA extracted from cytoplasmic polysomal fractions) microarray profiles were available. We found that a general, profound uncoupling between transcriptome and translatome gene espression variations emerges as a rule. Moving to ontological analysis of differentially expressed genes, an approach based on semantic similarity between Gene Ontology terms has shown that only in the minority of the datasets the semantic distance between transcriptome and translatome representations of each of the compared datasets outlies the distribution of the same measure computed between disparate pairwise transcriptome representations. These results severely question the information completenesss of transcriptome profiles in directly representing cell phenotypes and in portraying cell activities.

3.1 Identification of DEGs

Datasets containing comparisons between polysomal and total RNA levels have been collected through extensive researches in literature and in the main microarray databases: GEO [1], ArrayExpress [2], Stanford Microarray Database [3]. Datasets without complete available raw data or without hybridization replicas for every experimental condition were excluded from the meta-analysis. The selected datasets are listed and described in Table 3.1. Though original data were organized in different experimental designs, in each one a two-group comparison (treated group vs. control group) between total and polysomal RNA levels was possible. Microarray data were analyzed using the R software environment for statistical computing (http://www.r-project.org/) and the Bioconductor library of biostatistical packages (http://www.bioconductor.org/). The expression levels for all arrays were calculated from raw data with the RMA (Robust Multi-chip Average) algorithm implemented in the Affy package of Bioconductor (http://www.bioconductor.org/packages/release/bioc/html/affy.html). Parallel normalization was carried out for total and polysomal RNA hybridizations. Probesets were associated to their corresponding Ensembl gene IDs. Ambiguous probesets, i.e. probesets associated to more than one Ensembl gene ID because of annotation imperfections or annotation changes in time, were filtered out from the analysis at this stage. Signals from multiple probesets associated to the same gene were averaged. To identify differentially expressed genes (DEGs) in either the total or the polysomal fractions, three different statistical approaches were addressed: Rank Product, SAM (Significance Analysis of Microarrays) and t-test. The Rank Product algorithm, implemented in the Bioconductor RankProd package

[1] http://www.ncbi.nlm.nih.gov/geo
[2] http://www.ebi.ac.uk/microarray-as/ae/
[3] http://smd.stanford.edu/

(http://www.bioconductor.org/packages/release/bioc/html/RankProd.html), uses a technique based on calculating rank products from replicate experiments. A permutation-based procedure is used to determine false discovery rate values, estimated by RankProd as "Percentages of False Positives" (pfp). A threshold of 0.2 on the pfp value was used to filter DEGs in either the total RNA or the polysomal RNA comparison.

DEGs were identified as belonging to transcriptomic or translatomic hybridizations.

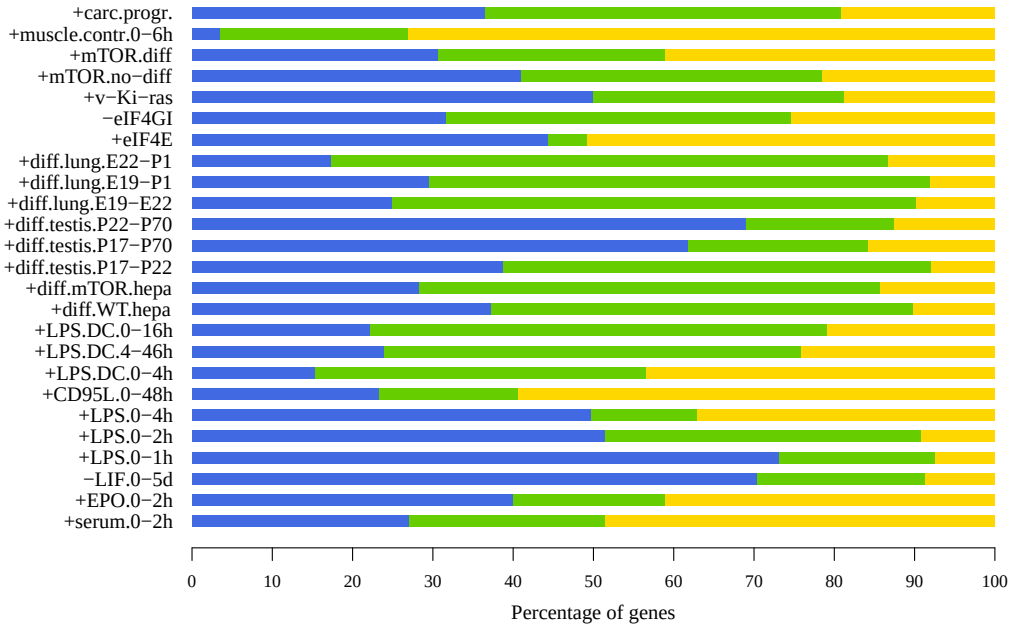


Figure 3.1: DEGs overlap between transcriptome and translatome levels. Each dataset in the bar graph is displayed to the left of its description in A. Genes are classified as DEGs by both transcriptome and translatome profilings (green), as DEGs only by transcriptome profiing (blue) or as DEGs only by translatome profiing (yellow). The bar length shows the relative proportion of DEGs in these three groups for each dataset.

Short ID	*Species*	Source	Description	Ref.
+serum.0-2h	*Mm*	MEFs	serum starvation release	[26]
+EPO.0-2h	*Mm*	I/11-R10 EPCs	erythroid EPO deprivation release	[27]
-LIF.0-5d	*Mm*	embrionic stem cells R1	stem cell differentiation through LIF removal	[28]
+LPS.0-1h	*Mm*	J774.1	macrophage LPS treatment	[29]
+LPS.0-2h	*Mm*	J774.1	macrophage LPS treatment	[29]
+LPS.0-4h	*Mm*	J774.1	macrophage LPS treatment	[29]
+CD95L.0-48h	*Mm*	neural stem cells	CD95L treatment	[30]
+LPS.DC.0-4h	*Hs*	dendritic cells	dendritic cell LPS treatment	[31]
+LPS.DC.4-46h	*Hs*	dendritic cells	dendritic cell LPS treatment	[31]
+LPS.DC.0-16h	*Hs*	dendritic cells	dendritic cell LPS treatment	[31]
+diff.WT.hepa	*Hs*	HepaRG	differentiation of WT hepatocytes	[32]
+diff.mTOR.hepa	*Hs*	HepaRG	differentiation of mTOR activated hepatocytes	[33]
+diff.testis.P17-P22	*Mm*	testis tissue	testis differentiation	[34]
+diff.testis.P17-P70	*Mm*	testis tissue	testis differentiation	[34]
+diff.testis.P22-P70	*Mm*	testis tissue	testis differentiation	[34]
+diff.lung.E19-E22	*Rn*	lung tissue	lung differentiation	[**?**]
+diff.lung.E19-P1	*Rn*	lung tissue	lung differentiation	[**?**]
+diff.lung.E22-P1	*Rn*	lung tissue	lung differentiation	[**?**]
+eIF4E	*Hs*	primary MECs	eIF4E overexpression	[35]
-eIF4GI	*Hs*	MCF10A	eIF4GI depletion	
+v-Ki-Ras	*Hs*	267B1/267B1-Ki-ras	v-Ki-ras transformation	[36]
+mTOR.no-diff	*Hs*	HepaRG	mTOR activation of proliferative hepatocytes	[33]
+mTOR.diff	*Hs*	HepaRG	mTOR activation of differentiated hepatocytes	[33]
+muscle.contr.0-6h	*Rn*	skeletal myocites	muscle subjected to high resistance contractions	[37]
+carc.progr.	*Hs*	SW480/SW620	carcinoma progression from primary cells to metastatic cells	[38]

Table 3.1: Datasets collection and classification on the basis of experimental perturbations applied. The red cluster indicates extracellular signaling events, the blue cluster is related to tissue differentiation, the green concerns genetic alterations of the translational machinery. Datasets are labelled by short names specifying perturbations and time points. Short names and color codes are used throughout the text to indicate the datasets belonging to each cluster. Species, biological sources, experimental settings and bibliographical references are summarized as well.

	DEGs numbers		
Short ID	**only tot**	**only poly**	**common**
+serum.0-2h	61	74	32
+EPO.0-2h	20	132	413
-LIF.0-5d	715	662	959
+LPS.0-1h	587	535	308
+LPS.0-2h	970	610	364
+LPS.0-4h	785	1063	628
+CD95L.0-48h	63	7	72
+LPS.DC.0-4h	477	1914	364
+LPS.DC.4-46h	761	1608	208
+LPS.DC.0-16h	805	2106	317
+diff.WT.hepa	513	138	93
+diff.mTOR.hepa	707	258	180
+diff.testis.P17-P22	523	721	106
+diff.testis.P17-P70	429	870	216
+diff.testis.P22-P70	339	478	92
+diff.lung.E19-E22	317	810	298
+diff.lung.E19-P1	461	998	464
+diff.lung.E22-P1	131	351	369
+eIF4E	178	132	452
-eIF4GI	626	168	467
+v-Ki-ras	276	212	49
+mTOR.no-diff	217	58	22
+mTOR.diff	138	41	17
+muscle.contr.0-6h	129	61	132
+carc.progr.	558	502	999

Table 3.2: Description of the parameters calculated from simulation 1

3.2 Ontological uncoupling

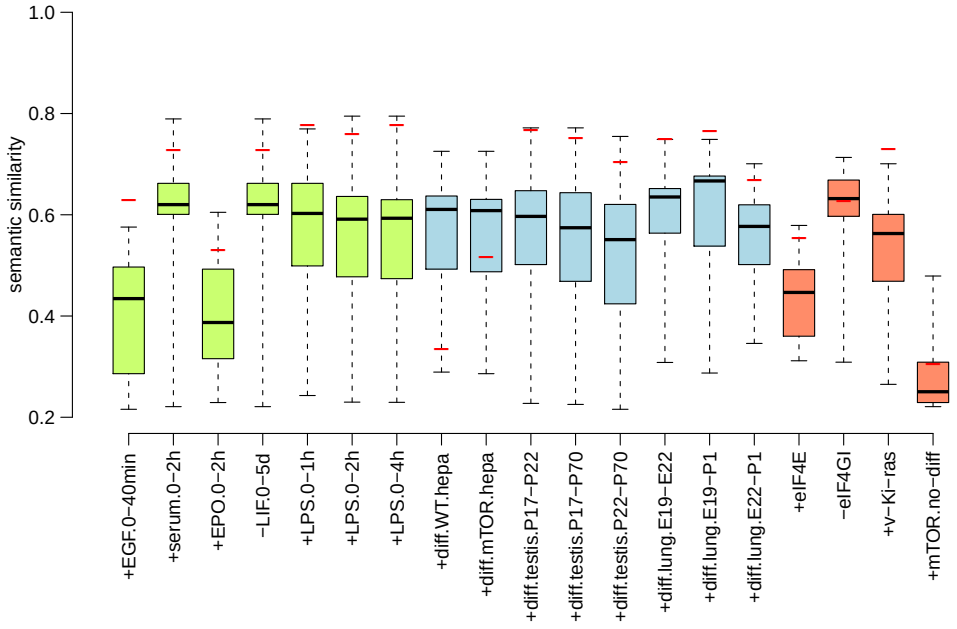


Figure 3.2: Semantic similarity between total and polysomal terms inside the same dataset is compared with the distribution of semantic similarities coming from pairwise comparisons between different datasets.

	GO terms statistics					
Short ID	**only tot**	**only poly**	**common**	**gain**	**loss**	**jaccard**
+serum.0-2h	29	35	22	0.31	0.17	0.52
+EPO.0-2h	13	77	13	0.83	0	0.17
-LIF.0-5d	70	159	62	0.58	0.05	0.37
+LPS.0-1h	119	84	82	0.02	0.31	0.68
+LPS.0-2h	172	101	93	0.04	0.44	0.52
+LPS.0-4h	203	133	115	0.08	0.40	0.52
+CD95L.0-48h	0	0	0	0	0	0
+LPS.DC.0-4h	160	144	125	0.11	0.20	0.70
+LPS.DC.4-46h	93	98	65	0.26	0.22	0.52
+LPS.DC.0-16h	189	177	139	0.17	0.22	0.61
+diff.WT.hepa	140	13	9	0.03	0.91	0.06
+diff.mTOR.hepa	151	19	19	0	0.87	0.13
+diff.testis.P17-P22	95	70	60	0.10	0.33	0.57
+diff.testis.P17-P70	114	70	56	0.11	0.45	0.44
+diff.testis.P22-P70	63	34	31	0.04	0.48	0.47
+diff.lung.E19-E22	146	109	88	0.13	0.35	0.53
+diff.lung.E19-P1	158	117	104	0.08	0.32	0.61
+diff.lung.E22-P1	74	38	34	0.05	0.51	0.44
+eIF4E	29	18	8	0.26	0.54	0.20
-eIF4GI	81	51	23	0.26	0.53	0.21
+v-Ki-ras	89	43	40	0.03	0.53	0.43
+mTOR.no-diff	13	2	2	0	0.85	0.15
+mTOR.diff	7	0	0	0	1	0
+muscle.contr.0-6h	30	41	14	0.47	0.28	0.25
+carc.progr.	74	111	57	0.42	0.13	0.44

Table 3.3: Number of significantly enriched GO terms found analyzing lists of DEGs coming from total and polysomal profiling

	KEGG-INTERPRO-PIR terms statistics					
Short ID	**only tot**	**only poly**	**common**	**gain**	**loss**	**jaccard**
+serum.0-2h	3	7	2	0.62	0.12	0.25
+EPO.0-2h	2	18	2	0.89	0	0.11
-LIF.0-5d	10	19	6	0.56	0.17	0.26
+LPS.0-1h	50	41	34	0.12	0.28	0.60
+LPS.0-2h	45	25	22	0.06	0.48	0.46
+LPS.0-4h	53	44	32	0.18	0.32	0.49
+CD95L.0-48h	0	0	0	0	0	0
+LPS.DC.0-4h	40	32	26	0.13	0.30	0.56
+LPS.DC.4-46h	17	26	14	0.41	0.10	0.48
+LPS.DC.0-16h	48	48	37	0.19	0.19	0.63
+diff.WT.hepa	54	6	4	0.04	0.89	0.07
+diff.mTOR.hepa	68	18	18	0	0.73	0.26
+diff.testis.P17-P22	17	11	7	0.19	0.48	0.33
+diff.testis.P17-P70	25	11	6	0.17	0.63	0.20
+diff.testis.P22-P70	10	6	6	0	0.40	0.60
+diff.lung.E19-E22	11	8	5	0.21	0.43	0.36
+diff.lung.E19-P1	6	5	3	0.25	0.38	0.38
+diff.lung.E22-P1	1	1	0	0.50	0.50	0
+eIF4E	12	10	5	0.29	0.41	0.29
-eIF4GI	18	13	4	0.33	0.52	0.15
+v-Ki-ras	46	13	12	0.02	0.72	0.26
+mTOR.no-diff	20	6	5	0.05	0.71	0.24
+mTOR.diff	20	0	0	0	1	0
+muscle.contr.0-6h	0	2	0	1	0	0
+carc.progr.	16	30	9	0.57	0.19	0.24

Table 3.4: Number of significantly enriched KEGG terms found analyzing lists of DEGs coming from total and polysomal profiling. The number of DEGs found only with transcriptome analysis, only with translatome analysis and in both analyses are visualized.

Chapter 4

Experimental validation of uncoupling

As we said before, translatome analysis by sucrose gradient centrifugation of cell lysates followed by microarray profiling of the polysomal and subpolysomal RNA fractions represents a way of both studying translational control networks and better approximating the proteomic representation of cells. It is an established notion that translational control takes place essentially at the translation initiation level, therefore the variation in abundance of a given mRNA species on polysomes can be directly related to the variation in abundance of the corresponding protein. Comparison of translatome profile changes with corresponding transcriptome profile changes can provide a measure of the degree of concordance between cellular controls affecting mRNA abundance and cellular controls affecting mRNA availability to translation. To provide a direct experimental evaluation of the phenomenon, we decided to study a classical example of transcriptional reprogramming of gene expression: Epidermal Growth Factor (EGF) treatment. This stimulus triggers a well known chain of intracellular transduction events, ultimately resulting in a multifaceted phenotypic spectrum of changes with prevalent induction of cell growth and proliferation. We subjected HeLa cells to serum starvation for 12h and then we

added EGF at final concentration of 1 microgram/ml, profiling before and after 40 minutes of treatment the transcriptome, the translatome, coming from the polysomal pool of mRNAs after sucrose gradient separation, and also the mRNA content of the subpolysomal pool, expected not to be actively translated.

Aim of the chapter. The aim of this experiment is to verify the results obtained by the bioinformatic analysis described in chapter 3, experimentally validating the existence of uncoupling between transcriptome and translatome variations as a general cellular process, and identifying which mechanisms and regulatory circuits are mostly responsible for the reprogramming of gene expression at the translational level.

Materials and methods. To confirm the uncoupling between transcriptome and translatome, proliferative induction triggered by Epidermal Growth Factor (EGF) after serum starvation in HeLa cells has been chosen as biological model. The alteration of this model has been performed by RNA interference, silencing 3 genes deeply involved in post-transcriptional control (4E-T, XRN1 e Dicer). After evaluating the degree of silencing at the protein level by Western blot the protocol of silencing and EGF treatment (40 min) has been performed in biological triplicate leading to the extraction of total RNA, polysomal RNA and subpolysomal RNA (the last two of these RNA classes have been obtained from cytoplasmic sucrose gradient separated fractions). All extracted RNAs have been hybridized on the Agilent-Whole Human Genome Microarray 4x44K platform to obtain gene expression profiles and to compare the significant differences.

Results and discussion. To identify translationally regulated RNAs, gene expression variations derived from polysomal (translatome profiling) and subpolysomal RNA has been compared with those obtained from total RNA (transcriptome profiling) by hybridization of RNA populations on microarrays. In EGF treated HeLa cells have been obtained 693 differ-

entially expressed genes (DEGs) only in transcriptome profiling and 1785 DEGs only in translatome profiling, with an overlapping of 226 (8.4%) genes, confirming an extensive uncoupling between transcribed RNA variations and RNA translation efficiency changes. In 4E-T silenced and EGF treated HeLa cells, 593 DEGs have been obtained for transcriptome and 430 DEGs for translatome, with an overlapping of 70 (6.4%) genes. In comparison with EGF treated HeLa cells, the overall reduction of DEGs, especially at the polysomal level can be imputed to P-bodies disassemblage obtained by 4E-T silencing. In Dicer silenced and EGF treated HeLa cells 1687 DEGs have been obtained for transcriptome and 1282 DEGs for translatome, with an overlapping of 109 (3.5%) genes, demonstrating a general shift of post-transcriptionally regulated genes, especially if we look at the identity of the top up-regulated polysomal genes. This uncoupling has been observed for all the experiments also examining the overlapping degree between the ontological terms associated to the populations of transcriptional and translational DEGs. By interrogating the main biological ontologies, the overlapping degree between the ontological terms associated to the populations of transcriptional and translational DEGs is extremely reduced in all the experiments, even null in Dicer silencing. Conclusions This experimental work confirms the general and profound uncoupling between transcriptome and translatome due to operative intelligence of polysomal machinery. A candidate able to trigger an expression reprogramming at the polysomal level and able to modulate this uncoupling has been identified with P-bodies compartment, where RNAs are transported by interacting RNA-Binding Proteins (RBPs) and microRNAs (miRNAs). This hypothesis has been studied by 4E-T and Dicer silencing, two key genes involved in P-bodies formation and in miRNA pathway.

4.1 Experimental design

Proliferative induction triggered by Epidermal Growth Factor (EGF) after serum starvation in HeLa cells has been chosen as biological model. The reference control consists in serum starved HeLa cells without EGF treatment, while the strong proliferative signal condition consists in HeLa cells treated with EGF for 40 min, as shown in figure 4.1. The "EGF release from starvation" protocol was carried out following the instructions given in [39] and [40]. Total RNA and polysomal/subpolysomal RNA are extracted from cells in each condition and hybridized on the Agilent-Whole Human Genome Microarray 4x44K platform to obtain gene expression profiles and to compare the significant differences. The goal is to observe significant changes in RNA levels in the two conditions and compare differences detected by transcriptome analysis with those detected by translatome analysis. Each gene falls in one of these possible outcomes:

- no change with EGF treatment
- significant changes detected only in total mRNA
- significant changes detected only in polysomal mRNA
- significant changes detected both in total and polysomal mRNA

Post-transcriptional alterations of this model were achieved through siRNA mediated silencing of genes selected for their relevance in the two main post-transcriptional mechanisms theoretically capable of generating widespread uncoupling: p-bodies and the miRNA pathway.

- p-bodies disassembly through 4E-T silencing, as reported by [41] and [42]
- p-bodies increase in size and number through XRN1 silencing, as reported in [43] and [44]

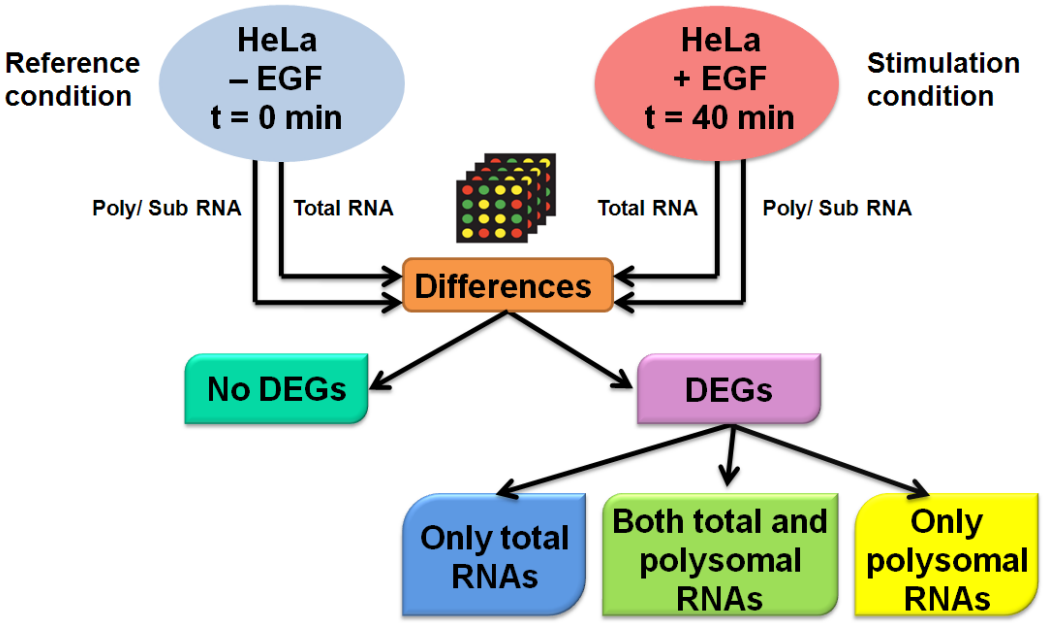


Figure 4.1: HeLA cells are treated with EGF: cellular extracts are collected at time t=0 and t=40, total RNA, polysomal RNA and sub-polysomal RNA are hybridized to Agilent microarray and significant differences between the two conditions are detected. mRNA levels for each gene can either be unaffected or affected by EGF treatment, in the latter case we define these genes DEGs. Comparing transcriptome and translatome mRNA levels, DEGs can be found just one of the two analyses (uncoupling) or in both.

- miRNA pathway suppression through Dicer silencing, as reported by [45] and [46]

The first step of this work has been setting the best experimental condition, especially for silencing protocol, to have the maximum silencing for 4E-T, XRN1 e Dicer. We set the following parameters:

- transfection time
- ratio between concentration of siRNA and Dharmafect
- serum starvation time

The degree of silencing has been checked at the protein level by Western blot, as shown in figure 4.2. We found the best transfection time to be 48h-72h with ratio siRNA-dharmafect equal to 100nM-2micrograms-mL for 4E-T and Xrn1 silencing, 75nM-2micrograms-mL for Dicer silencing. Serum starvation time before EGF treatment was set to 12h. From this last western blots we checked if the silencing was effective also after EGF treatment (40 min). The results show that silencing has been maintained. Residual expression percentages, shown in figure 4.2, are 2-4% for 4E-T, 24-30% for XRN1 and 11-12% for Dicer. Since XRN1 silencing was clearly less efficient than 4E-T and Dicer, we decided to not use it for the following experiments.

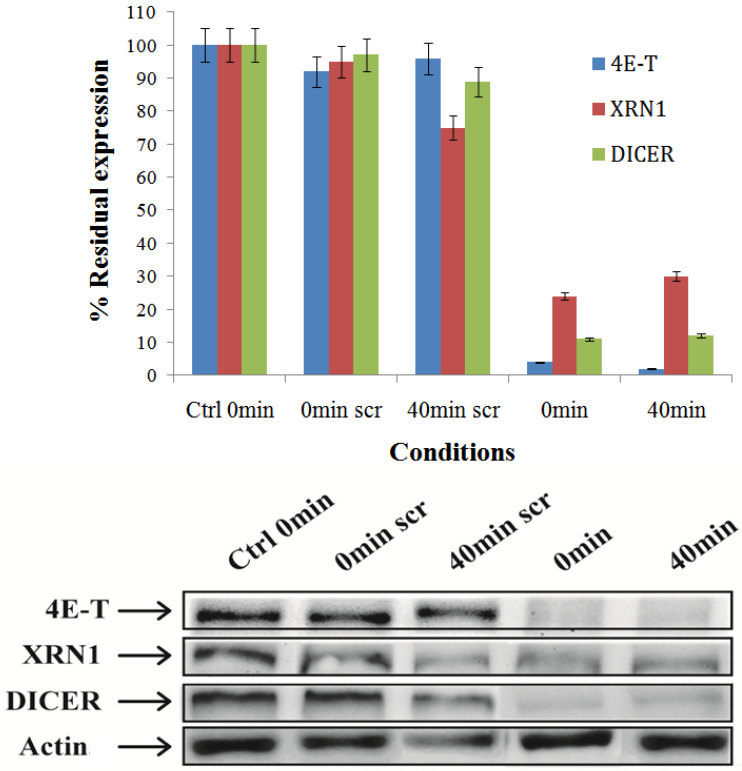


Figure 4.2: Silencing efficiency, expressed as percentage of residual expression, has been measured at the protein level for 4E-T, XRN1 and DICER with western blots, before and after EGF treatment.

After evaluating the degree of silencing at the protein level by Western blot, the protocol of silencing and EGF treatment (40 min) was performed in biological triplicate to extract total RNA, polysomal RNA and subpolysomal RNA. Figure 4.3 shows polysomal profiles obtained from HeLa cells after sedimentation through sucrose gradient centrifugation. For the not silenced sample in the first column (Mock) after EGF treatment we can observe the disappearance the 80s peak and an increase of polysomal peaks. For the 4E-T silenced sample we can observe a decrease of polysomal peaks in both EGF treated and not treated samples, and disappearance of the 80s peak probably due to ribosome recruitment. For the Dicer silencing sample, an increase in both monosomal and polysomal peaks can be observed after EGF treatment. It is important to stress that these first experimental results agree with the general results we will get from microarray signals, in terms of number of differentially expressed genes detected in the three samples.

4.2 Experimental procedures

The comparison between transcriptional and polysomal profiling was used for the discovery of general and mRNA-specific changes in the translation state of the serum starved HeLa cells transcriptome in response to EGF stimulus. To identify translationally regulated mRNA molecules, gene expression signals derived from the polysomal and subpolysomal RNA populations were compared by microarrays analysis to those obtained from unfractionated total RNAs. Polysomal RNA, subpolysomal RNA and total RNA were isolated from HeLa cells serum starved and treated with EGF. Cells lysates were collected before (t = 0 min) and after (t = 40 min) EGF treatment. All experiments were run in biological triplicates.

HeLa cells were seeded on adherent plates and serum starved for 12h

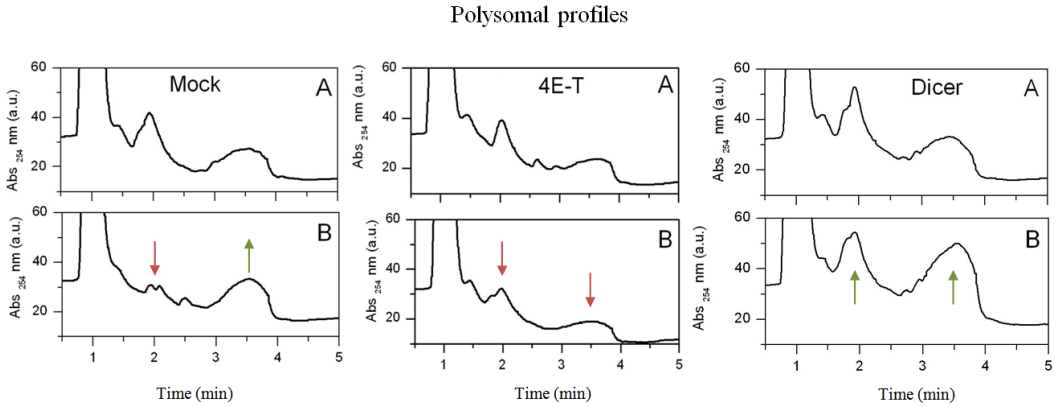


Figure 4.3: Polysomal profiles, measured as 254nm absorbance after sucrose gradient centrifugation of HeLa extracts before and after EGF treatment. Without silencing (Mock) an increase in the peak corresponding to the polysomal fraction and a corresponding decrease in the 80s peak are observed upon EGF treatment. With 4E-T silencing, both 80s and polysomal peak are not affected by EGF treatment. On the contrary, with Dicer signalling both 80s and polysomal peaks show an increase with EGF treatment.

with DMEM, 0.5% FBS, 2mM glutamine. Cells were treated for 40 minutes with recombinant human epidermal growth factor (hEGF) at final concentration of 1 microgram/ml. HeLa cells were cultured in Dulbecco Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 2mM glutamine, 100 units/ml penicillin, and 100 mg/ml streptomycin at 37 C in a humidified atmosphere of 5% CO2.

Total RNA was extracted using TRIZOL reagent according to the manufacturer's protocol. Briefly, the aqueous phase was used for RNA precipitation with an equal volume of isopropanol. The RNA pellet was washed once with 75% ethanol, then air-dried and re-dissolved in 20 microliters of RNase-free water. RNA was quantified using a spectrophotometer and its quality was checked by agarose gel electrophoresis and by Agilent 2100 Bioanalyzer platform, following the manifacturers guidelines for sample preparation and analysis of data (Agilent 2100 Bioanalyzer 2100 Expert

User's Guide [1].

For polysomal extraction, cells were washed once with phosphate buffer saline (PBS) and treated directly on the plate with 300 l lysis buffer [10 mM NaCl, 10 mM MgCl2, 10 mM TrisHCl, pH 7.5, 1% Triton X-100, 1% sodium deoxycholate, 0.2 U/microliter RNase inhibitor (Fermentas) and 1 mM dithiothreitol] and transferred to an Eppendorf tube. After a few minute incubation on ice with occasional vortexing, the extracts were centrifuged for 10 min at 12000 g at 4C. The supernatant was stored at 80C or loaded directly onto a 1550% linear sucrose gradient containing 30 mM TrisHCl, pH 7.5, 100 mM NaCl, 10 mM MgCl2, and centrifuged in an Sorvall rotor for 120 min at 40000 rpm. Polysomal and subpolysomal fractions were collected monitoring the absorbance at 254 nm and treated directly with proteinase K. After phenolchloroform extraction and isopropanol precipitation, polysomal RNA was resuspended in 30 microliters of water and then repurified with RNeasy kit (Qiagen, Hilden, Germany). RNA quality was assessed by agarose gel electrophoresis and the Agilent 2100 Bioanalyzer platform.

Microarray hybridization, blocking and washing were performed according to Agilent protocol: One-Color Microarray-Based Gene Expression Analysis (Quick Amp Labeling)[2].

Hybridized microarray slides were scanned with an Agilent DNA Microarray Scanner (G2505C, Agilent Technologies, Santa Clara, CA) at 5 micron resolution with the manufacturers software (Agilent ScanControl 8.1.3). The scanned TIFF images were analyzed numerically using the Agilent Feature Extraction Software version 10.7.7.1 according to the Agilent standard protocol GE1-107-Sep09.

[1] http://www.agilent.com

[2] https://www.chem.agilent.com/Library/usermanuals/Public/G4140-90040.pdf

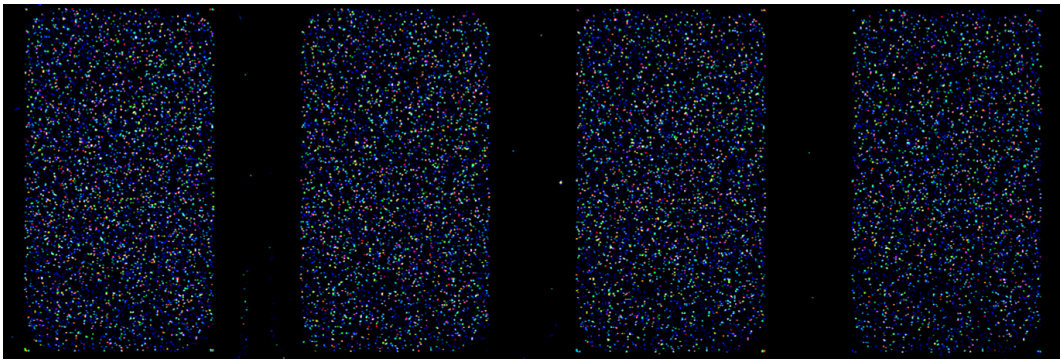


Figure 4.4: Agilent scanned slide showing with in color scale the fluorescence intensity signals coming from four different hybridizations.

	mock	**4E-T**	**Dicer**
Total RNA 0 min	100%	100%	100%
Polysomal RNA 0 min	18%	27%	49%
Subpolysomal RNA 0 min	82%	73%	49%
Total RNA 40 min	100%	100%	100%
Polysomal RNA 40 min	36%	27%	51%
Subpolysomal RNA 40 min	64%	73%	51%

Table 4.1: Percentages of polysomal and subpolysomal RNA quantities with respect to total RNA quantities, calculated on the basis of RNA quantities (micrograms) measured with nanodrop

Series GSE20277 UPDATE **Click here to create a reviewer access link**

Status	**Private until May 09, 2010**
Title	Translatome and transcriptome profiling of EGF response in HeLa cells
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Translatome analysis by sucrose gradient centrifugation of cell lysates followed by microarray profiling of the polysomal and subpolysomal RNA fractions represents a way of both studying translational control networks and better approximating the proteomic representation of cells. It is an established notion that translational control takes place essentially at the translation initiation level, therefore the variation in abundance of a given mRNA species on polysomes can be directly related to the variation in abundance of the corresponding protein. Comparison of translatome profile changes with corresponding transcriptome profile changes can provide a measure of the degree of concordance between cellular controls affecting mRNA abundance and cellular controls affecting mRNA availability to translation. To provide a direct experimental evaluation of the phenomenon, we decided to study a classical example of transcriptional reprogramming of gene expression: Epidermal Growth Factor (EGF) treatment. This stimulus triggers a well known chain of intracellular transduction events, ultimately resulting in a multifaceted phenotypic spectrum of changes with prevalent induction of cell growth and proliferation. We subjected HeLa cells to serum starvation for 12h and then we added EGF at final concentration of 1 µg/ml, profiling before and after 40 minutes of treatment the transcriptome, the translatome, coming from the polysomal pool of mRNAs after sucrose gradient separation, and also the mRNA content of the subpolysomal pool, expected not to be actively translated. Keywords: translatome profiling, polysomal profiling, polysomal RNA, translational control, translational profiling, polysome profiling, post-transcriptional regulation, EGF starvation release.
Overall design	The comparison between transcriptional and polysomal profiling was used for the discovery of general and mRNA-specific changes in the translation state of the serum starved HeLa cells transcriptome in response to EGF stimulus. To identify translationally regulated mRNA molecules, gene expression signals derived from the polysomal and subpolysomal RNA populations were compared by microarrays analysis to those obtained from unfractionated total RNAs. Polysomal RNA, subpolysomal RNA and total RNA were isolated from HeLa cells serum starved and treated with EGF. Cells lysates were collected before (t = 0 min) and after (t = 40 min) EGF treatment. All experiments were run in triplicates.
Contributor(s)	Tebaldi T, Viero G, Pegoretti I, Adami V, Re A, Quattrone A
Citation missing	*Has this study been published? If so, please notify GEO.*
Submission date	Feb 11, 2010
Last update date	Feb 12, 2010
Contact name	Toma Tebaldi
E-mail(s)	tebaldi@science.unitn.it
Organization name	University of Trento
Department	Centre for Integrative Biology
Lab	Laboratory of Translational Genomics

Figure 4.5: Microarray data have already been submitted to GEO and accepted as satisfying the MIAME standard. The picture shows a snapshot of the GEO data series containing our hybridization raw and processed signals.

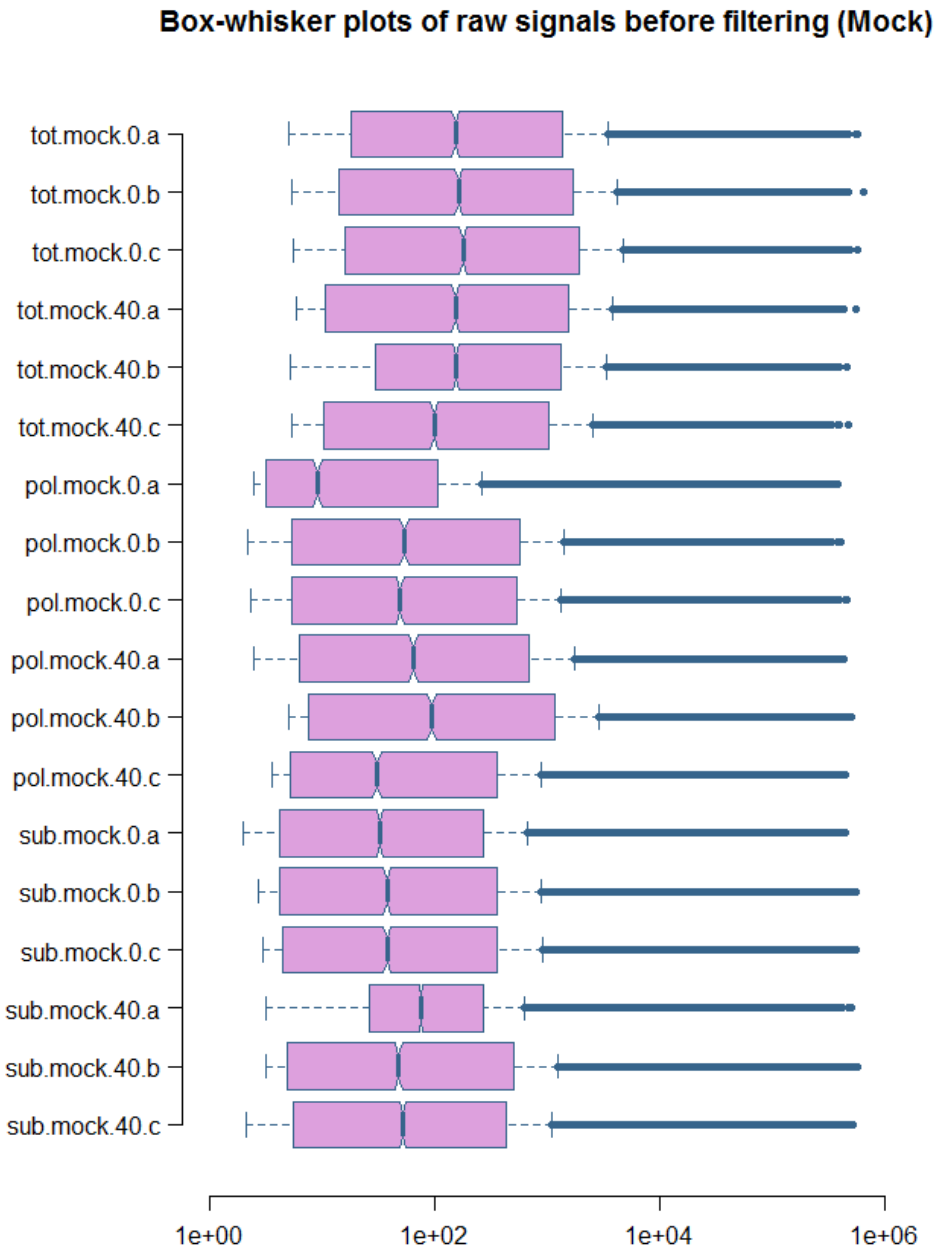


Figure 4.6: Raw signals before normalization are shown for each array belonging to the EGF induction experiment without any silencing. Analogous graphs were obtained for the arrays belongigng to the 4E-T silencing experiment and the Dicer silencing experiment.

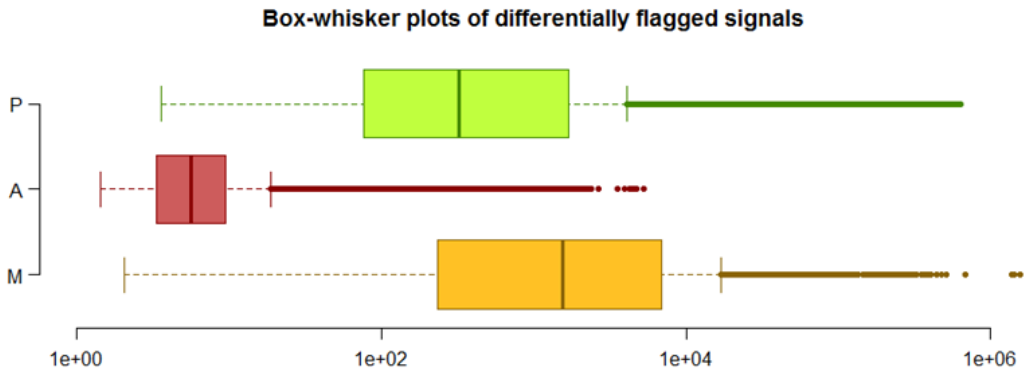


Figure 4.7: Distribution of raw intensity signals associated to detection calls of Agilent features: Absent, Present and Marginal, the last of which included unreliable spots whose signals were removed from following analyses.

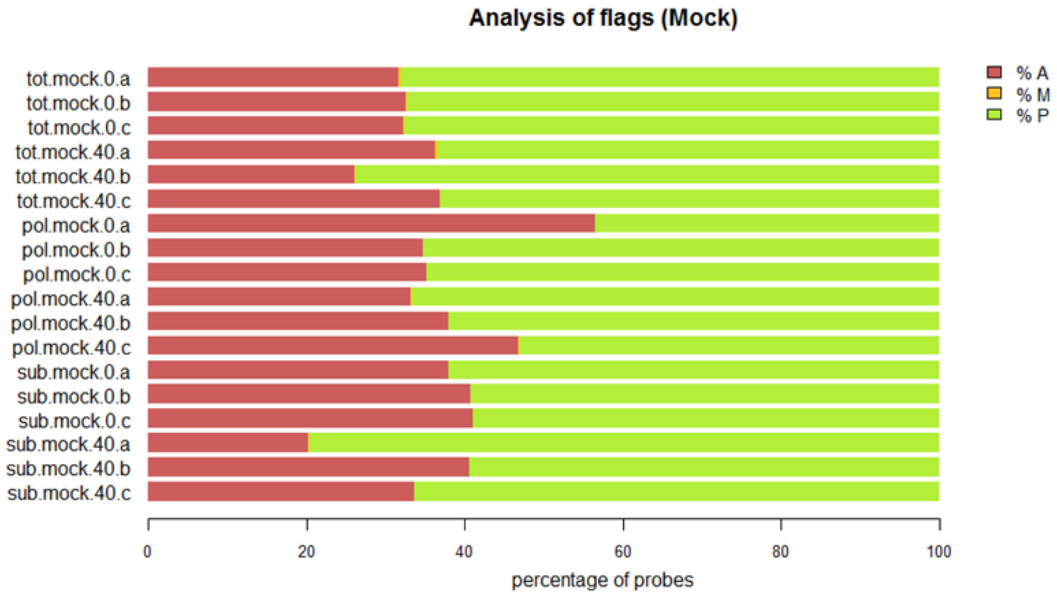


Figure 4.8: Percentages of Agilent features associated to different categorical values of detection call: Absent, Present and Marginal. Each bar represent a single hybridization. Data are visualized only for arrays belonging to the EGF experiment without any silencing. The percentage of absent features is influenced by the level of background signal on the array, which can vary according to experimental hybridization conditions.

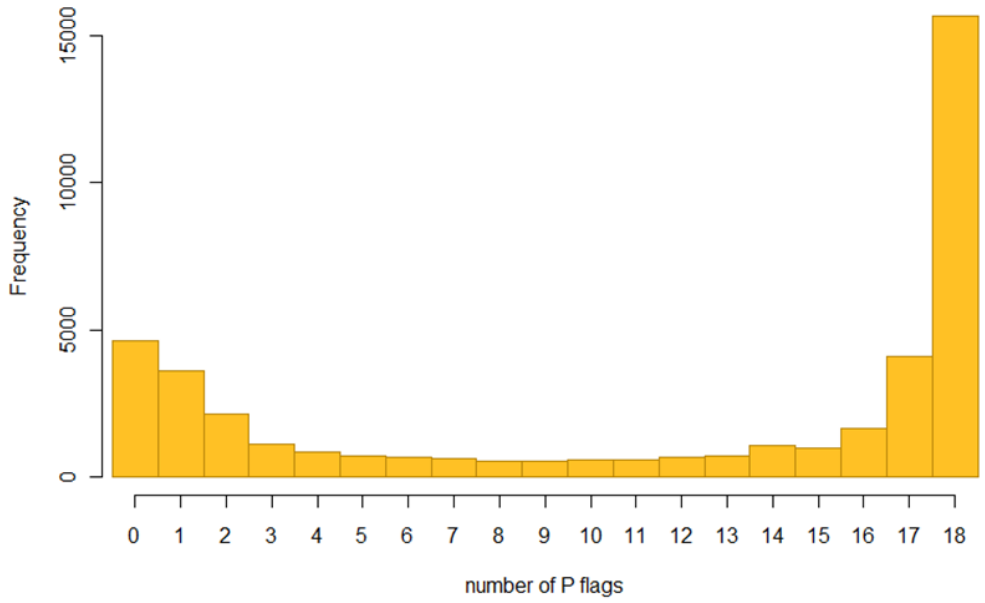


Figure 4.9: Histogram of the number of Present calls associated to each feature in the 18 experiments belonging to the EGF induction experiment without any silencing (Mock). The majority of features has 18 out of 18 Present calls. The filtering procedure removes features without 2 out of 3 Present calls in the biological replicas of at least one out of six experimental conditions.

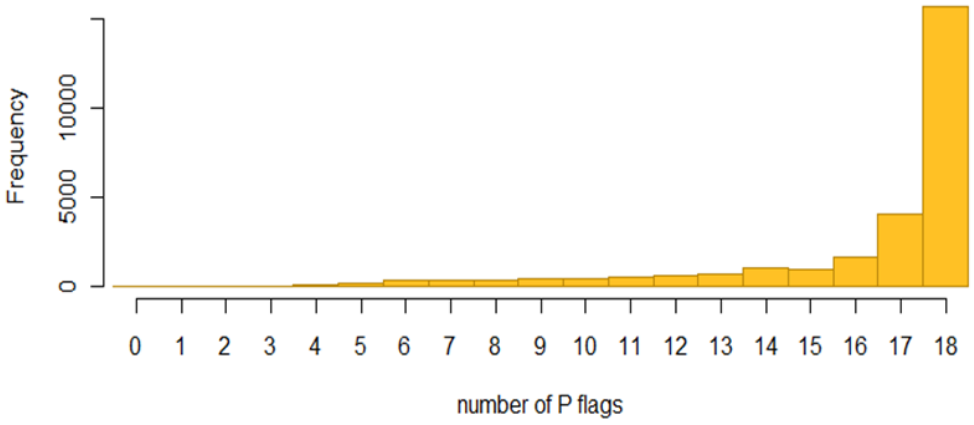


Figure 4.10: Histogram of the number of Present calls associated to each feature in the 18 experiments belonging to the EGF induction experiment without any silencing (Mock) after the filtering procedure. 11025 features were called as Absent in the majority of hybridizations and did not fulfill the filtering requirements described before, therefore they were removed from the analysis.

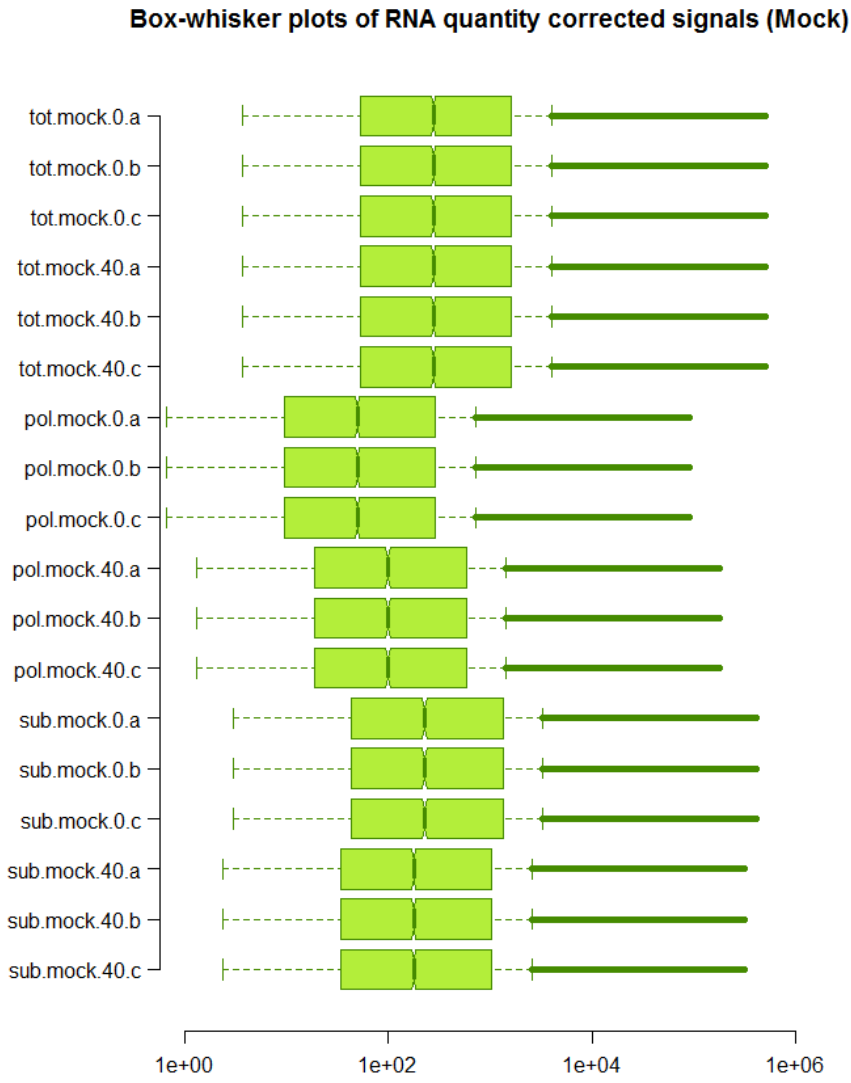


Figure 4.11: Distribution of array signal intensities after filtering of absent features, quantile normalization and correction of signals according to RNA quantities listed in table 4.1

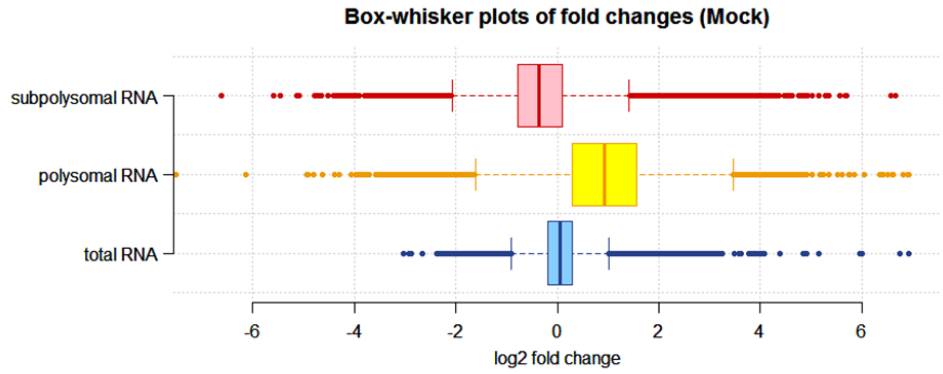


Figure 4.12: The distributions of log2 Fold Changes for the 30075 features considered after filtering of absent flags. The distribution of total RNA fold changes (transcriptome) is centered around 0, while the distribution of polysomal fold changes is centered around 1, reflecting the observed increase in polysomal content of HeLa cells after EGF treatment. On the other hand, the distribution of subpolysomal fold changes is centered around -0.5. The distribution of polysomal fold changes is also more dispersed, and this is reflected by the higher number of differentially expressed genes detected at the polysomal level.

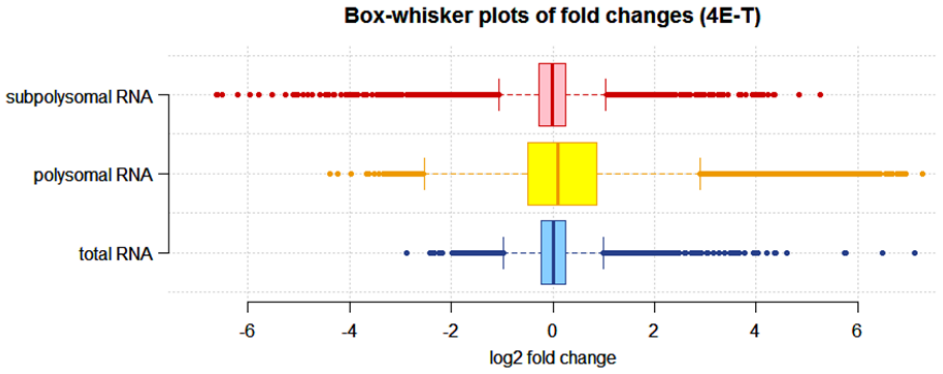


Figure 4.13: The distributions of log2 Fold Changes for the 29950 features considered after filtering of absent flags.All the three distributions are centered around 0, polysomal fold changes are still more dispersed.

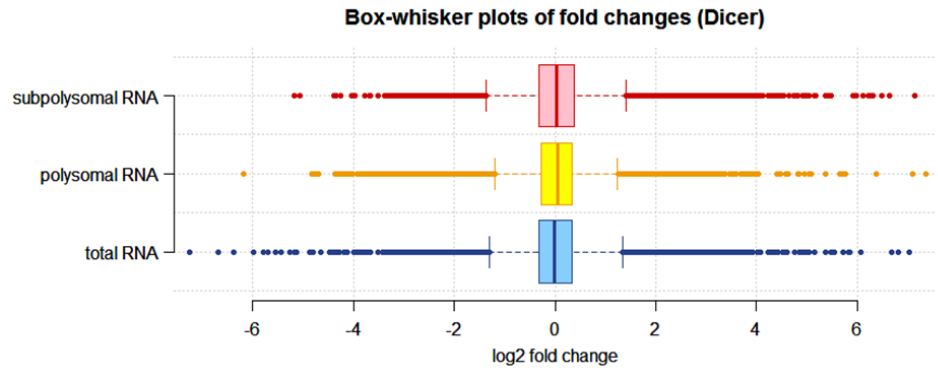


Figure 4.14: The distributions of log2 Fold Changes for the 29987 features considered after filtering of absent flags. All the three distributions are centered around 0 and have the same level of dispersion.

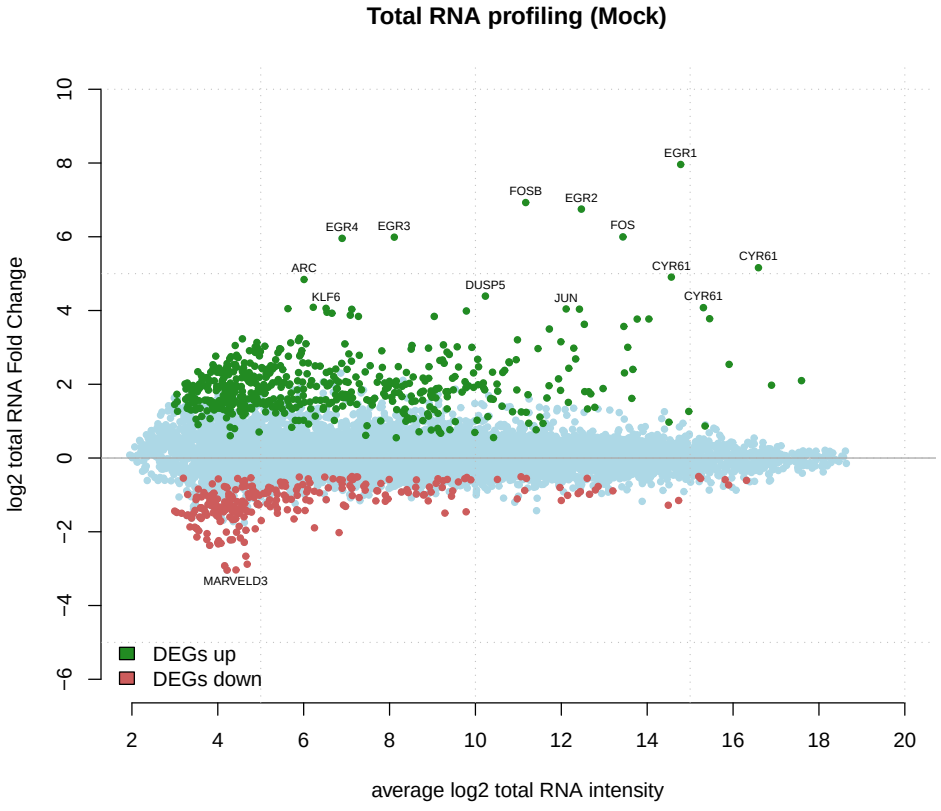


Figure 4.15: The figure shows up-expressed and down-expressed genes upon EGF treatment, identified as differentially expressed genes by RankProd.

Bibliography

[1] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561563, 1970.

[2] Alwin Köhler and Ed Hurt. Exporting RNA from the nucleus to the cytoplasm. *Nature reviews. Molecular cell biology*, 8(10):761–73, 2007.

[3] Arianne Heinrichs. Nuclear transport: Exit for fly mRNA. *Nature Reviews Molecular Cell Biology*, 9(2):92–93, 2008.

[4] Florence Besse and Anne Ephrussi. Translational control of localized mRNAs: restricting protein synthesis in space and time. *Nature reviews. Molecular cell biology*, 9(12):971–80, 2008.

[5] Yaron Shav-Tal and Robert H Singer. RNA localization. *Journal of cell science*, 118(Pt 18):4077–81, 2005.

[6] Kazuko Nishikura. Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nature reviews. Molecular cell biology*, 7(12):919–31, 2006.

[7] Nicole L Garneau, Jeffrey Wilusz, and Carol J Wilusz. The highways and byways of mRNA decay. *Nature reviews. Molecular cell biology*, 8(2):113–26, 2007.

[8] J Guhaniyogi and G Brewer. Regulation of mRNA stability in mammalian cells. *Gene*, 265(1-2):11–23, marzo 2001.

[9] V Narry Kim, Jinju Han, and Mikiko C Siomi. Biogenesis of small RNAs in animals. *Nature reviews. Molecular cell biology*, 10(2):126–39, 2009.

[10] S E Wells, P E Hillner, R D Vale, and a B Sachs. Circularization of mRNA by eukaryotic translation initiation factors. *Molecular cell*, 2(1):135–40, luglio 1998.

[11] Richard J Jackson, Christopher U T Hellen, and Tatyana V Pestova. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature reviews. Molecular cell biology*, 11(2):113–27, 2010.

[12] Nadia Amrani, Matthew S Sachs, and Allan Jacobson. Early nonsense: mRNA decay solves a translational problem. *Nature reviews. Molecular cell biology*, 7(6):415–25, 2006.

[13] B.M. Lunde, C Moore, and G Varani. RNA-binding proteins: modular design for efficient function. *Nature Reviews Molecular Cell Biology*, 8(6):479490, 2007.

[14] Tariq M Rana. Illuminating the silence: understanding the structure and function of small RNAs. *Nature reviews. Molecular cell biology*, 8(1):23–36, 2007.

[15] D Melamed and Y Arava. Genome-Wide Analysis of mRNA Polysomal Profiles with Spotted DNA Microarrays. *Methods in Enzymology*, 431(07):177–201, 2007.

[16] Q Zong, M Schummer, L Hood, and D R Morris. Messenger RNA translation state: the second dimension of high-throughput expression screening. *Proceedings of the National Academy of Sciences of the United States of America*, 96(19):10632–6, settembre 1999.

[17] Qiang Tian, Serguei B Stepaniants, Mao Mao, Lee Weng, Megan C Feetham, Michelle J Doyle, Eugene C Yi, Hongyue Dai, Vesteinn Thorsson, Jimmy Eng, David Goodlett, Joel P Berger, Bert Gunter, Peter S Linseley, Roland B Stoughton, Ruedi Aebersold, Steven J Collins, William a Hanlon, and Leroy E Hood. Integrated genomic and proteomic analyses of gene expression in Mammalian cells. *Molecular & cellular proteomics : MCP*, 3(10):960–9, 2004.

[18] D Greenbaum, C Colangelo, K Williams, and M. Gerstein. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol*, 4(9):117, 2003.

[19] B M Bolstad, R A Irizarry, M Astrand, and T P Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[20] Rainer Breitling, Patrick Armengaud, Anna Amtmann, and Pawel Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*, 573(1-3):83–92, 2004.

[21] Fangxin Hong and Rainer Breitling. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics (Oxford, England)*, 24(3):374–82, 2008.

[22] C Barreau, L Paillard, and H B Osborne. AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Res.*, 33(22):7138–7150, 2005.

[23] Paul Anderson and Nancy Kedersha. RNA granules: post-transcriptional and epigenetic modulators of gene expression. *Nature reviews. Molecular cell biology*, 10(6):430–6, 2009.

[24] Jack D Keene. RNA regulons: coordination of post-transcriptional events. *Nature reviews. Genetics*, 8(7):533–43, 2007.

[25] Regula E Halbeisen and André P Gerber. Stress-Dependent Coordination of Transcriptome and Translatome in Yeast. *PLoS biology*, 7(5):e105, 2009.

[26] M Kenzelmann, S Maertens, M Hergenhahn, S Kueffer, A Hotzwagenblatt, L Li, S Wang, C Ittrich, T Lemberger, R Arribas, S Jonnakuty, M.C. Holstein, W Schmid, N Grets, H.J. Grone, and G Schutz. Microarray analysis of newly synthesized RNA in cells and animals. *PNAS*, 104(15):6164–6169, 2007.

[27] Godfrey Grech, Montserrat Blazquex-Domingo, Andrea Kolbus, Walbert J Bakker, Ernst W Mullner, and Hartmut Beug. Igbp1 is part of a positive feedback loop in stem cell factor dependent , selective mRNA translation initiation inhibiting erythroid differentiation. *Blood*, 112(7):2750–2760, 2008.

[28] Prabha Sampath, David K Pritchard, Lil Pabon, Hans Reinecke, Stephen M Schwartz, David R Morris, and Charles E Murry. A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation. *Cell stem cell*, 2(5):448–60, 2008.

[29] Hiroshi Kitamura, Masatoshi Ito, Tomoko Yuasa, Chisato Kikuguchi, Atsushi Hijikata, Michiyo Takayama, Yayoi Kimura, Ryo Yokoyama, Tomohiro Kaji, and Osamu Ohara. Genome-wide identification and characterization of transcripts translationally regulated by bacterial lipopolysaccharide in macrophage-like J774.1 cells. *Physiological genomics*, 33(1):121–32, 2008.

[30] Nina S Corsini, Ignacio Sancho-Martinez, Sabrina Laudenklos, Dsire Glagow, Sachin Kumar, Elisabeth Letellier, Philipp Koch, Marcin Teodorczyk, Susanne Kleber, Stefan Klussmann, Benedict Wiestler, Oliver Brstle, Wolf Mueller, Christian Gieffers, Oliver Hill, Meinolf Thiemann, Matthias Seedorf, Norbert Gretz, Rolf Sprengel, Tansu Celikel, and Ana Martin-Villalba. The Death Receptor CD95 Activates Adult Neural Stem Cells for Working Memory Formation and Brain Repair. *Cell Stem Cell*, 5(2):178–190, agosto 2009.

[31] Maurizio Ceppi, Giovanna Clavarino, Evelina Gatti, Enrico K Schmidt, Aude de Gassart, Derek Blankenship, Gerald Ogola, Jacques Banchereau, Damien Chaussabel, and Philippe Pierre. Ribosomal protein mRNAs are translationally-regulated during human dendritic cells activation by LPS. *Immunome research*, 5:5, 2009.

[32] Romain Parent and Laura Beretta. Translational control plays a prominent role in the hepatocytic differentiation of HepaRG liver progenitor cells. *Genome biology*, 9(1):R19, 2008.

[33] Romain Parent, Deepak Kolippakkam, Garrett Booth, and Laura Beretta. Mammalian target of rapamycin activation impairs hepatocytic differentiation and targets genes moderating lipid homeostasis and hepatocellular growth. *Cancer research*, 67(9):4337–45, 2007.

[34] Naoko Iguchi, John W Tobias, and Norman B Hecht. Expression profiling reveals meiotic male germ cell mRNAs that are translationally up- and down-regulated. *Proceedings of the National Academy of Sciences of the United States of America*, 103(20):7712–7, maggio 2006.

[35] Ola Larsson, Shunan Li, Olga a Issaenko, Svetlana Avdulov, Mark Peterson, Karen Smith, Peter B Bitterman, and Vitaly a Polunovsky.

Eukaryotic translation initiation factor 4E induced progression of primary human mammary epithelial cells along the cancer pathway is associated with targeted translational deregulation of oncogenic drivers and inhibitors. *Cancer research*, 67(14):6814–24, 2007.

[36] Jean Spence, Brendan M Duggan, Colleen Eckhardt, Michael McClelland, and Dan Mercola. Messenger RNAs under differential translational control in Ki-ras-transformed cells. *Molecular cancer research : MCR*, 4(1):47–60, 2006.

[37] Y.-W. Chen, G. a Nader, K. R Baar, M. J Fedele, E. P Hoffman, and K. a Esser. Response of rat muscle to acute resistance exercise defined by transcriptional and translational profiling. *The Journal of Physiology*, 545(1):27–41, 2002.

[38] Alessandro Provenzani, Raffaele Fronza, Fabrizio Loreni, Alessia Pascale, Marialaura Amadio, and Alessandro Quattrone. Global alterations in mRNA polysomal recruitment in a cell model of colorectal cancer progression to metastasis. *Carcinogenesis*, 27(7):1323–33, 2006.

[39] Blagoy Blagoev, Shao-En Ong, Irina Kratchmarova, and Matthias Mann. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nature biotechnology*, 22(9):1139–45, 2004.

[40] Jesper V Olsen, Blagoy Blagoev, Florian Gnad, Boris Macek, Chanchal Kumar, Peter Mortensen, and Matthias Mann. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, 127(3):635–48, 2006.

[41] Ana Eulalio, Isabelle Behm-Ansmant, and Elisa Izaurralde. P bodies: at the crossroads of post-transcriptional pathways. *Nature reviews. Molecular cell biology*, 8(1):9–22, 2007.

[42] Maria a Ferraiuolo, Sanjukta Basak, Josee Dostie, Elizabeth L Murray, Daniel R Schoenberg, and Nahum Sonenberg. A role for the eIF4E-binding protein 4E-T in P-body formation and mRNA decay. *The Journal of cell biology*, 170(6):913–24, 2005.

[43] Georg Stoecklin, Thomas Mayo, and Paul Anderson. ARE-mRNA degradation requires the 5'-3' decay pathway., gennaio 2006.

[44] D Ingelfinger, R Luhrmann, and T Achsel. The human LSm1-7 proteins colocalize with the mRNA-degrading enzymes Dcp1 / 2 and Xrnl in distinct cytoplasmic foci. *RNA*, 8:1489–1501, 2002.

[45] Changning Liu, Baoyan Bai, Geir Skogerbø, Lun Cai, Wei Deng, Yong Zhang, Dongbo Bu, Yi Zhao, and Runsheng Chen. NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic acids research*, 33(Database issue):D112–5, 2005.

[46] Robinson Triboulet, Bernard Mari, Yea-Lih Lin, Christine Chable-Bessia, Yamina Bennasser, Kevin Lebrigand, Bruno Cardinaud, Thomas Maurin, Pascal Barbry, Vincent Baillat, Jacques Reynes, Pierre Corbeau, Kuan-Teh Jeang, and Monsef Benkirane. Suppression of microRNA-silencing pathway by HIV-1 during virus replication. *Science (New York, N.Y.)*, 315(5818):1579–82, 2007.

[47] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–5, 2007.

[48] Olivier Bodenreider and Robert Stevens. Bio-ontologies: current trends and future directions. *Briefings in bioinformatics*, 7(3):256–74, 2006.

[49] Daniel L Rubin, Nigam H Shah, and Natalya F Noy. Biomedical ontologies: a functional perspective. *Briefings in bioinformatics*, 9(1):75–90, 2008.

[50] Mikel Egaña Aranguren, Sean Bechhofer, Phillip Lord, Ulrike Sattler, and Robert Stevens. Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. *BMC bioinformatics*, 8:57, 2007.

[51] Dilvan a Moreira and Mark a Musen. OBO to OWL: a protege OWL tab to read/save OBO ontologies. *Bioinformatics (Oxford, England)*, 23(14):1868–70, 2007.

[52] John Day-Richter, Midori a Harris, Melissa Haendel, and Suzanna Lewis. OBO-Edit–an ontology editor for biologists. *Bioinformatics (Oxford, England)*, 23(16):2198–200, 2007.

[53] B Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. Rector, and C. Rosse. Relations in biomedical ontologies. *Genome Biology*, 6(5):R46, 2005.

[54] Tina Glisovic, Jennifer L Bachorik, Jeongsik Yong, and Gideon Dreyfuss. RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters*, 582(14):1977–86, 2008.

[55] Grant H Jacobs, Augustine Chen, Stewart G Stevens, Peter a Stockwell, Michael a Black, Warren P Tate, and Chris M Brown. Transterm: a database to aid the analysis of regulatory sequences in mRNAs. *Nucleic acids research*, 37(Database issue):D72–6, 2009.

[56] Paul Flicek, Bronwen L Aken, Benoit Ballester, Kathryn Beal, Eugene Bragin, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Julio Fernandez-Banet, Leo Gordon, Stefan Gräf, Syed Haider, Martin Hammond, Kerstin Howe, Andrew Jenkinson, Nathan Johnson, Andreas Kähäri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Felix Kokocinski, Gautier Koscielny, Eugene Kulesha, Daniel Lawson, Ian Longden, Tim Massingham, William McLaren, Karine Megy, Bert Overduin, Bethan Pritchard, Daniel Rios, Magali Ruffier, Michael Schuster, Guy Slater, Damian Smedley, Giulietta Spudich, Y Amy Tang, Stephen Trevanion, Albert Vilella, Jan Vogel, Simon White, Steven P Wilder, Amonida Zadissa, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé M Fernández-Suarez, Javier Herrero, Tim J P Hubbard, Anne Parker, Glenn Proctor, James Smith, and Stephen M J Searle. Ensembl's 10th year. *Nucleic acids research*, 38(Database issue):D557–62, 2010.

[57] Volker Haarslev and R. Möller. Racer: An owl reasoning agent for the semantic web. In *Proceedings of the International Workshop on Applications, Products and Services of Web-based Support Systems, in conjunction with the*, page 9195. Citeseer, 2003.

[58] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, Kimberly a Marshall, Katherine H Phillippy, Patti M Sherman, Rolf N Muertter, and Ron Edgar. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic acids research*, 37(Database issue):D885–90, 2009.

[59] Helen Parkinson, Misha Kapushesky, Nikolay Kolesnikov, Gabriella Rustici, Mohammad Shojatalab, Niran Abeygunawardena, Hugo Berube, Miroslaw Dylag, Ibrahim Emam, Anna Farne, Ele Hol-

loway, Margus Lukk, James Malone, Roby Mani, Ekaterina Pilicheva, Tim F Rayner, Faisal Rezwan, Anjan Sharma, Eleanor Williams, Xiangqun Zheng Bradley, Tomasz Adamusiak, Marco Brandizi, Tony Burdett, Richard Coulson, Maria Krestyaninova, Pavel Kurnosov, Eamonn Maguire, Sudeshna Guha Neogi, Philippe Rocca-Serra, Susanna-Assunta Sansone, Nataliya Sklyar, Mengyao Zhao, Ugis Sarkans, and Alvis Brazma. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic acids research*, 37(Database issue):D868–72, 2009.

[60] Rafael a Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, 4(2):249–64, aprile 2003.

[61] Tom Payne, Colin Hanfrey, Amy L Bishop, Anthony J Michael, Simon V Avery, and David B Archer. Transcript-specific translational regulation in the unfolded protein response of Saccharomyces cerevisiae. *FEBS letters*, 582(4):503–9, 2008.

[62] Vivian L MacKay, Xiaohong Li, Mark R Flory, Eileen Turcott, G Lynn Law, Kyle a Serikawa, X L Xu, Hookeun Lee, David R Goodlett, Ruedi Aebersold, Lue Ping Zhao, and David R Morris. Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone. *Molecular & cellular proteomics : MCP*, 3(5):478–89, 2004.

[63] D Shenton, J B Smirnova, J N Selley, K Carroll, S J Hubbard, G D Pavitt, M P Ashe, and C M Grant. Global translational responses to

oxidative stress impact upon multiple levels of protein synthesis. *J. Biol. Chem.*, 281(39):29011–29021, 2006.

[64] J B Smirnova, J N Selley, F Sanchez-cabo, K Carroll, A A Eddy, J E Mccarthy, S J Hubbard, G D Pavitt, C M Grant, and M P Ashe. Global gene expression profiling reveals widespread yet distinctive translational responses to different eukaryotic translation initiation factor 2B-targeting stress pathways. *Molecular and Cellular Biology*, 25(21):9340–9349, 2005.

Appendix A

The USER Ontology

The USER (Untranslated Sequence Elements for Regulation) ontology is a structured controlled vocabulary designed to describe post-transcriptional regulation mechanisms of gene expression, which take place in every cell in order to control protein production. It has been designed for the annotation and classification of transcripted but untranslated sequence elements, for example 3' and 5' untranslated regions (UTR) on mRNAs or noncoding RNAs, involved in post-transcriptional controls. It provides a standardized set of terms, definitions, relationships and axioms that facilitate the formal and consistent analysis of data about post-transcriptional regulation and that will make possible the automated reasoning over their contents.[1] This ontology has been developed, together with the results of both computational and experimental analyses, as a basis to create a new bioinformatics knowledgebase which will enable the generation of predictions on the probability that a particular mRNA molecule is post-transcriptionally regulated. The necessity of such an instrument has increased after the progressive discovery in the last years of the influence of post-transcriptional controls, such as the microRNA expression silencing system, on gene expression and other important cellular mechanisms. An important goal for this database will be the collection and the integration of different types

[1] `http://www.obofoundry.org/`

of data about possible controls affecting mRNA molecules (sequence and structural data coming from appendix **??**, presence of particular motifs or three-dimensional folds, expression data from polysomal profiling experimentscollected in chapter 3 and in appendix B, and generated in chapter 4. These data sets have to be connected and merged into a single functional context to allow the analysis of their correlation, and a useful way to structure and connect data elements avoiding redundancy is the development of an ontology that represents and conceptualizes particular knowledge domains. The general structure of the ontology, with classes and properties, has been defined. A first draft of the ontology (USER-OBO) has been written following the OBO foundry guidelines (the Open Biomedical Ontologies Foundry is a collaborative project with the goal of creating a suite of reference ontologies in the biomedical domain) [47]. This version is interoperable and orthogonal to other OBO ontologies such as the Gene Ontology (the most known and used bio-ontology: a controlled vocabulary to describe gene and gene product attributes in any organism)[2] or the Sequence Ontology (an ontology suitable for describing features and attributes of biological sequences)[3]. As a second step the USER ontology has been manually translated in the OWL-DL Knowledge Representation language (USER-OWL).

A.1 Use of biomedical ontologies

Ontologies are specific (theoretical or computational) artifacts expressing the intended meaning of a vocabulary in terms of primitive categories and relations describing the nature and structure of a domain of discourse. Biologists have always classified the phenomena they observed in the biological world around them (medieval bestiaries, lists of ways in which

[2] `http://www.geneontology.org/`

[3] `http://www.sequenceontology.org/`

people died, Linnaean classification of species), but only the advent of bioinformatics has caused the birth of the first computer-based conceptual models addressed to the biomedical knowledge domain, in order to share unambiguously what is known about the world of biomedicine. The Gene Ontology, due to community involvement, clear goals, limited scope, simple structure, continuous evolution and immediate applications, has been the most successful biomedical ontology, being responsible of the publication of more than half of all the ontology papers in Pubmed in the last years [48], [49]. A map of some well known biological ontologies is represented in figure A.1, taken from [4].

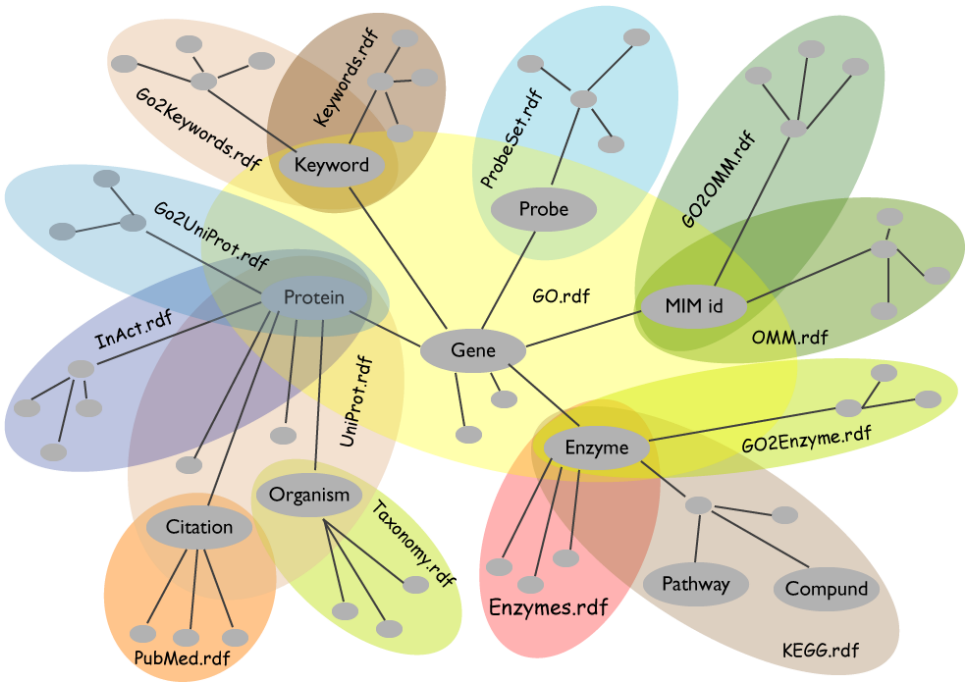


Figure A.1: The image shows a map of connections between biological knowledge domains modelled by some of the most known biological ontologies and centered aroung the Gene Ontology.

[4] http://www.w3.org/2007/Talks/0403-Tampere-IH/

A.2 Comparison between OBO and OWL

Different Knowledge Representation languages provide different means to make statements about knowledge to be captured. Different languages have varying expressivity and computational properties according to their semantics. The comparison between OBO and OWL formats can be considered a symbol of the broader comparison between two alternative approaches towards ontologies, adopted respectively by biologists and computer scientists. In this sense the comparison can be properly introduced by this citation from [50]:

"The bio-ontology community falls into two camps: first biology domain experts, who actually hold the knowledge we wish to capture in ontologies; second, ontology specialists, who hold knowledge about techniques and best practice on ontology development. In the bio-ontology domain, these two camps have often come into conflict, especially where pragmatism comes into conflict with perceived best practice (for example the insistence of computer scientists on a well-defined semantic basis for the Knowledge Representation language being used)."

Today the Gene Ontology and a significant number of bio-ontologies are in the OBO-format, which has evolved to support the needs of the bio-ontologies under the Open Biomedical Ontologies umbrella and aims to have human readability, ease of parsing, extensibility and minimal redundancy. The Gene Ontology and all other OBO ontologies are formalized as a Directed Acyclic Graphs (DAG), a structure with nodes (classes) and edges (relations). Every term has a definition given in natural language, not intended to be used by an automated reasoning tool to draw new inferences. The OBO-format is a very successful format for biomedical ontologies and it's used by most GO-based data analysis tools. The drawback is that ontologies in the OBO format typically lack computational definitions

to differentiate a term from other similar terms. This leaves the task of maintaining ontology integrity entirely on human developers because tools such as automated reasoners can't be used properly. The OBO-format is moving towards an increased expressiveness and the new version, OBO 1.2, even if it has not yet been adopted by most OBO ontologies, can support some reasoning but it still lacks OWL expressiveness and a DL reasoner. As pointed out in [51] once OBO ontologies are converted to OWL, they are available to a wider user community and they can make use of automatic reasoners, especially when logical statements such as necessary and sufficient definitions for classes are added. A mapping between OBO format files into appropriate OWL constructs and predefined annotation tags is being attempted by the National Center for Biomedical Ontologies (NCBO) [5] in a joint effort of OBO developers and ontology experts. For the USER ontology this mapping has been performed manually.

A.3 The USER-OBO Ontology

The USER-OBO ontology is structured as a directed acyclic graph (DAG), which is similar to a hierarchy, but differs in that a child term can have many parents, or less specialized, terms. New terms and their location within the ontology are proposed and then approved or rejected by an open group of individual on the web. The USER-OBO ontology has been developed and can be viewed using the editor software OBO-Edit [52]. Basic categories and relations in this format are necessary for the USER-OBO ontology to co-exist and co-operate with other OBO foundry ontologies. Ontological relations have not been chosen arbitrary: even if an "is-a" overloading can lead to overgeneralization or reduction of sense, the choice has been made according to the current standard use of biological ontolo-

[5] http://bioontology.org/

gies.

A.3.1 OBO terms

In the OBO format, the USER ontology consists of a controlled vocabulary of terms which represent classes, and a restricted set of relationships between these terms. Terms are organized in a hierarchy. Each entry in OBO ontologies consists of these anatomical parts:

Unique ID: a numerical identifier (e.g. USER: 0000003)

Term name: defined following name conventions (e.g. mRNA)

Synonyms: variant names that have the same meaning as the term (e.g. messenger-RNA). In the latest version of the OBO format there is the possibility to specify precision and coverage of a synonym selecting one option among: "related", "exact", "narrow" or "broad".

Textual definition: a human-readable and not computable definition that concisely states the biological meaning of the term (e.g. RNA molecule which contains the information ribosomes will use to produce a protein. It doesn't contain introns. It includes UTRs and coding sequences).

Database reference: reference of the definition: an organization, a book, a PubMed ID or some other source. (e.g. PMID17544019)

Parentage: computer readable parent-child relationships with other terms of the ontology (e.g. is-a: USER:0000002)

In order to reduce lexical confusion and render the ontology more computer-friendly, the terms of the USER ontology have been defined following some naming conventions, commonly shared by the other OBO ontologies:

- Term names are always singular, as they represent universals, not specific instances of concepts.
- Term names do not include spaces: underscores are used to separate the words in phrases (for example aminoacid-structure-motif).
- Numbers are spelled out in full (for example three-prime-UTR).
- Periods, points, slashes, hyphens and brackets are not allowed in term names.
- Common abbreviations, used in the molecular biology community, can be included in term names (for example snoRNA, RRM).

Synonyms are employed to record the variant names that have the same meaning as the term. Their usage facilitates the searching of the ontology. There is no limit to the number of synonyms a term can have, and they don't need to adhere to the previous naming conventions.

A.3.2 OBO relationships

While a controlled vocabulary is merely a collection of predefined terms that are used to describe the data, an ontology also formally specifies the relationships between its terms. This feature makes the data, labeled with the terms of an ontology, an admissible input for a software capable of logical inference. Currently the USER-OBO ontology uses three basic kinds of relationships between its terms: is-a, part-of and associated-with. The firs two relationships are defined in the OBO relationship types ontology, the last one is still a matter of discussion in the Sequence Ontology development but, providing a way to integrate heterogeneous data into the single context of post-transcriptional regulation of gene expression, it is

particularly significant in the light of the goals this ontology wants to pursue. The information introduced by this relation can be more efficiently conveyed using OWL-DL object properties.

is-a: it is a transitive, reflexive, anti-symmetric relationship. is-a is a simple class-subclass relationship, where A is-a B means that A is a subclass of B; for example, miRNA (microRNA) is-a ncRNA (non coding RNA). This relationship has a directional nature and establishes a sort of hierarchy among the terms of the ontology: inferences as to what something is proceed from the leaves towards the root of the ontology.

part-of : it is a transitive, reflexive, anti-symmetric relationship. This relationship belongs to mereology, the discipline dealing with parts and their respective wholes. C part-of D means that whenever C is present, it is always a part of D, but C does not always have to be present. For example a motif is part-of a biological-sequence: motifs are always part of a biological sequence, but not all biological sequences have motifs. Part-of relationships are not valid in both directions: every part-of relationship logically implies the inverse has-part relationship between the two terms.

associated-with: it is a symmetric relationship. It means being related to or accompanying, joined. Whenever there is strong evidence suggesting that the presence of a particular motif (a continuant) is highly correlated to the presence of one or more molecular functions (processes) in the cell, such as the bond to another molecule, then that function is associated with that motif, and vice versa. Any motif can also be associated with an effect on the behavior of the cell (for example the biological processes which are affected by the binding of a particular molecule to a binding motif, are associated with that mo-

tif, and vice versa). This relationship is not genuinely ontological, but dependent on experimental assumptions and methods [53].

A.3.3 Content description

The USER ontology presently contains more than one hundred terms. This section provides, with the help of illustrations and definitions, a general description of the shape, the organizational choices and the main contents of the USER ontology. Different portions of the ontology, each representing an important branch of biological annotation, will be presented and discussed: the path to the ontology root (the biological-sequence term) will always be present in the pictures, in order to let the reader orient himself.

Kinds of biological sequences: the two main kinds of biological sequence taken into consideration are transcripts (RNA molecules generated from DNA through the transcription process) and polypeptide chains (generated from mRNA through the translation process). DNA is not considered, as being not concerned in post-transcriptional regulations.

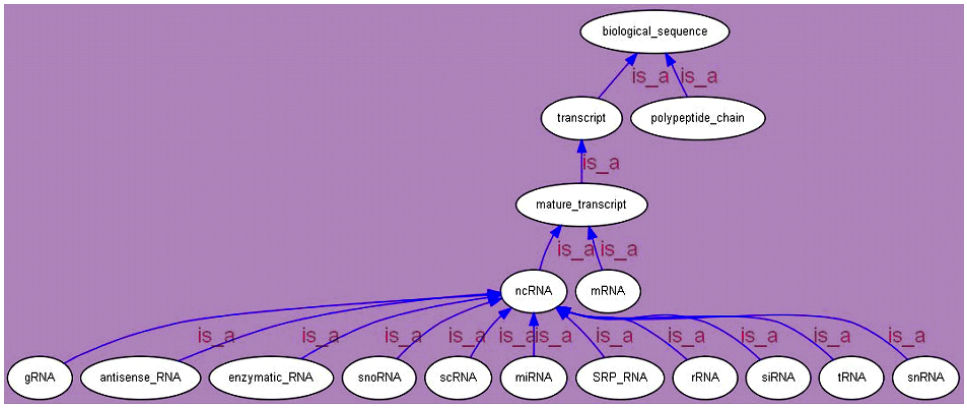


Figure A.2: A section of the USER Ontology describing different kinds of non-coding RNAs.

Kinds of aminoacid motifs: motifs can be defined as the functionally relevant parts of biological sequences. The aminoacid motifs selected

and considered in the USER ontology are the protein domains believed to be involved in post-transcriptional controls and thus dealing directly with RNA molecules. A rich classification of these domains can be found in literature [13], [54].

Kinds of RNA motifs: RNA motifs can be classified into two broad classes depending on the feature which is functional and evolutionary conserved: sequence or structure. RNA sequence motifs and their definitions are taken from the Transterm database [55], which provides access to mRNA sequences and regulatory elements (65 different motifs are contained in the latest Transterm version). Many of these motifs are primarily located in the 5' or 3' UTRs of mRNA sequences: they have been reported less commonly in coding sequences. RNA main structural motifs have been extracted from the SCOR database [6], the NDB database [7] and the RNA Ontology Consortium website [8].

Motif functions: by definition every biological motif can be joined to specific molecular functions, which establishes the biological role of the motif and why it has been selected and conserved by evolution. As the ontology is centered on post-transcriptional mechanisms concerning mRNA, binding functions are primarily considered. The three major binding classes, RNA-binding, small-molecule-binding or protein-binding, specify the biological molecule with which the binding motif interacts. This portion of the USER Ontology is integrabile with the Molecular Function tree in Gene Ontology.

Post-transcriptional effect : motifs in certain mRNAs have been shown to have a positive or negative effect on many functions and post-translational controls in cells, described in this section of the ontology.

[6] http://scor.lbl.gov/

[7] http://ndbserver.rutgers.edu/

[8] http://roc.bgsu.edu/

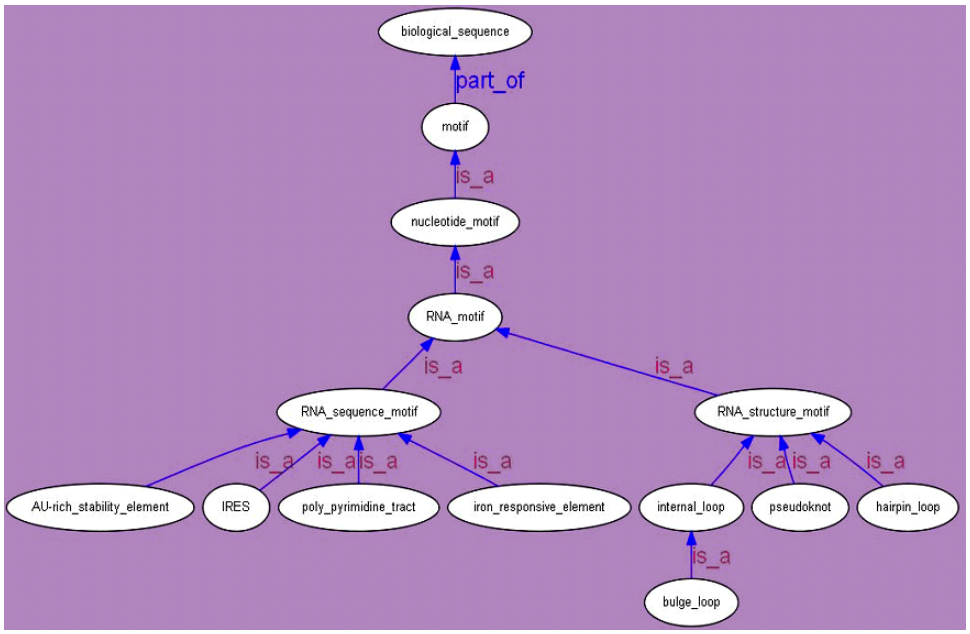


Figure A.3: A section of the USER Ontology describing the kinds of motifs which can be traced in RNA sequences.

A.4 The USER-OWL Ontology

The USER ontology has been first built in OBO format using the OBO-Edit software, and then it has been recast in OWL-DL language using the Protg Software [9]. OWL-DL is a sub-language of OWL (Web Ontology Language) [10] characterized by computational tractability and an expressiveness which falls between that of OWL-Lite and OWL-Full and is based on Description Logics. OWL-DL allows a Description Logic Reasoner to check the consistency of the ontology and automatically compute the ontology class hierarchy. OWL ontologies consist of Individuals, Properties and Classes. Individuals in the USER ontology are principally RNA sequences corresponding to real transcripts (any biological species can be considered). Classes are sets which contain individuals, described when possible using formal descriptions that state precisely the requirements for

[9] `http://protege.stanford.edu/`
[10] `http://www.w3.org/TR/owl-features/`

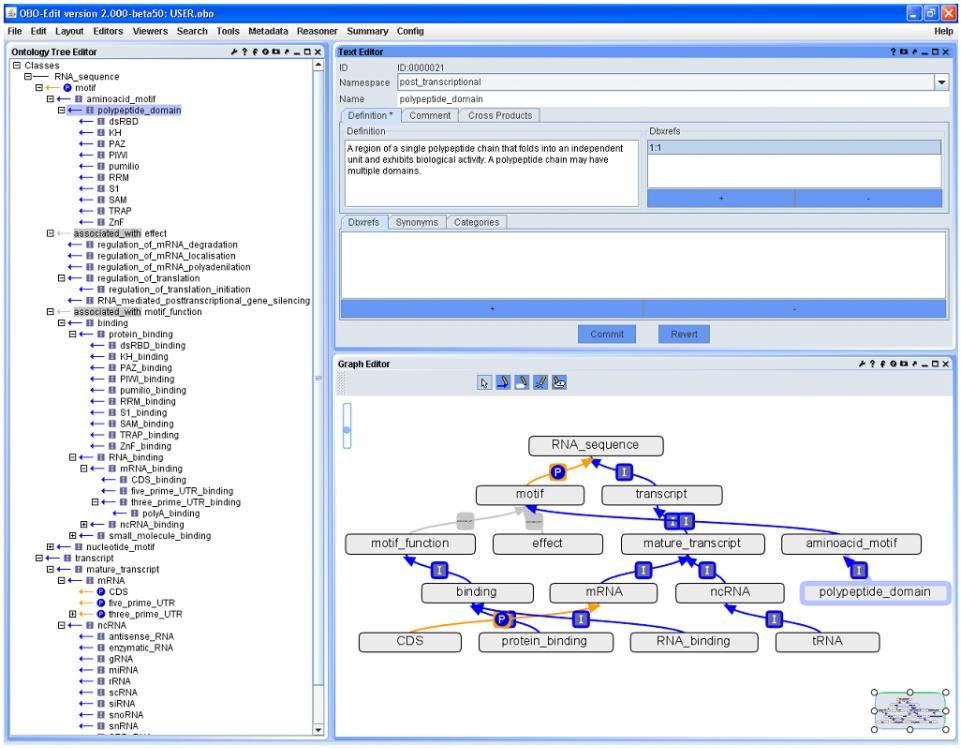


Figure A.4: Screenshot taken from the OboEdit software showing an overall representation of the USER-OBO ontology.

membership of the class. Classes in the USER-OWL ontology correspond to OBO terms in the USER-OBO ontology: class names are thus directly created from OBO term names, following the same name conventions. Also the superclass-subclass hierarchical structure follows the same is-a structure found in the USER-OBO ontology, while part-of and associated-with relations have been translated in different OWL properties. The following sections introduce and discuss the new features introduced in the USER ontology during the OWL-DL conversion, thanks to the increased expressivity of this language.

A.4.1 Disjointness and covering constraints

While in the OBO format if a parent class has more than one child there is no way to distinguish between possibly overlapping and disjoint classes,

OWL-DL allows this specification. In the USER-OWL ontology this specification has been added every time a supporting reliable biological knowledge is present. For example mRNA and ncRNA are disjoint classes of RNA sequences because an instance of mature-transcript can't belong to both these two classes. On the contrary, inside the ncRNA class, miRNA and siRNA haven't been specified as disjoint because the real boundary between these two categories of functional RNAs hasn't been outlined yet, both at the conceptual level and at the instance levels. This means that some individuals belonging to both these classes could exist. For the same reason sequence-RNA-motif and structure-RNA-motif are not disjoint classes. Covering constraints have been added to the USER-OWL ontology where appropriate. If we have three classes A, B and C and classes B and C are subclasses of class A, a covering axiom that specifies that class A is covered by class B and by class C means that a member of class A must be a member of B and-or C (class A is the union of the classes being covered). The OBO representation doesn't allow these axioms, and their use has been limited also in the USER -OWL ontology because such axioms require more knowledge than is usually available in biology. For example we can't presently be sure that every ncRNA belongs to one of the subclasses specified in the USER ontology since other subclasses are probably going to be discovered in the future, and therefore a covering constraint has not been added in this case. On the contrary we can say that a mature-transcript must be either a mRNA or a ncRNA, therefore a covering axiom relating these three classes has been added.

A.4.2 Object properties

Object properties represent binary relationships between individuals. OWL allows the meaning of properties to be enriched through the use of property characteristics such as functional, inverse functional, transitive or symmet-

ric. In OWL properties may also have sub-properties, therefore it's possible to create hierarchies of properties. Properties may have a domain and a range specified: a property links individuals from its domain to individuals from its range. These are the object properties in the USER-OWL ontology; they have been named following OWL property naming conventions. Domain and range of each property has been specified. Each of these object properties has a corresponding inverse property, with swapped domain and range.

hasMotif: links an individual belonging to the class RNA-sequence or RNA-part to an individual belonging to the class motif. Inverse property: isMotifOf.

hasPart: links an individual belonging to the class mRNA to an individual belonging to the class RNA-part. This property has three sub-properties: hasCDS, hasThreePrimeUTR and hasFivePrimeUTR, which link an individual belonging to the class mRNA to an individual belonging respectively to the class CDS, the class three-prime-UTR and the class five- prime-UTR. These three subproperties are functional, since every mRNA sequence has one and only one CDS, three prime UTR and five prime UTR. Inverse property: isPartOf, isCDSOf, isThreePrimeUTROf, isFivePrimeUTROf.

hasFunction: links an individual belonging to the class motif to an individual belonging to the class function. Inverse property: isFunctionOf.

hasEffect: links an individual belonging to the class motif to an individual belonging to the class effect. Inverse property: isEffectOf.

A.4.3 Property restrictions

In OWL properties are used to create restrictions, which help to define classes in a computer understandable way. Restrictions are used to restrict the individuals that belong to a class. Quantifier restrictions are composed of a quantifier (commonly the existential quantifier or the universal quantifier) a property and a filler (usually a class or a composition of classes). Existential restrictions have been added to the USER-OWL ontology: for example an existential restriction has been added to the class motif specifying that it must be a motif of some RNA-sequence (along the isMotifOf property). These are other examples of existential restrictions used in the USER-OWL ontology:

- Class function must be a function of some motif (along the isFunctionOf property)
- Class effect must be an effect of some motif (along the isEffectOf property)
- Class aminoacid-motif must be a motif of some CDS (along the isMotifOf property)

Universal restrictions are more difficult to assign because they require a more precise underlying biological knowledge, with respect to existential restrictions. These are examples of universal restrictions used in the USER-OWL ontology:

- Class RNA-sequence has only motifs which belong to the motif class (along the hasMotif property)
- Class motif has only functions which belong to the function class (along the hasFunction property)

- Class CDS is only part of individuals belonging to the mRNA class (along the isCDSOf property)

These restrictions have been inferred from the textual definitions present in the USER-OBO ontology and have been added as necessary conditions to primitive classes in the ontology. They represent a way in which human-readable knowledge can be to some extent converted into computer-readable knowledge. The hope is that, as biological knowledge grows and becomes more accurate, more information can be poured into formal restrictions.

A.4.4 Defined classes

Necessary and sufficient conditions have not been used in the original OBO-derived USER ontology classes. Some defined classes have been created to facilitate the extraction of biologically meaningful information from the ontology and to help the creation of restrictions which define other classes. For example it is very useful to define a subclass of motif called mRNA-binding-motif, whose members are all motifs which have at least an mRNA-binding function. This condition is necessary and sufficient to define the RNA-binding-motif class and converts this from a Primitive Class into a Defined Class, on which the DL-reasoner can perform automatic classification. The defined class mRNA-binding-motif is useful also as filler in restrictions which describe other classes: for example an individual belonging to the class miRNA or siRNA must have some mRNA-binding-motif (existential restriction along the hasMotif property). Another biologically interesting defined class is composed of mRNA which can be regulated at the same time by microRNAs and by RNA binding proteins. These mRNA must possess at least one or more miRNA-binding-motif and at least one protein-binding-motif: we can call this class mixture-regulated-

mRNA. This class is defined as a subclass of mRNA having two existential restrictions acting along the hasMotif property.

A.4.5 Annotation properties

Natural language definitions associated with terms in the USER-OBO ontology cannot be translated directly in OWL-DL axioms. However it is possible to capture them using annotation properties: OWL allows classes, properties and individuals to be annotated with various species of information, for example comments or references to other resources. Assertions on annotation properties act as comments and are not taken into account from a DL point of view, yet they can be displayed to the biologist as a piece of information on classes, just as in OBO ontologies. The most suitable annotation property for labeling a term with its id is rdfs:label, while the most suitable annotation property for labeling a term with its textual definition is rdfs:comment. The annotation rdfs:seeAlso can be used to identify related resources. Synonyms from the OBO format can be traduced using assertions on annotation properties or creating equivalent classes.

A.4.6 Populating classes with individuals

Populating USER-OWL classes with real biological entities implies the annotation of biological RNA sequences. This work is being accomplished for the human genome using several bioinformatics resources such as the already seen Transterm [55] or the Ensembl Genome Browser [56]. At the end of this process the RNA-sequence class will contain all the human transcribed sequences currently known. Every individual is currently identified by a unique ENSEMBL transcript ID (e.g. ENST00000229384). The following tab summaries the number of individuals populating some of the

mature-transcript subclasses of the USER-OWL ontology:

Class	**Individuals**
mRNA	21528
miRNA	1472
rRNA	333
snoRNA	758
snRNA	1288

Table A.1: Number of individuals, retrieved from the Ensembl Genome browser, belonging to different classes of mRNAs and ncRNAs.

The next step is the creation of relationships between individuals, according to the object properties of the USER-OWL ontology. Once the ontology classes have been populated with individuals and relationships, hasValue restrictions can be used in class descriptions. For example there is the possibility to create classes of all the motifs associated to any individual mRNA sequence along the property isMotifOf.

A.4.7 Use of the RACER reasoner

Ontologies described using OWL-DL can be processed by a reasoner, one of whose main services is consistency checking: based on the conditions of a class the reasoned can check whether or not it is possible for the class to have any instances: a class is considered inconsistent if it cannot possibly have any instances. The DIG compliant reasoner RACER [57] has been applied to check the USER-OWL ontology consistency: corrections have been made in order to make the current version of the ontology free of any inconsistency. Another standard service offered by the reasoner is to test whether or not one class is a subclass of another class, relying upon class definitions: by performing such tests on all the classes of the ontology the reasoner can automatically compute the inferred ontology class hierarchy.

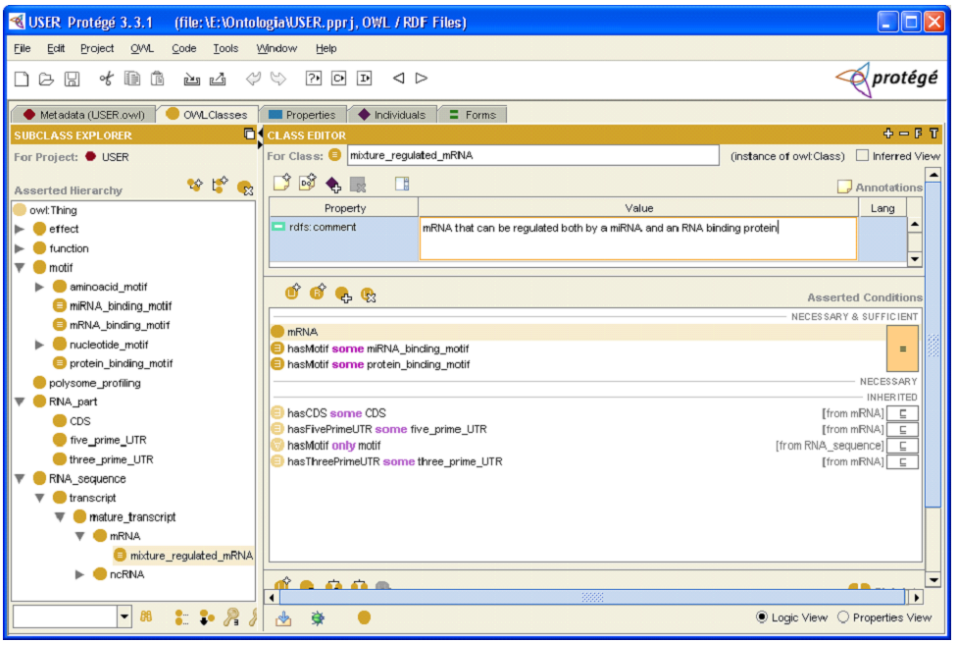


Figure A.5: Screenshot taken from the Protg software showing necessary and sufficient conditions for the class mixture-regulated-mRNA.

This inferred hierarchy can be compared to the already existing asserted hierarchy. The RACER reasoner has been applied to the USER-OWL ontology to perform automatic classification.

Appendix B

Bayesian inference of RBP-mRNA interactions

Translational controls of gene expression are strongly mediated by RNA-binding proteins (RBPs), therefore change over time among different conditions in polysomal mRNA profiles should be mirrored by changes in concentrations of effector RBPs.

We first decided to consider a linear model relating changes in RPB polysomal levels to the difference between changes in polysomal RNA concentrations and changes in total RNA concentrations. We assume also that RBP protein levels run parallel to changes in RBP polysomal concentrations. The goal of the analysis is inferring the network structure, i.e. the interactions among RBPs and mRNAs.

$$x_t^i(t) - x_p^i(t) = \sum_j A_j^i y_p^j(t) + \epsilon \tag{B.1}$$

We considered also the possibility of a linear model relating directly changes in RBP polysomal levels to changes in polysomal RNA concentrations, without considering total RNA concentrations. In this way the model turns into:

$$x_p^i(t) = \sum_j A_j^i y_p^j(t) + \epsilon \tag{B.2}$$

Another proposal is to use as input parameter the quantitative uncoupling values determined after principal component analysis on polysomal and total fold change values, as seen in chapter 3

$$PC2_p^i(t) = \sum_j A_j^i y_p^j(t) + \epsilon \tag{B.3}$$

The last option is to use as independent variable the ratio between sub-polysomal levels and polysomal levels (subtracting polysomal levels from total levels in all the datasets where the sub-polysomal signal is not available).

$$x_p^i(t) - x_s^i(t) = \sum_j A_j^i y_p^j(t) + \epsilon \tag{B.4}$$

The outcome of this work will be the identification of putative actions of given RBPs on translational efficiency modifications of given mRNAs, detected by comparison between translatome and transcriptome profiling techniques. It's quite relevant to note that the inferred action of RBPs on mRNAs is not dependent on their direct phisical interactions during translation. This statement justifies the observed degree of discrepancy between the model outcomes and some RIP-chip experimental results. Following an accepted model of translational regulation which considers translation initiation the fundamental rate limiting step of protein synthesis, these RBPs should be able to produce uncoupling between transcriptome and polysomal mRNA variations by mainly affecting their translational initiation efficiency. Described molecular mechanisms include: subtraction of mRNA availability for the formation of the closed loop for polysomal mRNA buffering and increased efficiency of assembly of pre-initiation complex, or 5'UTR scanning for polysomal mRNA magnification of steady state

variations. Note that in both cases the proteins involved as effectors are not necessarily RBPs (e.g. eIF4G). In these cases the expected effect should be less relying on phisical protein-RNA interactions (for example TCD4 in mammals is a repressor of eIF4E helicase, which on turn entangles mRNAs 5'UTRs increasing translational initiation efficiency).

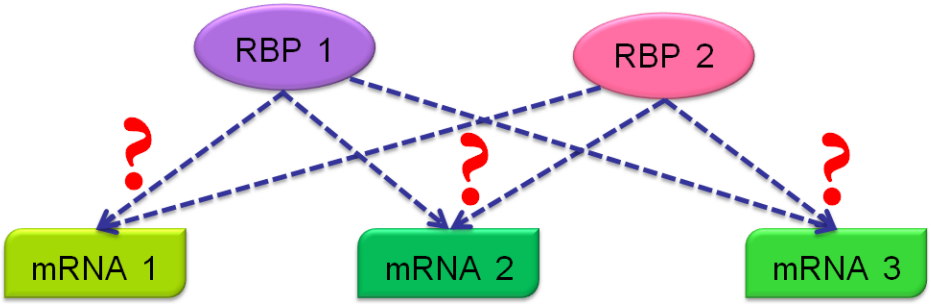


Figure B.1: This scheme represents the problem of inferring specific connections between mRNAs and RNA binding proteins able to regulate, directly or indirectly, their translation efficiency. Inference on post-transcriptional networks is based on high-throughput variables, such as polysomal profiling signals coming from the experiments described in chapter 3.

B.1 Clustering of RBPs

Since observations (polysomal data coming from different experiments) are less than the number of RBPs (a few hundreds), an idea is to cluster the RBP profiles. The updated model would be the following:

$$x_t^i(t) - x_p^i(t) = \sum_j A_j^i \mathbf{c}_p^j(t) + \epsilon \tag{B.5}$$

$$y_p^i(t) \sim \sum_{i=1}^{K} \pi^i \mathcal{N}(\mathbf{c}^i(t), \sigma_i^2) \tag{B.6}$$

Possible RBP clustering criteria are:

- a priori clustering using sequence homology information
- a posteriory clustering using behaviour similarity in polysomal profiles

Proteins to be included as regulators of translational efficiencies could include not only RNA binding proteins, but also more general translational modulators, such as translation initiation factors or proteins involved in pathways known to influence post transcriptional regulation such as the mTOR pathway.

B.2 Graphical model

Figure B.2 represents the graphical model of the Bayesian network adopted to extract from polysomal data putative RBP-mRNA connections.

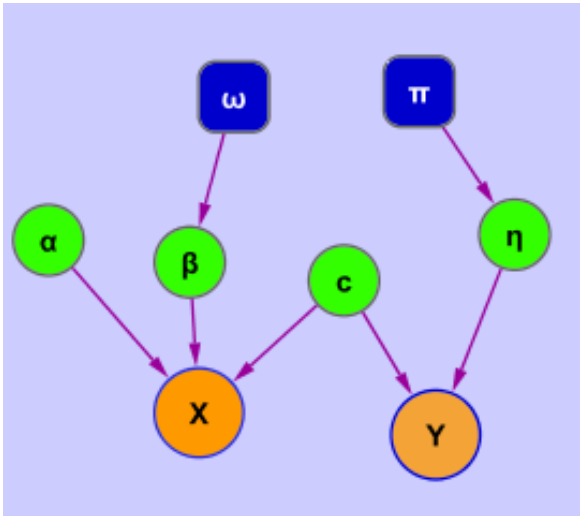


Figure B.2: Bayesian network used to infer relationships between RBPs and mRNAs

We assume that the entries in the matrix A_j^i are of product form $\alpha_j^i \beta_j^i$ where α_j^i is normally distributed with mean 0 and variance 1, while β_j^i is 0 or 1 with constant probability ω. This last parameter controls the expected parsity of the network.

The joint probability, according to the Bayesian network used to model post-transcriptional controls, can be expressed as:

$$p(x, y, \alpha, \beta, c, \eta) = p(x|\alpha, \beta, c) \cdot p(y|c, \eta) \cdot p(\alpha) \cdot p(\beta) \cdot p(c) \cdot p(\eta) \quad \text{(B.7)}$$

A simplified version of the model doesn't consider the clustering step: the simplified bayesian network is represented in figure B.3.

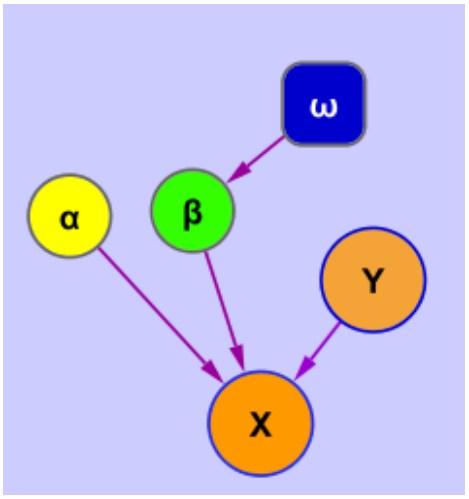


Figure B.3: Simplified Bayesian network used to infer relationships between RBPs and mRNAs without clustering RBPsaccording to their similar behaviour in polysomal profiling experiments.

The joint probability of the model without clustering is:

$$p(x, y, \alpha, \beta) = p(x|\alpha, \beta, y) \cdot p(y) \cdot p(\alpha) \cdot p(\beta) \quad \text{(B.8)}$$

B.3 Gibbs sampler

A Gibbs sampler can be used to estimate the joint posterior probability over the network structure, the latent RBP primitive profiles and the latent RBP cluster allocations. For the Gibbs sampler the conditional posterior over all the unobserved variables in the model are needed:$\alpha_j^i, \beta_j^i, \mathbf{c}, \eta$.

$$p(\alpha_j^i|\beta_j^i, \mathbf{c}^j, x^i, A_{-ij}, \Theta) \propto p(x^i|\alpha_j^i, \beta_j^i = 1, A_{-ij}, \mathbf{c}^j, \Theta)\, p(\alpha_j^i) \tag{B.9}$$

$$p(\beta_j^i = 1|\alpha_j^i, \mathbf{c}^j, x^i, A_{-ij}, \Theta) \propto p(x^i|\alpha_j^i, \beta_j^i = 1, A_{-ij}, \mathbf{c}^j, \Theta)\, p(\beta_j^i = 1) \tag{B.10}$$

$$p(\eta|x, y, A, c) \propto p(\eta|y, c) \tag{B.11}$$

$$p(\mathbf{c}^j|x, y, A, \eta, \Theta) = p(\mathbf{c}^j)\, p(x|A, \mathbf{c})\, p(y|\mathbf{c}, \eta, \Theta) \tag{B.12}$$

The equation describing how x data are modelled from y data is:

$$x_i = \sum_j \alpha_j^i \beta_j^i y_j + \varepsilon \tag{B.13}$$

The following equation describes how the β parameters in the Gibbs sampler are updated at each iteration:

$$p(\beta_j^i = 1|\alpha_j^i, \mathbf{c}^j, x^i, A_{-ij}, \Theta) \propto p(x^i|\alpha_j^i, \beta_j^i = 1, A_{-ij}, \mathbf{c}^j, \Theta)\, p(\beta_j^i = 1) \tag{B.14}$$

$$p(x^i|\alpha_j^i, \beta_j^i = 1, A_{-ij}, \mathbf{c}^j, \Theta) \sim N(\sum_j \alpha_j^i \beta_j^i c_j, \varepsilon) = \frac{1}{\sqrt{2\pi\varepsilon}} \exp\left\{-\frac{1}{2\varepsilon}(x_i - \sum_j \alpha_j^i \beta_j^i c_j)^2\right\} \tag{B.15}$$

The following equation describes how the α parameters in the Gibbs sampler are updated at each iteration:

$$p(\alpha_j^i|\beta_j^i, \mathbf{c}^j, x^i, A_{-ij}, \Theta) \sim N(m, \xi) \tag{B.16}$$

$$\frac{1}{\xi^2} = \frac{c_j^2}{\varepsilon} + 1 \tag{B.17}$$

$$m = \frac{c_j}{\varepsilon} \cdot (x_i - \sum_{\hat{j} \neq j} \alpha_{\hat{j}}^i \beta_{\hat{j}}^i c_{\hat{j}}) \cdot \xi^2 \tag{B.18}$$

Update of η parameters in the Gibbs sampler is described by the following equation:

$$p(\eta_l | y, \mathbf{c}_l) = \frac{\pi_l \cdot exp\left\{-\frac{1}{2}(y - \mathbf{c}_l)^2\right\}}{\sum_{k=1}^K \pi_k \cdot exp\left\{-\frac{1}{2}(y - \mathbf{c}_k)^2\right\}} \tag{B.19}$$

Update of $\mathbf{c}$ parameters in the Gibbs sampler:

$$p(\mathbf{c}^i | x, y, A, \eta, \Theta) \sim N(m, \xi) \tag{B.20}$$

$$\frac{1}{\xi^2} = \Sigma_j \frac{(\alpha_j^i \beta_j^i)^2}{\sigma^2} + \sum_l \eta_{li} + 1 \tag{B.21}$$

$$m = \left(\frac{\Sigma_j (x_j - \Sigma_{\hat{i}} \alpha_{\hat{i}}^j \beta_{\hat{i}}^j \mathbf{c}_{\hat{i}}) \cdot \alpha_j^i \beta_j^i}{\sigma^2} + \sum_l y_l \cdot \eta_{li} \right) \cdot \xi^2 \tag{B.22}$$

B.4 Data structures

This section describes the data structures used to implement computationally the Gibbs sampler.

$$\alpha \begin{bmatrix} \alpha_1^1 & . & . & \alpha_1^C \\ . & & & . \\ . & & & . \\ \alpha_T^1 & . & . & \alpha_T^C \end{bmatrix} \begin{array}{l} \downarrow number\ of\ transcripts \\ \\ \\ \\ \rightarrow number\ of\ clusters \end{array} \tag{B.23}$$

$$\beta \begin{bmatrix} \beta_1^1 & . & . & \beta_1^C \\ . & & & . \\ . & & & . \\ \beta_T^1 & . & . & \beta_T^C \end{bmatrix} \begin{array}{l} \downarrow number\,of\,transcripts \\ \\ \\ \rightarrow number\,of\,clusters \end{array} \tag{B.24}$$

$$\mathbf{X} \begin{bmatrix} x_1^1 & . & . & x_1^O \\ . & & & . \\ . & & & . \\ x_T^1 & . & . & x_T^O \end{bmatrix} \begin{array}{l} \downarrow number\,of\,transcripts \\ \\ \\ \rightarrow number\,of\,observations \end{array} \tag{B.25}$$

$$\mathbf{Y} \begin{bmatrix} y_1^1 & . & . & y_1^O \\ . & & & . \\ . & & & . \\ y_R^1 & . & . & y_R^O \end{bmatrix} \begin{array}{l} \downarrow number\,of\,RBPs \\ \\ \\ \rightarrow number\,of\,observations \end{array} \tag{B.26}$$

$$\mathbf{c} \begin{bmatrix} c_1^1 & . & . & c_1^O \\ . & & & . \\ . & & & . \\ c_C^1 & . & . & c_C^O \end{bmatrix} \begin{array}{l} \downarrow number\,of\,clusters \\ \\ \\ \rightarrow number\,of\,observations \end{array} \tag{B.27}$$

$$\eta \begin{bmatrix} \eta_1^p \\ . \\ . \\ \eta_R^p \end{bmatrix} \begin{array}{l} \downarrow number\,of\,RBPs \\ \\ \\ where\,\eta^p \in \{\eta^1 ... \eta^C\} \end{array} \tag{B.28}$$

B.5 Algorithm implementation with synthetic data

In order to test the learning capability of the implemented model, synthetic data were generated as:

- synthetic α(*Alfa_real*) are generated sampling from a normal distribution with mean=0 and standard deviation=1
- synthetic β(*Beta_real*) are generated placing 1 values for all β corresponding to α absolute values above a certain quantile (modulated by the sparseness parameter ω), and placing 0 values for all other β.
- synthetic c(*Cluster_real*) are generated sampling from a mixture of two normal distributions with mean=$\pm m$ and standard deviation=sd (where m and sd are user-defined parameters)
- synthetic η(*Eta_real*) are generated with equal probability for every RNA binding protein to belong to any cluster
- synthetic polysomal fold changes y(*Y_real*) are generated sampling from normal distributions with mean equal to the values of the cluster the RNA binding protein belongs to, and standard deviation=sdy (a user defined parameter)

$$y_r^o \sim N(\eta_r^p \cdot \mathbf{c}^{\mathrm{o}}, \mathrm{sdy}) \tag{B.29}$$

- synthetic delta fold changes of transcripts x(*X_real*) are generated using *Cluster_real*, *Alfa_real* and *Beta_real* data and adding a certain percentage of error, defined by the *errorpar* parameter, on the mean variance on delta fold change observations for every gene
- initial c and η data (*Cluster_start*) and (*Eta_start*) are generated applying the k-means clustering algorithm on *Y_real* data (the number of clusters is defined by the user)

- initial $\alpha$and β data (*Alfa_start*) and (*Beta_start*) are created by regression on *X_real* and *Cluster_start* data.

The algorithm shows a good learning performance on synthetic data, as testified in figure B.4 by the ROC curve obtained on the β parameter, which models the probability of having interaction between RBPs and mRNAs.

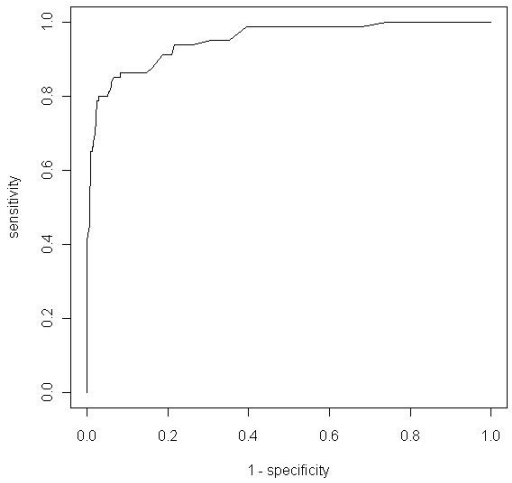


Figure B.4: ROC curve associated to the β parameter threshold chosen as significative to infer an interaction between RBPs and mRNAs

B.6 Algorithm sperimentation with yeast data

Polysomal yeast data were retrieved from literature and microarray databases (GEO [58], ArrayExpress [59]). Following the procedure already described in chapter 3,raw data were normalized with the RMA method (Robust Multichip Average) [60]. Fold changes relative to yeast open reading frames were calculated for total and polysomal RNA profiling. A brief description of the datasets included in the analysis follows:

1. Transcript-specific translational regulation in the unfolded protein response of Saccharomyces cerevisiae [61]. Cells are treated with DTT,

polysomal and sub-polysomal RNA fractions are extracted and quantified with Affymetrix genechips before the treatment and after 1 hour.

2. Gene expression analyzed by high-resolution state array analysis and quantitative proteomics: response of yeast to mating pheromone [62]. Cells are treated with αfactor, polysomal and total RNA fractions are extracted and quantified with custom yeast ORF microarrays before the treatment and after 30 minutes.

3. Translation profiling of yeast caf20 mutants. Wild type cells are compared with mutant cells, polysomal and total RNA fractions are extracted and quantified with Affymetrix genechips.

4. Global Translational Responses to Oxidative Stress Impact upon Multiple Levels of Protein Synthesis [63]. Cells are treated with H_2O_2, polysomal and total mRNA fractions are extracted and quantified with Affymetrix genechips before the treatment and after 15 minutes.

5. Global gene expression profiling reveals widespread yet distinctive translational responses to different eukaryotic translation initiation factor 2B-targeting stress pathways [64]. Cells are aminoacid starvated, polysomal and total mRNA fractions are extracted and quantified with Affymetrix genechips before the treatment and after 20 minutes.

6. Global gene expression profiling reveals widespread yet distinctive translational responses to different eukaryotic translation initiation factor 2B-targeting stress pathways [64]. Cells are treated with butanol, polysomal and total mRNA fractions are extracted and quantified with Affymetrix genechips before the treatment and after 10 minutes.

Appendix C

The mRNA relay model of gene expression

After 40 minutes of EGF stimulation the majority of variations in gene expression (DEGs) are seen only at the polysomal level (72.4%), and the degree of overlapping between translatome DEGs and transcriptome DEGs is very small (3.2%). Despite this, the over-represented GO themes are largely overlapping (18.4% identity overlap). This is really weird. Translatome reprogramming is profoundly rewiring the transcriptome program in terms of the specific DEGs, but not in terms of potential final phenotypic outcomes, it is a sort of "fake reprogramming". It is something like the nucleus is delivering the exact final message, but this message, before arriving at destination, has to change in shape without affecting the content. This analysis suggests the existence of fake reprogramming.

From the 4E-T silencing experiment (4E-T is necessary for P-bodies formation) we see that the "fake reprogramming" activity is disrupted by the loss of P-bodies, because it induces a marked decrease of translatome DEGs. Consistently with fake reprogramming, there is no change in over-represented GO themes after inhibition of P-body formation. This experiment therefore confirms fake reprogramming and locate it in P-bodies. Another confirmation comes from the inconsistency of results we observe

between polysomal and subpolysomal signals of the same genes. In our analysis, these two DEGs profiles are not in favour of an obvious model of transfer of mRNAs from the subpolysomal compartment to the polysomal. But if we assume that fake reprogramming takes place and that the newly transcribed mRNAs after EGF stimulation are targeted to P-bodies and subsequently degraded, we should not expect direct transfer to the polysomal compartment. This analysis confirms fake reprogramming suggesting that it involves a step of degradation (possibly in P-bodies) of the newly synthetized mRNAs.

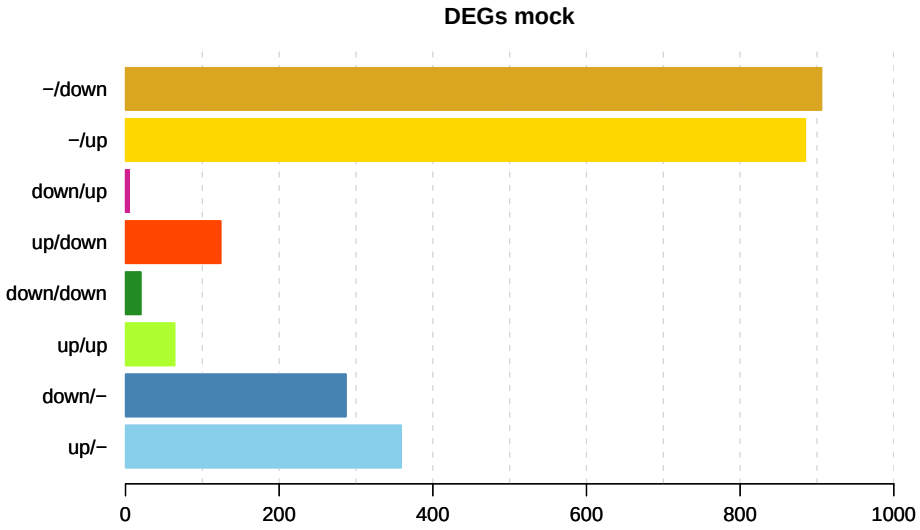


Figure C.1: Number of DEGs for the different classes in the mock experiment

C.1 A model for the mechanism of fake reprogramming

In this model positive variations of mRNA quantities are all "absorbed" by P-bodies (we do not say anything here for simplicity about negative

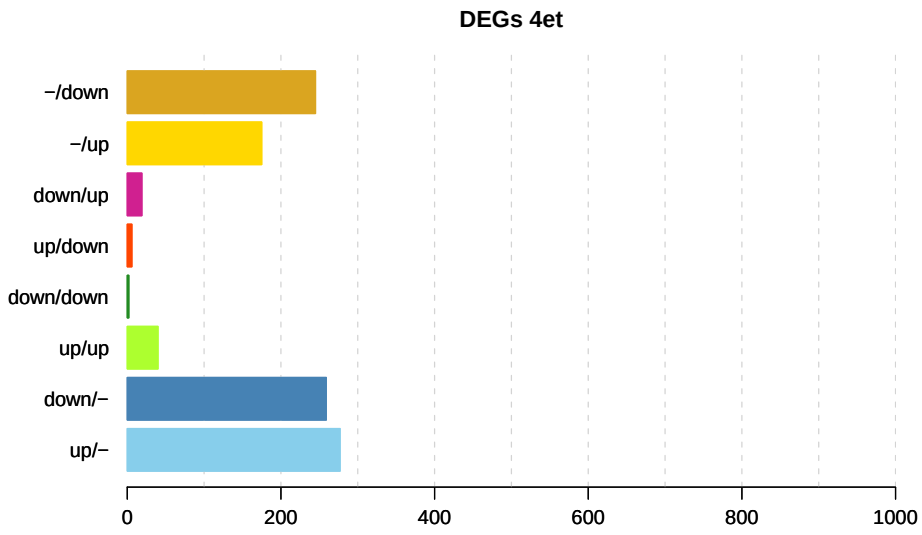


Figure C.2: Number of DEGs for the different classes in the 4-ET silencing experiment

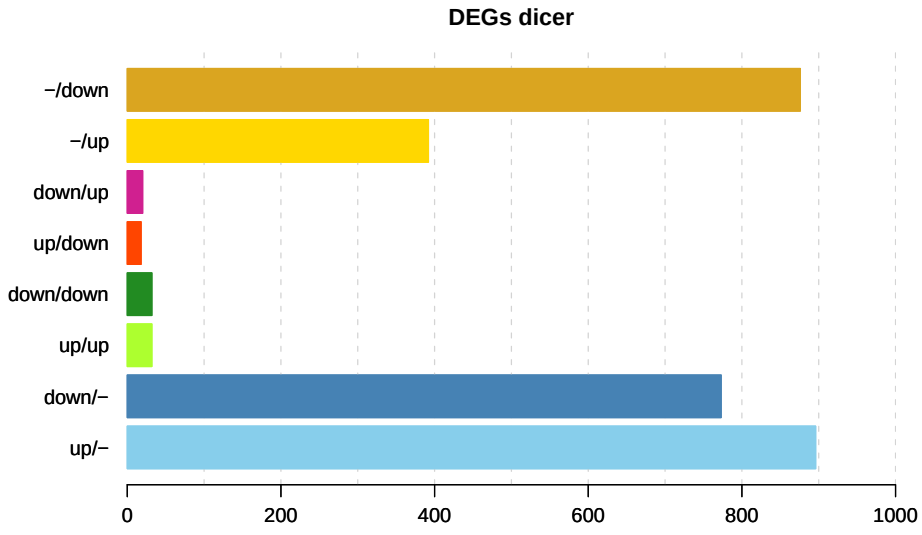


Figure C.3: Number of DEGs for the different classes in the Dicer silencing experiment

variations, which should inserted later in the model). This implies that every incremental mRNA species induced in the nucleus by EGF signaling is

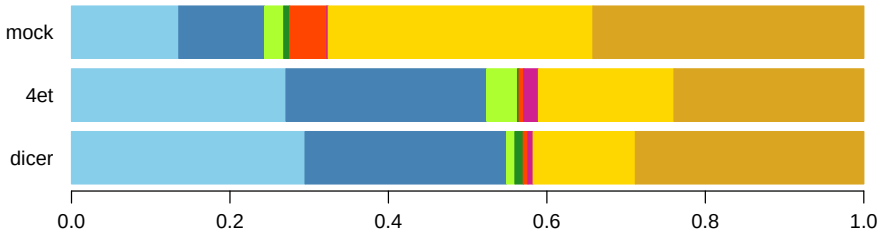


Figure C.4: Percentage of DEGs for the different classes

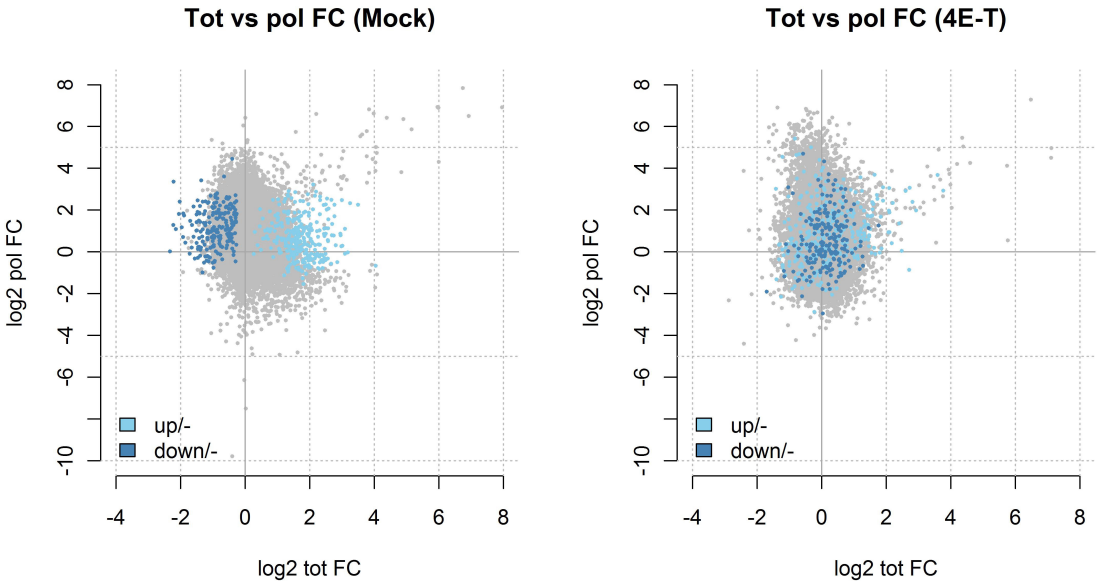


Figure C.5: Movements of mock transcriptome DEGs upon 4E-T silencing

directed in the cytoplasm to P-bodies. In P-bodies a completely unknown mechanism determines the "recognition" by each entered mRNA of other mRNAs already stored in P-bodies which are ontologically compatible with the first (i.e., whose protein products, altogether, perform a similar function), finally determining their release toward the polysome. The "input" mRNA in the P-bodies is instead kept there or, most likely, degraded.

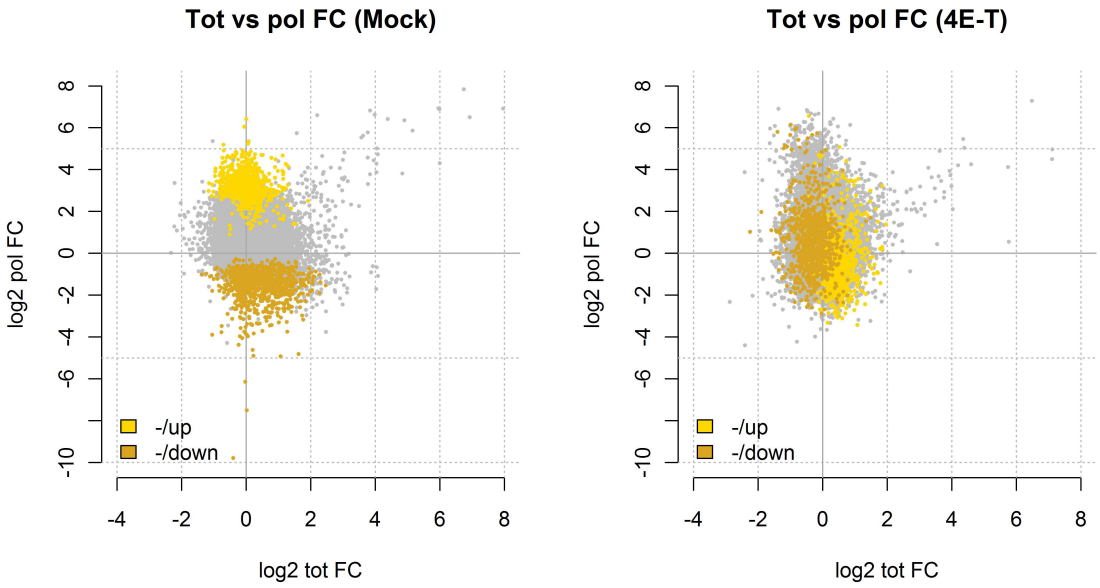


Figure C.6: Movements of mock translatome DEGs upon 4E-T silencing

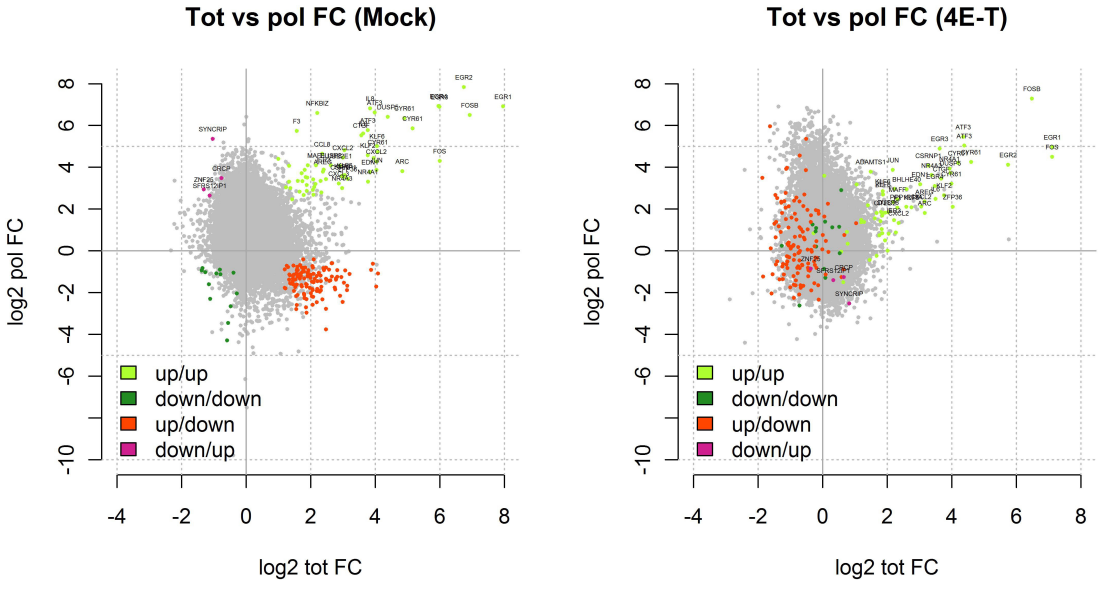


Figure C.7: Movements of mock common DEGs upon 4E-T silencing

1. What is the code for ontological compatibility

2. And how could be mRNA-mRNA recognition realized?

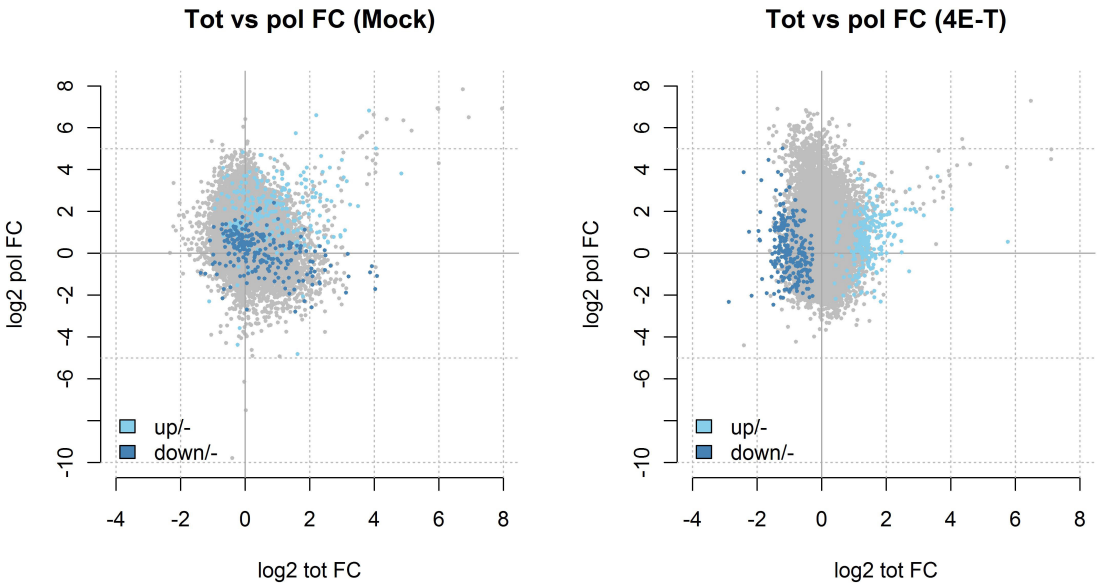


Figure C.8: Movements of 4E-T transcriptome DEGs upon 4E-T silencing

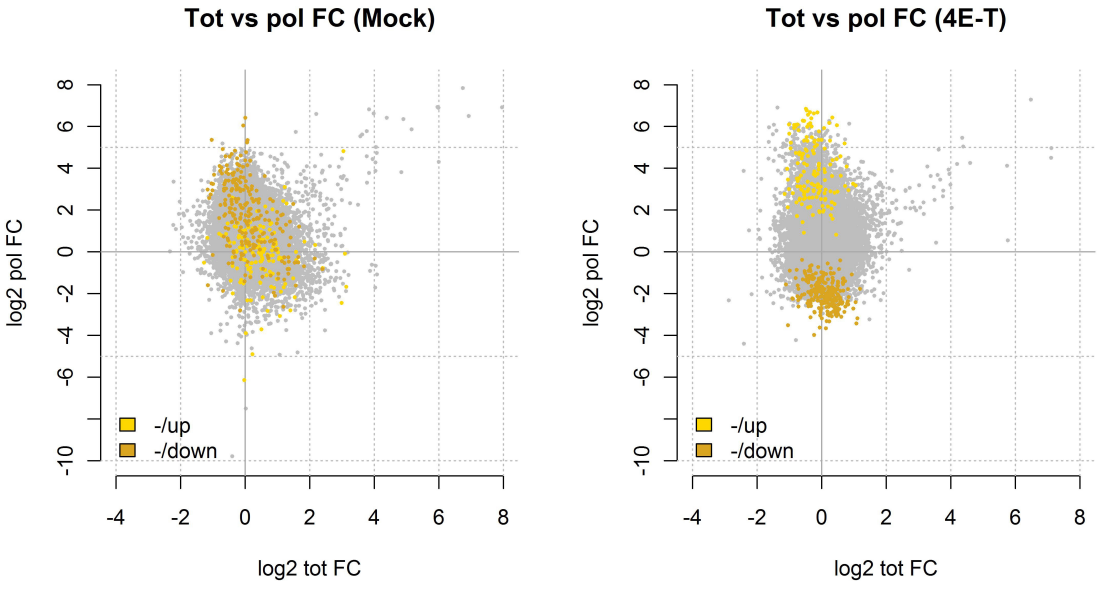


Figure C.9: Movements of 4E-T translatome DEGs upon 4E-T silencing

(1) There are various possibilities, but two hypotheses are the more likely.

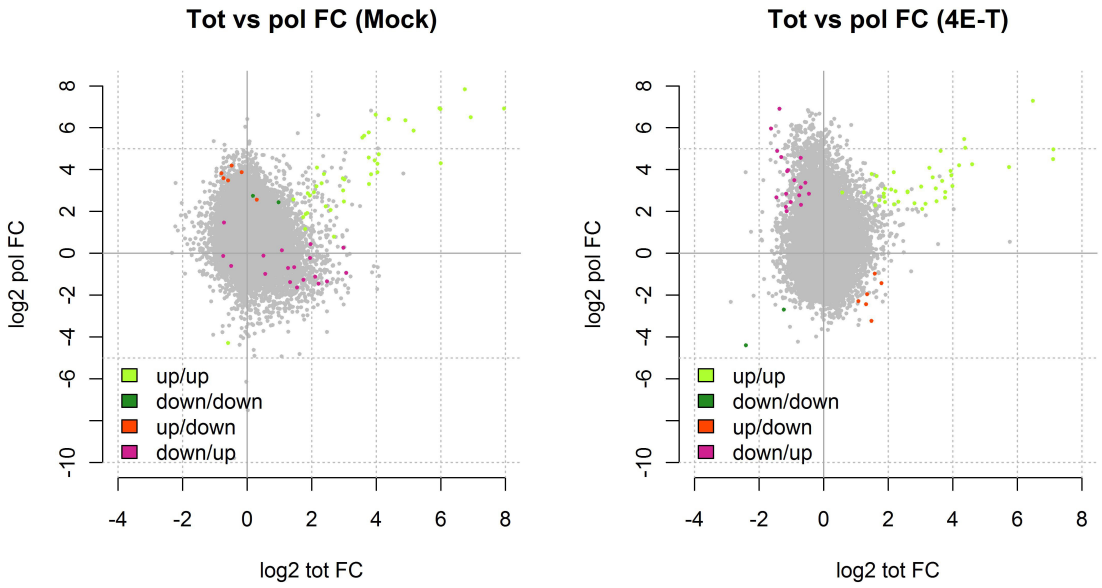


Figure C.10: Movements of 4E-T common DEGs upon 4E-T silencing

1a The code is based on CDS paralogy. In this case an mRNA entering in the P-bodies recognizes close gene paralogs, coding for members of the same protein family. Members of the same family of proteins perform usually similar functions, but are expressed sometimes alternatively in different tissues. Verification: superimposing a paralog annotation mask to translatome and transcriptome DEG profiles should allow decoding.

1b The code is based on 5' or 3' UTR recognition sites. In this case what is recognized are these sites, and genes ontologically compatible share these sites. Verification: we could start from our map of phylogenetic footprints (PFs) of 5' and 3' UTRs, and superimpose them to to translatome and transcriptome DEG profiles.

(2) For the (1a) hypothesis we need an activity binding to a CDS and finding homolog CDSs. For the (1b) hypothesis we need the same activities binding instead to PFs and finding matching ones (not necessarily homolog

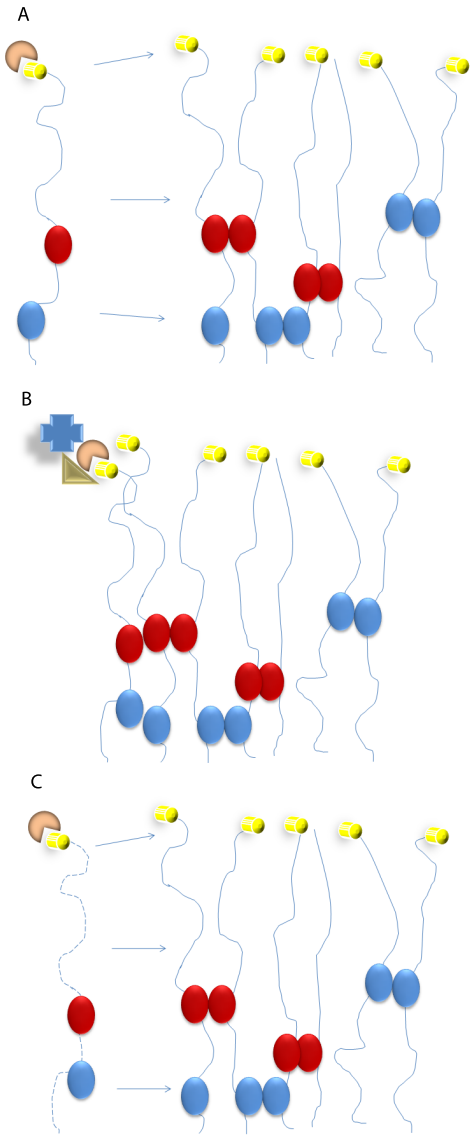


Figure C.11: A: The newly transcribed mRNA arrives at the P-body directed by 4E-T (orange) and interact with a preexisting mRNA/RBP lattice by an interaction code made by cis sequences and bound RBPs. B: The newly arrived mRNA recruits the decapping/degradation machinery because of the presence of 4E-T. C:Degradation of the newly arrived mRNA and release of the lattice from P-bodies through an unknown mechanism

ones). A ncRNA recognizing in trans two or more mRNA sites or an RBP (or an RBP complex) recognizing with different domains two or more

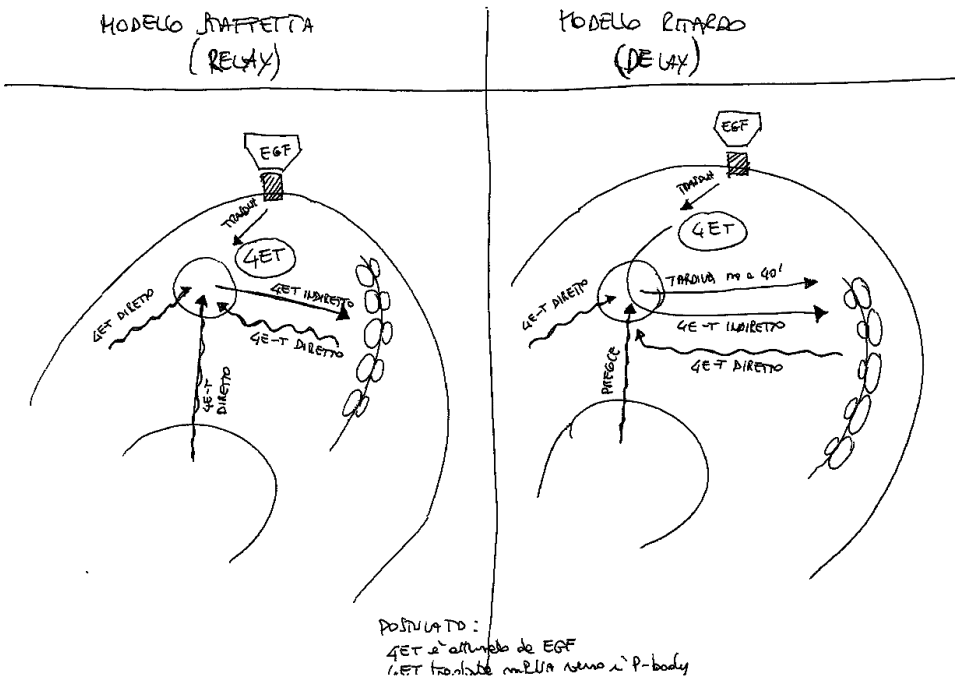


Figure C.12: Relay versus delay model

mRNA 5' or 3' UTRs could perform this. Complexes of different RBPs with different mRNAs have already been described, sort of ribonucleoprotein lattices present in the cytoplasm. What we need here in both cases is that

a all newly transcribed mRNAs after the EGF stimulus are targeted to P-bodies.

b the arrival of a new mRNA in the P-body determines the release of ontologically compatible mRNAs from the P-body.

c degradation of this new mRNA.

What is the mechanism of this process?

(a) If the targeting molecule is a protein or a complex of proteins, which is likely, we need that this protein binds to all newly transcribed mRNAs. It could therefore be component of the eIF4F complex, or of the exon-junction complex, both present in all mRNAs. A very interesting candidate could

be exactly 4E-T, the protein we have targeted just because it destroys P-bodies. 4E-T is a predominantly cytoplasmic protein which binds to eIF4E and directs mRNAs into P-bodies, colocalizing with the P-body decapping factors and also negatively controlling mRNA stability [42]. 4E-T is a very poorly studied protein (only about ten papers). 4E-T could be at least one fundamental protein of a complex present in P-bodies devoted to the recognition and degradation steps of fake reprogramming, involved in the first step of P-body transport of mRNAs once produced in the nucleus. Targeting 4E-T is therefore an efficient way not only to disrupt P-bodies but also to disrupt fake reprogramming, it could be the transporter of the process.

(b) For this we need fist of all a coupler between mRNAs. The coupler, as said, could be a ncRNA, a protein or protein complex, or both. The coupler should be able to act as a bridge between the original incoming mRNA and the ontologically compatible mRNAs already present in the P-body. I think that the of the two presented hypothesis for the code the 1b, recognition by the 5' or 3' UTR sites is the most likely because it can allow great flexibility of coupling. Indeed, a CDS paralogy coupler is only able to aggregate mRNAs of proteins belonging to the same family, while a coupler based on 5' or 3' UTR sites has a great degree of flexibility, it could couple whatever types of mRNAs irrespective of their CDS sequence. To be noted also that the first type of coupler should be more able to tolerate mismatches on a great variety of different sequences, while the second could act on more specific and shorter sequence stretches. In the choice between an RBP-based or a ncRNA-based coupling machinery, RBPs could be more likely to be involved, for their flexibility in forming coupling pairs by interacting with each other so to constitute a complex lattice, but ncRNAs could be involved also in the recognition for the simplicity of a RNA-based bridge. They could also be involved together in the

coupler. A selective releasing system for the mRNAs identified by the code to be released is also necessary. Their specific lattice should be broken , in some way. (c) A triggered mRNA degrading system could be exactly that already described in P-bodies. But how the system could recognize the incoming mRNA as to be degraded? It could be that it bears a specific tag, which could simply be for example 4E-T bound to eIF4E, while the mRNAs already stored in P-bodies are not tagged, for some reason.

C.2 Decoding

The verification of CDS paralogy is simple, while that of 5'/3' UTR site involvement rather complex. From the available data we can simply isolate those mRNAs which after EGF stimulation are translatome increased when other mRNAs are transcriptome increased. The overlapping should be minimal or absent, after having eliminated the highly expressed transcription factors at the top of our list. The code is inside these genes. Let us suppose that no matching between mRNAs associate in proteins families and these two groups of mRNAs is present. We then need to test the second hypothesis. We isolate the PFs of the 5' and 3' UTRs of these mRNAs and we have to establish if there is a preferential matching between patterns of PFs in the two datasets. We basically need to establish clusters of transcriptome increased and translatome increased mRNAs based on these matching patterns, and these clusters should be at that time ontologically compatible. This finding could verify the hypothesis.

C.3 Mechanistic verification of fake reprogramming

We need say three examples of EGF-induced genes buffered in the polysomes but which elicit in our dataset the polysomal increase of ontologically com-

patible mRNAs. We consider the three best clusters. We identify the transcriptome increased mRNA with the higher degree of connections with translatome increased mRNAs. We repeat the EGF experiment and measure these mRNAs in a more precise way, by real-time PCR, and in a say 4-time kinetic. We should see the fake reprogramming in a clearer way and confirm it in these three selected cases, provided that one mRNA is sufficient to trigger the P-body release of other mRNAs. Then, for one of these cases, we stably transfect the complete cDNA of the gene (chosen with sort 5' and 3' UTRs) in the same HeLa cells under a tetracycline-inducible promoter, and we activate the expression of the gene in a controlled and dose-response way. We look then at the other mRNAs in polysomes. We then repeat the same experiments using a luc reporter construct with the 5'3' UTRs of the transcriptionally induced gene and a renilla reporter construct with the 5'3' UTRs of one of the P-body released, polysomally activated mRNAs. We make deletion mutants of these two 5'/3' UTR PFs in order to identpfy the cis regions necessary for the supposed coupling. Looking for the fake reprogramming machinery. If it exists, it should be in the P-bodies. We need to look in the literature about all the demonstrated components of P-bodies (are there enrichment protocols and already published mass spec papers?) and use all the available interactome maps to look for associations of RBPs (1a hypothesis). For the 1b hypothesis we should identify instead bridging ncRNAs between our clusters of ontologically compatible genes derived from the translatome/transcriptome DEGs comparison. Why fake reprogramming? Why a cell should being so complicate in genetically determining phenotypic transitions? Why should a cell discard transcriptome DEG programs which should already be effective, without polysome overlapping and fake reprogramming in determining the phenotype? Some hypotheses:

a The RNA world. Our cells come from a world without DNA, therefore

translatome DEGs where possibly at the origin the only way to instruct phenotypic changes. The crystallization of genetic information into genomes made of DNA allowed its efficient preservation, and could have determined a sort of "triggering role" for transcription in transition states: if you need to change state, remind it to the polysome through transcriptome DEGs which elicit the release from P-bodies of the really active mRNAs.

b If we exclude a small number of "trojan mRNAs" (those which are very strongly increased, which are not buffered by the polysomes) - by the way, there should be a verifiable positive relationship between positive DEG degree and tendency to be coupled genes - which typically in our case are transcription factors instructing the second wave of the EGF program, the other increased mRNAs following the delivery to the nucleus of the signaling pathway could be not enough to trigger the complex first cell reaction. P-bodies become therefore a sort of amplification/resonance system, working in such a way to get a number of mRNAs saying something in input and to release a bigger number of ontologically compatible mRNAs, already transcribed, for translation. The only drawback of this mechanism is that the system looses the original mRNAs, they only work as first relay runners in this sort of two-stage relay race. Possibly this happens because the machinery doing this has degraded the original mRNAs in order to release the ontologically compatible mRNAs.