Studying Post Transcriptional Regulation Layers: Modeling the Dynamics of Molecular Interactions in Living Cells

Thesis for the degree of “Doctor of Philosophy”

By

Shelly Mahlab-Aviv

SUBMITTED TO THE SENATE OF
THE HEBREW UNIVERSITY OF JERUSALEM

June 2018
THIS WORK WAS CARRIED OUT UNDER THE SUPERVISION OF:

PROF. MICHAL LINIAL
Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisors Prof. Michal Linial and Prof. Nati Linial, for the continuous support of my PhD study, for their patience, motivation, and immense knowledge. Prof. Michal Linial provided my first introduction to the fascinating field of computational biology. Fortunately, I was guided and inspired by her and by Prof. Nathan Linial, who always lighten my way and encourage my independent thinking. Our discussions had much more than scientific and computational insights for me.

Next, I would like to thank Prof. Ruthi Sperling and Prof. Liran Carmel for giving me an opportunity to join their research, and benefit from their experience and knowledge.

Last, but not least, I would like to thank Ofer, my wonderful husband, for his endless love and support. And of course, our three beautiful children – Shira, Ronny and Omer, who inspired me just by their presence and put into perspective the important things in life.
Abstract

Gene expression regulation is one of the most important processes in the cell. It is composed of several levels that are extensively studied covering the regulation of initiation of transcription, mRNA processing, translocation from the nucleus, regulation within the cytoplasm and post-translation regulation. However, the overall understanding of the layer of post-transcription regulation that involves microRNA (miRNA) remains fragmented. MiRNA are small non-coding RNA, usually of 22-25 nucleotide in length. In humans, there are about 1900 miRNA genes and about 2500 mature miRNAs. It is estimated that about 60% of the mRNAs are targeted by miRNAs. Most of the miRNA regulation results in mRNA decay and translational inhibition, and is done by base-pair complementarity, where the elementary unit is a miRNA binding site (MBS) at the 3’-UTR of the mRNA. Importantly, the interaction of miRNAs with their targets creates a complex network where one miRNA can recognize multiple MBS (often on hundreds different mRNAs), and each mRNA can be occupied by one and often many miRNAs on its multiple MBS. The dynamics and global properties of such complex network is poorly understood. In most studies, the effect of any specific miRNA was investigated by monitoring the outcome on the expression levels of its predicted target genes. Many of our current knowledge of miRNA is based on analyzing the results from miRNAs overexpression or knockout of a specific miRNA. It was shown that in general, the extent of downregulation of mRNA by binding of to the cognate miRNA is moderate, and in most cases very small. A competition on miRNA binding sites was proposed in the context of the ceRNA hypothesis. In my work I developed a generic framework that simulates the miRNA–mRNA interactions using an unbiased cell based probabilistic approach. The final outcome of the simulation is the retention rate of each gene after simulating the miRNA- mRNA interactions while considering the miRNA-MBS affinity, the initial quantities of miRNA and mRNA in the cells and the availability and degree of occupation of MBS. Surprisingly, exhaustive over expression simulations of 250 different miRNA expose a list of genes that their abundance was significantly reduced in almost any overexpression settings (referred as cross-miRNA sensitive genes), and another larger list of genes that are very stable and were indifferent to the overexpression condition (referred as cross-miRNA stable genes). Moreover, I show that different cell types have different sets of sensitive genes while they share a large fraction of the stable genes. We also found unexpected differences in the sensitivity of specific genes to the degree of overexpression (in a range of 1-1000 folds). Overall, I show that gaining insights on cell states in
term of its miRNA and mRNA levels is achieved by simulating the probabilistic and stochastic nature of the miRNA-mRNA interactions. Although the classical miRNA regulation is taken place in the cytoplasm, we further investigate their regulatory effect in the nucleus. We characterize short RNAs from human HeLa cells found in the supraspliceosome fraction (SF), a nuclear dynamic machine in which pre-mRNA processing occurs. We sequenced small RNAs extracted from the SF and identified sequences that were aligned to different region of miRNA genes. We found that about 60% of those sequences were mature miRNA. However, the profile of the mRNA was not identical to the expression profile of miRNA in the cytoplasm. Among the different miRNA that were found, 95 were not detected in the cytoplasm, and the abundance of additional 58 strongly differed from that in cellular extract. This finding implies the possible overlooked functionality of miRNA within the supraspliceosome. Another level of gene expression regulation is carried out at the level of protein translation efficiency. While miRNAs are expected to act also at the levels of ribosome initiation and translational attenuation, I tested directly the evolutionary signal that encode the efficiency of translation elongation. Protein translation is the most expensive operation in dividing cells from bacteria to humans. Therefore, managing the speed and allocation of resources is subject to tight control. From bacteria to humans, clusters of relatively rare tRNA codons at the N’-terminal of mRNAs have been implicated in attenuating the process of ribosome allocation, and consequently the translation rate in a broad range of organisms. In my work, I demonstrate that proteins translated by free- or ER-bound ribosomes exhibit different overall properties in terms of their translation efficiency and speed in yeast, fly, plant, worm, bovine and human. We noted that only secreted or membranous proteins that are characterized by having a Signal peptide (SP) are specified by having segments of “slow” tRNA at their N’-terminal, followed by abundant codons that are considered “fast”. The definition of slow and fast codons is a reflection of their tRNA affinities and their copy numbers. We postulate that “hard-coded” signals within the secretory proteome assist the steps of protein maturation and folding. Specifically, “speed control” signals for delaying the translation of a nascent protein fulfill the co- and post-translational stages such as membrane translocation, proteins processing and folding.

In this thesis, I will present my research of different layers of post-transcription regulation in living cells, focusing on the impact of both miRNA and translation regulation.
<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1. Pre-transcriptional Regulation</td>
<td>2</td>
</tr>
<tr>
<td>1.2. Post-transcriptional Regulation</td>
<td>3</td>
</tr>
<tr>
<td>1.2.1. RNA Processing</td>
<td>3</td>
</tr>
<tr>
<td>1.2.2. RNA Stability</td>
<td>4</td>
</tr>
<tr>
<td>1.2.3. Translation Regulation</td>
<td>5</td>
</tr>
<tr>
<td>1.3. Thesis Outline</td>
<td>8</td>
</tr>
<tr>
<td><strong>2. Pre-microRNA-aligned sequences in the supraspliceosome</strong></td>
<td>10</td>
</tr>
<tr>
<td>2.1. Appendix</td>
<td>40</td>
</tr>
<tr>
<td><strong>3. Stochastic Modeling of miRNA Regulation</strong></td>
<td>48</td>
</tr>
<tr>
<td>3.1. The Translation machinery is immune from miRNA perturbations: A cell-based probabilistic approach</td>
<td>48</td>
</tr>
<tr>
<td>3.2. Sensitivity of Gene Sets to miRNA Regulation: A Cell-Based Probabilistic Approach</td>
<td>90</td>
</tr>
<tr>
<td>3.3. Appendix</td>
<td>102</td>
</tr>
<tr>
<td><strong>4. Computational Analysis of Translation Elongation Rate</strong></td>
<td>113</td>
</tr>
<tr>
<td>4.1. Conservation of the relative tRNA composition in healthy and cancerous tissues</td>
<td>113</td>
</tr>
<tr>
<td>4.1.1. Appendix</td>
<td>127</td>
</tr>
<tr>
<td>4.2. Speed Controls in Translating Secretory Proteins in Eukaryotes - an Evolutionary Perspective</td>
<td>135</td>
</tr>
<tr>
<td>4.2.1. Appendix</td>
<td>152</td>
</tr>
<tr>
<td><strong>5. Discussion</strong></td>
<td>157</td>
</tr>
<tr>
<td>5.1. Computational model of miRNA regulation</td>
<td>157</td>
</tr>
<tr>
<td>5.2. Computational view of translation efficiency</td>
<td>158</td>
</tr>
<tr>
<td>5.3. Gene sets display different regulation sensitivity</td>
<td>159</td>
</tr>
<tr>
<td>5.4. Regulation in localizations</td>
<td>160</td>
</tr>
<tr>
<td>5.5. Quantitative view of regulation</td>
<td>162</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

This chapter presents a brief introduction to the biological issues addressed in my thesis, namely, investigating and comprehending the different layers of the post transcriptional regulation in cells in health and disease.

According to the central biological dogma, genes encoded in the DNA are transcribed into RNA in the nucleus. For protein coding genes, the RNAs, referred to as messenger RNAs (mRNAs), are translocated to the cell’s cytoplasm. In the cytoplasm, the mRNAs are being translated by the ribosome to form functional proteins. However, in steady state, the mRNA constitutes only 3%-7% of the total RNA in the cell (Palazzo and Lee, 2015). The rest of the expressed RNA, referred as non-coding RNA (ncRNAs), has regulatory and supportive functionalities on mRNA expression and availability for protein translation. The vast majority of these non-coding RNA (80%-90%) is the ribosomal RNAs, 10%-15% are transfer RNAs (tRNAs), and a complex composition of ncRNAs that account for the minor fraction of the expressed RNAs include miRNA, snoRNA, circular RNA and long ncRNA (lincRNAs) (Eddy, 2001; Livny et al., 2008). It is believed that the composition and types of expressed proteins dictate the cell identity and functionality. When the cell undergoes any internal or external perturbations, cell must rapidly response by expressing the relevant genes. This is the basis for the ability of cells to cope with changing conditions as occurs in cases of differentiation and switching developmental stages, viral infection and cancerous transition. It is often the outcome of a response to external ligands such as changes in nutrients, oxygen availability, hormones, toxins or pathogens. Therefore, gene expression is one of the most regulated process from bacteria to humans. The principle design underlying several layers of post-transcriptional regulation is the heart of my thesis.
1.1. Pre-transcriptional Regulation

The first layer of regulation is pre-transcriptional regulation. There are various levels of pre-transcriptional regulations. The first is chromatin accessibility. The chromatin is tightly packaged into DNA binding protein complexes of nucleosomes. The nucleosome positions on the DNA in view of regulatory regions of promoters and enhancers are crucial as it dictates the chromatin accessibility and therefore the possibility for binding the transcription factors and enable or suppress the RNA transcription (Lee et al., 2004; Schones et al., 2008). At the high level of the chromosome, the structure of the chromatin may be regulated between different chromosomal regions, resulting open and condensed regions. Moreover, chromatin accessibility is also regulated by DNA methylation patterns, that are cell identity and cell state specific (Bird, 1986; Zemach et al., 2010). In recent years, due to the technological breakthrough in revealing the 3D view of the nucleus, the dynamics of the chromatin is linked to cancer development, stemness of cells, ageing and hormonal regulations throughout life.

Within the accessible regions, transcription is further regulated by different sets of transcription factors and connectivity of remote and proximal promoters with an increasing understanding on the role of thousands of enhancers throughout the genome. Transcription factors (TFs) are proteins that bind specific DNA sequences in close proximity to genes and may induce or repress their
transcription (Razin, 1984; Mitchell and Tjian, 1989; Latchman, 1997; Struhl, 1999). They are often located within hundreds of nucleotides from the actual transcription start sites (TSS).

1.2. Post-transcriptional Regulation

Upon RNA transcription, there are further several layers of RNA regulation.

1.2.1. RNA Processing

The first layer of post-transcriptional regulation occurs at the nucleus. The first stages of RNA (the pre-mRNA) processing right after or even along its transcription include capping, addition of poly-A tail and splicing for removing introns. The capping of the pre-mRNA involves the addition of 7-methylguanosine on the 5’ end and is mediated by specialized enzymes. The cap protects the 5’ end of the primary transcript from ribonucleases and is thus vital for the mRNA stability (Ramanathan, Robb and Chan, 2016). Polyadenylation is the addition of the poly (A) tail, a stretch of non-template addition of adenine to the mRNA at its 3’ end after the cleavage of the original 3’ end. There are cases where the addition of the poly (A) tail may occur at different sites of the same pre-mRNA transcript, resulting alternative polyadenylated transcripts from the same gene. The poly (A) tail protects the mRNA from enzymatic degradation from the 3’ end and promotes transcription termination, export of the mRNA from the nucleus, and involved in effective translation (Sachs, 1990; Guhaniyogi and Brewer, 2001; Proudfoot, Furger and Dye, 2002). The splicing of the pre-mRNA is a process where introns from the primary transcript are removed at conserved sequences that include the splice sites (Clancy, 2008). The splicing process occurs in cellular machines called spliceosomes. The spliceosome is assembled from small nuclear ribonucleoproteins (snRNPs) and SR proteins that catalyzing the splicing process (Will and Lührmann, 2011). In addition to introns removal, there are many instances of alternative splicing, in which different exons of the transcripts are spliced, thus resulting different processed transcripts from the same pre-mRNA transcript (Berget, Moore and Sharp, 1977; Fong and Bentley, 2001).
1.2.2. RNA Stability

The stability of the mRNA molecule after the first stages of RNA processing is further regulated by the process of RNA interference (RNAi). In this regulatory process, small interfering RNA molecules (siRNA) that are usually of 21- and 22-nucleotides length, interact via base pairing with the target mRNA. This interaction may cause gene destabilization and lead to gene silencing, and therefore have great impact of gene suppression (Elbashir et al., 2001). siRNA regulation was first discovered in nematodes (Fire et al., 1998). Synthetic siRNA is used in many cases of gene manipulations and therapeutics such as gene knocked (Whitehead, Langer and Anderson, 2009). Another class of small RNA that have a dominant role in gene expression regulation are microRNA (miRNA). miRNAs are a class of small endogenous RNAs of 21-25 nucleotides in length that regulate gene expression by partial base-pairing to the mRNA sequences, mostly on the 3’ UTR region of the targeted gene (Pasquinelli, 2012; Ameres and Zamore, 2013; Moore et al., 2015). In multicellular organisms, miRNAs act post-transcriptionally by causing destabilization and degradation of mRNAs, as well as interfering with the translation machinery (Filipowicz, Bhattacharyya and Sonenberg, 2008; Chekulaeva and Filipowicz, 2009; Eichhorn et al., 2014; Bartel, 2018). There are also cases that miRNA may result mRNA deadenylation and decay (Djuranovic, Nahvi and Green, 2012). It is estimated that miRNA constitute as many as 1% - 10% of all predicted genes in nematodes, files and mammals (Lai et al., 2003), and are also exist in plants (Reinhart et al., 2002). miRNAs have a major regulatory role in defending cells against parasitic nucleotide sequences such in the cases of viral infection (Yiliang et al., 2012) and cancer transformation (Lu et al., 2005; Bertoli, Cava and Castiglioni, 2015). With the progress of computational and experimental techniques the number of detected miRNA increased, and today there are ~1900 known miRNA genes in human. miRNAs are transcribed by either RNA polymerase II or RNA polymerase III as a long RNA primary transcript called the primary miRNA (pri-miRNA), (Winter et al., 2009), Figure 1.2. Pri-miRNA transcripts may be edited by ADAR1 and ADAR2 changing adenosine (A) to inosine (I), causing alteration in base pairing and structure, as well as further sequence-based processing (LUCIANO et al., 2004; Yang et al., 2005; Blow et al., 2006). In the next step, still within the nucleus, the pri-miRNA is further processed by the microprocessor complex that contains the double-stranded RNA binding protein DGCR8 and the ribonuclease III enzyme Drosha (Filippov et al., 2000; Wu et al., 2000; Denli et al., 2004). This
processing results stem-loop structures of ~70 nucleotides long that are known as the pre-miRNAs. However, there are some exceptional cases in which Drosha is not involved in the pre-miRNA processing. miRtrons are a collection of miRNAs that encoded in an intron of a hosting gene. In those cases, the pre-miRNA is released after splicing (Berezikov et al., 2007; Okamura et al., 2007; Ruby, Jan and Bartel, 2007) or processed by trimming one of their extended tails (tailed miRtrons). Pre-miRNAs are exported to the cytoplasm by Exportin-5 in complex with Ran-GTP (Yi et al., 2003). In the cytoplasm, the miRNA maturation step occurs with the binding of the pre-miRNA to the RISC loading complex (RLC). This complex is composed of the RNase Dicer and the double stranded RNA-binding proteins TRBP and PACT, and the core component Agronaute-2 (Ago-2). The pre-miRNA is cleaved by dicer into 22-nucleotide duplex. Upon cleavage, Dicer and its interactors TRBP and PACT dissociate from the miRNA duplex. Helicases that associate with the RISC complex unwind the miRNA duplex, and only one strand is usually incorporated into the RISC complex via its binding to Ago-2, where its interaction with the target mRNA occurs, while the other one is degraded. The interaction of the miRNA is done via the seed region (nucleotides 2 to 8) that requires full complementarity to the miRNA binding site (MBS) of the target mRNA (Lewis et al., 2003; Doench and Sharp, 2004; Lewis, Burge and Bartel, 2005).

MiRNAs expression profiles are different between different cell types and different cell states such as development, proliferation, apoptosis and tumorigenesis, which implies of their crucial regulatory role and have an informative information of the identity of the cell. (Pasquinelli et al., 2000; Lu et al., 2005; Volinia et al., 2006; Zhao and Srivastava, 2007; Ribeiro et al., 2014; Suzuki, Young and Sharp, 2017).

Another important property of the miRNA regulatory system is its many-to-many nature. One miRNA may target several mRNAs, and one mRNA may be targeted by several miRNAs, resulting a form of competing endogenous RNAs (Cesana et al., 2011). This mode of regulation results in undirected effects of miRNAs and mRNAs expression profiles, by sequestering the availability of one miRNA by different transcripts with the same miRNA binding sites.

### 1.2.3. Translation Regulation

The expressed proteins in the cells are the dominant factors that dictate cell’s behavior and properties. As many regulatory processes affect the availability of the mRNA were described
above, the next step of protein expression depends on the translation machinery. The translation process is one of the most energetically expensive process in the cell (Plotkin and Kudla, 2010; Tuller et al., 2010; Gingold and Pilpel, 2011). The translational regulation allows rapid and acute changes in the cellular proteomic profile. The process of translation is carried out by the ribosomes, and can be divided into initiation, elongation and termination. The most critical phase in the regulation is at the stage of translation initiation, where the start codon AUG is identified. In eukaryotes, the start codon is usually identified by a scanning mechanism of the mRNA. Once it is found, it is placed at the peptidyl (P) site of the ribosome and undergoes base pairing with the Met-tRNAi. This process is mediated by different eukaryotic initiation factors (eIFs) (Pisarev, Hellen and Pestova, 2007). For example, one of the major regulatory mechanism of translation is

Figure 1.2 The miRNA processing pathway, taken from (Winter et al., 2009)
the phosphorylation of eIF2 resulting a decrease in Met-tRNAi eIF2 complex and this decrease in translation initiation. During the elongation process, the ribosomal RNA (rRNA) molecules in the ribosome machinery are bounded to the mRNA. During the mRNA-tRNA interaction, the amino acid carried by the tRNA is transferred into the growing polypeptide. This process is mediated by several elongation factors (Sonnenberg and Hinnebusch, 2009). Each type of tRNA has a complementary sequence to any codon called the anticodon, and according to the genetic code, it is bounded to a different amino acid. Since there are cases where one amino acid may be coded by several synonymous codons, there may be several different tRNA that carry the same amino acid. The relative genomic abundance of the different synonymous codons varies in different organisms (Sharp and Matassi, 1994; Stenico, Lloyd and Sharp, 1994). Moreover, synonymous codon usage in different genes tends to be related to their level of expression, and highly expressed genes usually encoded by the more abundant tRNAs (Anderson, 1969; Bulmer, 1987).

Ribosome profiling techniques of ribosome density shows that different genes have different translation efficiency and dynamic (Arava et al., 2003; Ingolia, Hussmann and Weissman, 2018). Setting the translation efficiency as the ratio between the ribosome attached to the mRNA and the total of mRNA fragments in yeast show a 100-fold range of translation efficiency between the different expressed genes. Moreover, the ribosome distribution along the mRNA is non unified. It was shown that ribosome density was 3-fold greater at the first 30-40 codons (Ingolia et al., 2009). This finding was the basis for the ‘ramp’ model, accordingly, the slow translation codons in the beginning of the coding sequence may avoid jamming of ribosomes once they passed it (Gingold and Pilpel, 2011), Figure 1.3. Ribosome translation efficiency depends on several factors, among them are availability of ribosomes, tRNA, amino acids, translation factors and cell energy. The availability of tRNA can be approximated by the genomic tRNA copy number (Dittmar et al., 2004; Tuller et al., 2010). The level of translation efficiency for each codon can be measured by the tRNA adaptation index (tAI), that considers the available tRNA pool and wobble roles (Reis, Savva and Wernisch, 2004). Moreover, the order of the codons within the same mRNA molecule also effect the translation efficiency (Cannarozzi et al., 2010). Therefore, the specie specific tRNA copy number and the codon usage and codon sequence of each mRNA has an underlying effect on its translation efficiency.
1.3. Thesis Outline

The remaining chapters of this thesis present the publications of my research and analysis of the different layers of gene expression regulation that I investigated in recent years.

Chapter 2 presents a quantitative analysis of miRNA abundance within the nucleus. Since miRNA biogenesis and pathway usually terminates at the cytoplasm, the presence of mature miRNA within the nucleus is a phenomenon that was mostly overlooked. Only in recent years, it became evident that a pool of miRNA is in the nucleus. I performed a rigorous analysis of the fragments in the nucleus. I found different unexpected fragments of the pre-miRNA within the nucleus, and specifically in the supraspliceosome complex. Although most of them where mature miRNA, the complexity of the fragment composition extend beyond the mature miRNAs. Comparing their abundance to the cytoplasm, shows different set of miRNAs. We hypnotized that those miRNAs have non-canonical regulatory roles within the nucleus, and their localization within the supraspliceosome complex may implies mRNA splicing related functionality.

Chapter 3 presents a system view of the miRNA – mRNA interaction network. In this work I built a stochastic model of this network, that simulates miRNA- mRNA interactions, according to the expression profiles of different cells. The outcome of this simulation presents the available mRNA
pool. The simulator also enables to perform over expression experiments of different genes by different over expression factors. I used the retention level of the mRNA as measure to reflect the regulation and some aspects of its dynamics. I found that an overexpression of miRNA has a direct effect on its target genes reducing their retention rate at the end of the simulation, also affect some of its non-target genes by increasing their final retention rates. An analysis of exhaustive overexpression simulations of all 248 miRNAs in 3 different types of cells was performed. I was able to show that there is a coherent set of genes that remain stable under any of the miRNA manipulations. This set of genes is enriched by translation annotations and is share by the three cell types despite their variability in the set of miRNA and genes and their compositions. However, for each cell type we found another set of genes that their retention was minimal and are shown to be extremely sensitive to the actual composition of the most abundant miRNA that specify each cell type. The design of cells was revealed by showing set of genes that are cell specific and other sets that are generic. The later one is associated with the translation machinery of any cell. Further analysis of the simulation results shows that different gene sets are affected by miRNA regulation in different rates. We also found the different gene sets are sensitive to different miRNA over expression factor.

Chapter 4 presents the analysis of translation efficiency in view of the tRNA adaptation index of different genes. In one work presented in this chapter we found that the genomic tRNA copy number is a good proxy for their abundance in the cell. We also found that in different cell states, i.e., healthy and cancerous cell-lines and tissues, although there is a difference in the total amount of expressed tRNA, their proportion is kept. Next, I discovered that mRNA that are translated by ribosomes that are localized on the endoplasmic reticulum (ER) membrane can be modeled by the ‘ramp’ notion, as their translation elongation pattern shows slowly translated codons at their ribosome localization sequence. However, I was able to show that the ramp model applied to these set of secretory proteins and it is not dominant feature for all other soluble proteins in cells from plants to humans and across many model systems. In Chapter 5 I conclude with a brief discussion of the different research presented and the new understanding on post-transcriptional regulation in living cells.
Chapter 2

Pre-microRNA-aligned sequences in the supraspliceosome

This chapter includes the unpublished research of the analysis of miRNA in the supraspliceosome
Pre-microRNA-aligned sequences in the suprasspliceosome

Mahlab-Aviv Shelly ¹, Peretz Ayelet R ², Boulus Ayub ², Carmel Liran ², Sperling Ruth ²,* and Linial Michal ³,*

¹The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel; ²Department of Genetics, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel; ³Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

*Corresponding author: SR, LM

The Hebrew University of Jerusalem,
Edmond J. Safra Campus, Givat Ram, Jerusalem 91904, ISRAEL

Telephone: +972-2-6585162 (SR); +972-54-8820035 (LM); FAX: 972-2-6523429

SR: r.sperling@mail.huji.ac.il
LM: michall@cc.huji.ac.il

Figures: 1-7

Tables: 1-3

Additional files: Table S1-S6; Figures S1-S4

Keywords: nuclear microRNA, miRBase, Ribonucleoproteins (RNPs), microRNA biogenesis, mirtron, small non-coding RNA, microprocessor
* Rich repertoire of pre-miRNA-aligned sequences are found at the supraspliceosome, several of them are exclusive to it.

* Supraspliceosome pre-miRNA-aligned sequences cover diverse regions of the pre-miRNA, including mature miRNAs.

* Novel potential functions for supraspliceosomal pre-miRNA-aligned sequences in pre-mRNA processing.

* The abundance and the base-pairing potential of miR-7704 at the supraspliceosome is correlated with the expression of HAGLR gene.
MicroRNAs (miRNAs) are short non-coding RNAs that negatively regulate the expression and translation of genes in healthy and diseased tissues. Herein, we characterize short RNAs from human HeLa cells found in the supraspliceosome, a nuclear dynamic machine in which pre-mRNA processing occurs. We sequenced small RNAs (<200 nt) extracted from the supraspliceosome, and identified 304 sequences that are aligned to pre-miRNAs genes. About two thirds of them are mature miRNAs, whereas the rest account for various defined regions of the pre-miRNA. Out of these pre-miRNA-aligned sequences, 95 were undetected in cellular extract, and the abundance of additional 58 strongly differed from that in cellular extract. Notably, we describe seven abundant miRNA-aligned sequences that overlap non-coding exons of their host gene. Furthermore, the abundant miR-99b may affect splicing of LINC01129 primary transcript through base-pairing with its exon-intron junction. The rich collection of pre-miRNA-aligned sequences at the supraspliceosome suggests overlooked nuclear functions. We show that the level of miR-7704 expression and the degree of HAGLR expression are correlated. We claim that in cases of extended base-pairing complementarity, such supraspliceosomal pre-miRNA-aligned sequences might have a role in transcription attenuation, maturation and processing.
INTRODUCTION

MicroRNAs (miRNAs) are small ~22 nt long molecules that have been implicated in regulating all cellular signaling pathways, and alterations in their expression level are associated with the development and progression of diseases, including cancer (1). Their main role is thought to be the negative regulation of gene expression and translation, mostly by base-pairing to the 3’-UTR of target mRNA transcripts in the cytoplasm (2,3).

The canonical biogenesis of miRNAs from RNA Polymerase II (Pol II) transcripts involves a number of steps. The first step occurs in the nucleus by the microprocessor, whose key proteins are DGCR8 and Drosha. DGCR8 binds the RNA molecule, while Drosha, an RNase III type enzyme, cleaves the primary (pri) miRNA transcript into a precursor (pre) miRNA stem-loop molecule of 70-80 bases (4-6). Next, the pre-miRNA is exported by Exportin-5 to the cytoplasm, where it is cleaved by the RNase III Dicer, yielding double stranded RNA that comprises the mature miRNA and its complement. The leading strand of the miRNA is then loaded on Argonaute protein (AGO), the key component of the RNA-induced silencing complex (RISC), which stabilizes the miRNA and directs its binding to its target gene (Reviewed in (3)).

Recent studies revealed the presence of mature miRNAs in the nucleus (7-9), suggesting nuclear functions in addition to the known cytoplasmic ones (10,11). Several studies provide evidence for active shuttling of miRNAs from the cytoplasm to the nucleus (12,13), involving Exportins (14). Although the function(s) of miRNAs in the nucleus are not yet well understood, recent reports demonstrated their involvement in a number of processes, such as the regulation of non-coding RNAs (15-18), and transcriptional silencing (19). Search for nuclear miRNAs complementarity in human promoters revealed additional targets, suggesting that they may function in transcription inhibition (20). Nuclear miRNAs were also shown to affect transcriptional activation, a process that requires the involvement of AGO and proximal promoter of non-coding transcripts (21). Analysis of miRNA-mRNA-AGO interactions, revealed substantial AGO-miRNA mapping to intronic sequences (22).

A large proportion of miRNA genes are located in introns of coding genes, albeit many are likely transcribed from their own Pol II dependent promoters (3,23,24). Multiple lines of evidence show strong links between splicing and miRNA processing. For example, splicing components were reported to co-sediment with the microprocessor and pre-miRNAs (25); processing of
miRNA-211 was shown to promote the splicing of its hosting intron (26); knockdown of AGO1, AGO2 and Dicer affect alternative splicing (AS) (13,27); splicing factors were found within chromatin-associated AGO proteins (13); nuclear small RNAs were shown to affect AS, and altered splicing requires AGO2 expression.

Most mammalian Pol II transcripts undergo pre-mRNA splicing (28). Notably, splicing and AS play a major role in the regulation of gene expression, and changes in splicing are common in many human diseases including cancer (29,30). Splicing, as well as additional processing events of Pol II transcripts, occurs in the cell nucleus within a huge (21 MDa) and highly dynamic machine known as the supraspliceosome (31-33). It is a ribonucleoprotein (RNP) complex, composed of four native spliceosomes, each similar to the in-vitro assembled spliceosome, that are connected by the pre-mRNA (34,35). The entire repertoire of nuclear pre-mRNAs, independent of their length and number of introns, is individually assembled in supraspliceosomes that include the five spliceosomal U snRNPs, and all known splicing factors (32,33). Supraspliceosome offers coordination and regulation for pre-mRNA processing events including AS (36,37) and other processing steps, such as 5’-end and 3’-end processing, and RNA editing (38). Thus, the supraspliceosome – the endogenous spliceosome - is involved in all nuclear processing activities of pre-mRNAs (31-33).

We hypothesized that there is a cross-talk between their processing and pre-mRNA splicing within the supraspliceosome, primarily for miRNAs residing on the same transcriptional unit. The main microprocessor components, Drosha and DGCR8 are found in supraspliceosomes (39). In addition, the presence in the supraspliceosome of several pre-miRNAs from intronic sequences have been previously reported (39,40).

Here we focus on supraspliceosomal small RNA sequences aligned to pre-miRNAs from HeLa cells. We refer to these sequences as miRNA gene aligned sequences (miR-GAS). Using small ncRNA (<200 nt) deep sequencing, we identified hundreds of miR-GAS sequences in the supraspliceosome. Two thirds are mature miRNAs and the rest represent other, typically well-defined, regions of the pre-miRNAs. We focus on miR-GAS that are exclusively expressed in the supraspliceosomal fraction (SF), and point out potential functions for such set of pre-miRNAs. Our results imply an unexplored function for miRNAs at the supraspliceosome in post-transcriptional gene regulation, splicing and nuclear pre-mRNA processing.
MATERIALS AND METHODS

Isolation of supraspliceosomes

All isolation steps were conducted at 4°C. Supraspliceosomes were prepared from nuclear supernatants enriched in supraspliceosomes as described previously (34). Briefly, nuclear supernatants were prepared from purified cell nuclei by microsonication of the nuclei and precipitation of the chromatin in the presence of excess of tRNAs. The nuclear supernatant was fractionated on 10-45% (vol/vol) glycerol gradients. Centrifugations were carried out at 4°C in an SW41 rotor run at 41 krpm for 90 min [or an equivalent \( w^2t = 2500 \) (\( w \) is in krpm; \( t \) is in hr)]. The gradients were calibrated using the tobacco mosaic virus as a 200S sedimentation marker. Supraspliceosome peak fractions were confirmed by Western Blot (WB) and by electron microscopy visualization.

Protein detection

Western blotting (WB) analyses were performed as previously described (41). We used anti-hnRNP G (kindly provided by Prof. Stefan Stamm, University of Kentucky, Lexington), visualized with horseradish peroxidase conjugated to affinity-pure Goat anti-Rabbit IgG (H+L; Jackson Immunoresearch, 1:5000). The MAb-104 that is directed against the phosphorylated epitopes of the SR proteins (42) was used as previously described (43).

RNA isolation from supraspliceosomes and deep sequencing

RNA was extracted from supraspliceosomes prepared from HeLa cells, as previously described (34). The integrity of the RNA was evaluated by an Agilent 2100 bioAnalyzer. For small RNA library construction, ~1 µg of RNA was used. After phosphatase and T4 polynucleotide kinase (PNK) treatments, the RNA was ethanol precipitated to enrich for small RNA, and small RNA libraries (in triplicates) were prepared according to NEBNext Small RNA Library Prep Set for Illumina (Multiplex Compatible) Library Preparation Manual. Adaptors were then ligated to the 5’ and 3’ ends of the RNA, and cDNA was prepared from the ligated RNA and amplified to prepare the sequencing library. The amplified sequences were purified on E-Gel® EX 4% Agarose gels (ThermoFisher # G401004), and sequences representing RNA smaller than 200 nt were extracted from the gel. The library was sequenced using the Illumina NextSeq 500 Analyzer. The sequencing
data, after removal the adaptors and filtering out low quality sequences, were aligned to mirBase (Release 21). In addition, the filtered high-quality fragments were mapped to the human transcriptome of hg19 gtf file from UCSC provided by Galaxy. The hg19 transcriptome contains 963,559 exons from 45,314 transcripts. Sequences aligned to the miRNA genes as compiled in miRBAse are reported.

**Small RNA preparation from HeLa cell extract**

Total RNA from HeLa cells was purified using QIAzol Lysis Reagent RNeasy Plus Universal Kit (QIAGEN, Germany). Sample has been transferred up to an RNeasy Mini-spin column and centrifuged for 15 sec at >8000 g at room temperature, and the supernatant was processed according to the manufacturer’s standard protocol. Samples with an RNA Integrity Number (RIN) of >8.5, as measured by Agilent 2100 Bioanalyzer, were considered for further analysis. Small RNA libraries (<200 nt) were generated using the Illumina TruSeq RNA V2 (Illumina) library protocol. The RNA was tested for quality and subjected to a library preparation of miRNA and mRNA according to TruSeq RNA standard protocol. Read length was 100 nt. Data used are derived from two biological duplicates.

**Next Generation Sequencing (NGS) analysis**

RNA extracted from the supraspliceosome fractions (SFs) in HeLa cells was taken from six independent library preparations. HeLa cells supraspliceosomes were taken from three independent biological preparations. Each preparation was split, and used to produce two libraries. NGS was performed for each sample on small RNA (<200 nt) molecules using standard Illumina Protocol. Each library consisted of ~25 million reads of maximum length 76 (see supplementary Table S1).

Raw data of the sequenced small RNA were trimmed using Cutadapt ver. 1.13. Low-quality reads were filtered out using FASTX toolkit. Reads from the SF were aligned against human genome hg19 and miRbase database (version 21) using TopHat 2.1.1, allowing 90% sequence identity and a maximum of two mismatches. Reads whose start or end position were mapped to miRNA genes were considered. High quality reads from the six SF preparations were combined. The correlation between the expression of all miRNAs from HeLa-1 to HeLa-6 ranges from 0.58
to 0.82. Each sample is associated with 100 to 170 pre-miRNA genes (Supplementary Figure S1). Out of the mapped reads, only reads of length \( \geq 17 \) were considered. miRNA gene aligned sequences (miR-GAS) refer to all mapped, high quality reads that are aligned to any of the pre-miRNA as defined by miRBase.

A significant correlation is calculated between the mapping schemes of high-quality reads against the sequences in miRBase and the human transcriptome (Supplementary Table S2). Furthermore, the correctly mapped reads between the two mapping methods are comparable (138.0 k and 138.8 k reads from mapping applied on miRBase, and the whole transcriptome, respectively).

Only miR-GAS with \( \geq 3 \) reads were considered. This threshold was used as the total reads assigned to SF is limited (all mapped reads are 139k, Supplementary Table S2). However, for improving the statistical analyses, higher thresholds were considered (as detailed in the text). Sequencing of the cellular extract (CE) used read length of 100 nt, that were mapped to miRNA genes using miRExpress 2.0. The total number of reads for the CE for miRNA library is 29.8 M reads among them 3.25 M reads are mapped to miRNAs (Supplementary Table S1).

**Validation of gene expression**

RNA was extracted from HeLa and MCF10A cells. The MCF10A, human mammary epithelial cell-line, was purchased from ATCC (CRL-10317). RT-PCR was performed on RNA extracted from HeLa and MCF10A cells and from nuclear supernatants of the above cells as described (59). The following sets of primers for HAGLR was used: Forward (exon 1) 5’-CGTCGGAGCGCGAGAATT-3’ and Reverse 5’-AAGGGCCCATTTCAGGCA-3’ (exon 2). The primers for HOXD1 are: Forward (exon 1) 5’-ATTTACCTCCGGCTACACTCG-3’ and Reverse: 5’-AGGTGCAAGCAGTTGGCTAT-3’ (exon 2). The identity of all PCR products was confirmed by sequencing. Each experiment was repeated at least 3 times. The relative abundance was quantified in view of the intensity of the \( \beta \)-actin that was used as a control. The \( \beta \)-actin Forward and Reverse primers, for an amplicon of 140 nt, are: 5’-CTGGAACGCTGGAAGTGACA-3’ and 5’-AAGGGACTTCTCTGTAACATGCA-3’, respectively.
RNA folding and Bioinformatics tools

We have used the RNAfold software as part of a set of web-tools for analyzing the thermodynamics of folding (44). We applied the program by default parameters. The set of miR-GAS that are longer than 22 nt were analyzed. The minimum free energy (MFE) that is calculated is reported along other quantities such as the number of predictions and the average free energy for the folded structure.

RESULTS

Hundreds of small RNA types are associated with the supraspliceosome

A large proportion of miRNA genes are harbored within introns of genes transcribed by Pol II. As processing of Pol II transcripts occur in supraspliceosomes, we hypothesized that key microprocessor components should be present in supraspliceosomes. Here, we wished to characterize the composition of supraspliceosome-associated small RNAs that align to pre-miRNAs (hereinafter miR-GAS). To this end, we prepared nuclear supernatants enriched with supraspliceosomes under native salt conditions from HeLa cells, and fractionated them on glycerol gradients as previously described [(34), see Materials and Methods]. The extraction protocol preserves the higher order splicing complexes as shown by electron microscopy (34,45). Supraspliceosomes sediment at 200 S; and the splicing factors SR proteins and hnRNP G, predominantly found in supraspliceosomes and associated with them, mark their position in these gradients (36,46). Figure 1 shows the results of the WB analyses across the gradient using antibodies directed against these splicing factors.
Figure 1. Splicing factors mark the supraspliceosome fraction (SF). WB analysis of the distribution across the glycerol gradient of splicing factors, previously shown to be predominantly found and associated with supraspliceosomes. Nuclear supernatants enriched for supraspliceosomes were prepared from HeLa cells and were fractionated in 10-45% glycerol gradients. Aliquots from odd gradient fractions were analyzed by WB using anti-hnPNP G and anti-SR antibodies. Supraspliceosomes peak in fractions 9-13. The distribution of hnRNP G (42 kDa), SRSF3 (SRp20, 20 kDa) and SRSF6 (SRp55, 55kDa) across the gradient is presented. TMV was used as a size marker for the sedimentation (200 S).

Next, we extracted small RNA (< 200 nt) from the SF (fractions 9-12, Figure 1). The RNA was used to build a barcoded library for sequencing by NGS (see Materials and Methods). Altogether, we analyzed six sequencing datasets, each library yielded 20 M to 30 M raw reads (Supplementary Table S1). Alignment of these supraspliceosomal sequences revealed a diverse collection of small RNA species, including pre-miRNAs, small nucleolar RNAs (SNORDs) (47), intronic sequences and more. In this study, we only consider reads that are aligned to pre-miRNA sequences as determined by miRBase (See Materials and Methods, Supplementary Table S1).

Overall, 504 miR-GAS were identified in the combined data from supraspliceosomes in HeLa cells (Supplementary Table S2). Only the 304 SF miR-GAS that are >=17 nt long, and are supported by >= 3 reads will be further discussed (see Materials and Methods, Supplementary Table S2). To test that we do not introduce a bias by mapping the NGS reads to miRBase sequences, we have repeated the analysis while mapping the reads to the entire human transcriptome (hg19 genome, see Materials and Methods). The two mappings gave consistent results (Spearman correlation r=0.78, p-value=3.3e-113, Supplementary Table S2).

Additionally, independent libraries were prepared from total small RNA (<200 nt) from HeLa cell extract (CE). The data from CE is used as a reference, representing the mature, unfractionated, cellular miRNAs in HeLa cells.
Figure 2. The collection of miR-GAS from HeLa in the supraspliceosomes. (A) Intersection of pre-miRNAs aligned sequences found in SF and CE. (B) Intersection of mature miRNAs aligned sequences found in the SF and the CE. Counting of miRNAs is a unified collection from the 6 independent libraries. Each miRNA is associated with >=3 reads. For >95% of all the miR-GAS evidence appears in at least 2 of the HeLa libraries.

Supraspliceosome-associated miRNA profile differs from the cellular one

A large population of human miRNAs are located in introns. These are either transcribed together with their host gene on the same transcriptional unit or at the opposite strand. Among them, many are transcribed by their own promoters, others are coupled to splicing as mirtrons or tailed mirtrons (48). The rest of the miRNAs are associated with intergenic regions and are single exon genes transcribed by Pol II as autonomous transcriptional unit. To test whether the collection of miRNAs identified at the SF are limited to intronic miRNAs, and thus are naturally involved with the splicing machinery during miRNA biogenesis, we compared the genomic location of the miRNAs that were identified in SF versus CE (Supplementary Figure S2). The SF is significantly enriched with miRNAs of intronic localization (72%) relative to the CE (62%; p-value = 0.00076). Note that the fraction of reads associated with miRNAs in the SF is only 4.2% of their number in the CE (Supplementary Table S1), suggesting that the impact of the SF on the CE (which also includes the nucleus and the supraspliceosome within it) is negligible. The substantial fraction (28%) of intergenic miR-GAS in the supraspliceosome indicates that the presence of these sequences in SF are not a mere reflection of pre-miRNA biogenesis but rather suggest a broader role in gene expression.
We assessed the overlap between the miRNAs identified in HeLa SF and CE, the latter being dominated by cytoplasmic miRNAs. Out of the 304 supraspliceosomal miR-GAS, 209 appear in both preparations (SF and CE), and 95 could only be identified in SF (Figure 2A). Focusing on mature miRNAs in the two fractions, we identified 187 mature miRNAs that are shared between SF and CE, and 30 that are exclusive to the SF (Figure 2B). In addition, there are many mature miRNAs (330) that appear in the CE but not in the SF (see Supplementary Table S3).

Table 1 lists the SF miR-GAS that are supported by >=20 reads. The most striking observation is the extremely high abundance of miR-6087 that accounts for 17% of the total reads associated with miR-GAS from the SF, yet it is virtually absent in the CE. Note that most miRNAs that are exclusive to the SF were recently identified, and are considered novel miRNAs (48,49). Nevertheless, a careful assessment of miR-6087 proposes that it is not a canonical miRNA. Other highly abundant SF miR-GAS that are undetected in CE include miR-633a, miR-7704 and miR-622. For a complete list of miR-GAS in SF and CE, and for the subset of mature miRNAs in the SF, see Supplementary Table S3.

As the total number of reads from the SF is substantially smaller relative to the CE (4.2%), we compared the ranking of pre-miRNAs from both preparations rather than their absolute amounts. For this analysis, we considered the 200 miRNAs from SF with a predetermined threshold of >=10 reads (Figure 3A), and identified those that exhibit substantial ranking difference in the SF compared to the CE (where 1 is the rank of the miRNA with the highest number of reads). The ranking of most miRNAs is roughly preserved. However, the ranking of 38 miRNAs (~25%) substantially differ between the SF and the CE (Supplementary Table S3). From them, 12 miRNAs show over-representation in the CE (e.g., miR-148a, let-7a-1, let-7a-2 and let-7a-3). However, the great majority show an opposite trend, and are over-represented in the SF. For example, miR-1246 is ranked 7 in the SF, but only 96 in the CE. Other extreme examples include miR-378C, miR-663b and miR-7704 (Figure 3A). We suggest that these miR-GAS are of a special interest, and their prevalence at the supraspliceosome indicates potentially overlooked functions in the nucleus.
Table 1. Top 23 miR-GAS that are detected in the SF but not in the CE.

<table>
<thead>
<tr>
<th>miRNA</th>
<th># reads in SF</th>
<th>% in the SF</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-6087</td>
<td>23686</td>
<td>17.04</td>
</tr>
<tr>
<td>miR-663a</td>
<td>2853</td>
<td>2.05</td>
</tr>
<tr>
<td>miR-7704</td>
<td>2798</td>
<td>2.01</td>
</tr>
<tr>
<td>miR-622</td>
<td>2436</td>
<td>1.75</td>
</tr>
<tr>
<td>miR-3960</td>
<td>226</td>
<td>0.16</td>
</tr>
<tr>
<td>miR-663b</td>
<td>210</td>
<td>0.15</td>
</tr>
<tr>
<td>miR-2861</td>
<td>142</td>
<td>0.10</td>
</tr>
<tr>
<td>miR-7110</td>
<td>94</td>
<td>0.07</td>
</tr>
<tr>
<td>miR-3648</td>
<td>81</td>
<td>0.06</td>
</tr>
<tr>
<td>miR-7705</td>
<td>65</td>
<td>0.05</td>
</tr>
<tr>
<td>miR-492</td>
<td>60</td>
<td>0.04</td>
</tr>
<tr>
<td>miR-639</td>
<td>55</td>
<td>0.04</td>
</tr>
<tr>
<td>miR-7641-1</td>
<td>54</td>
<td>0.04</td>
</tr>
<tr>
<td>miR-7641-2</td>
<td>48</td>
<td>0.03</td>
</tr>
<tr>
<td>miR-3654</td>
<td>48</td>
<td>0.03</td>
</tr>
<tr>
<td>miR-8086</td>
<td>40</td>
<td>0.03</td>
</tr>
<tr>
<td>miR-1244-3</td>
<td>38</td>
<td>0.03</td>
</tr>
<tr>
<td>miR-3652</td>
<td>35</td>
<td>0.03</td>
</tr>
<tr>
<td>miR-4792</td>
<td>30</td>
<td>0.02</td>
</tr>
<tr>
<td>miR-7706</td>
<td>27</td>
<td>0.02</td>
</tr>
<tr>
<td>miR-5047</td>
<td>23</td>
<td>0.02</td>
</tr>
<tr>
<td>miR-4444-1</td>
<td>22</td>
<td>0.02</td>
</tr>
<tr>
<td>miR-4709</td>
<td>21</td>
<td>0.02</td>
</tr>
</tbody>
</table>
For testing whether the observed phenomenon is dominated by mature miRNAs, we repeated the above analysis for mature miRNAs only (Figures 3B, Supplementary Table S2). For SF miRNA-GAS that are aligned mostly to mature mRNAs we observed, as expected, no change between Figures 3A and 3B (e.g., miR-148a, let-7a-1, let-7a-2 and let-7a-3, Supplementary Table S3). It is evident that the ranking bias (Figure 3A) applies also to miR-GAS that are not aligned to mature miRNA sequences (e.g., miR-19b-1, miR-622). These non-canonical sequences strongly point at an additional miRNA processing mechanism, presumably within the supraspliceosome.

Table 2 compares the most abundant miR-GAS in the SF to their cognate miRNAs in the CE. Note that the top 10 miR-GAS account for over 70% of all miR-GAS reads from the SF. Among the most prevalent miR-GAS are variants of let-7, miR-21 and miR-100, which are also extremely abundant in CE.

Table 2 also indicates the percentage of mature miRNAs that are associated with each of the identified miRNA-GAS (see Supplementary Table S4).

**Figure 3.** Comparison of miR-GAS abundance ranking in SF and CE. (A) Comparison of the rank order for 200 miR-GAS with a threshold of >10 reads in each, in view of their rank in CE. Note that rank 1 is associated with the most abundant miRNA sequence and rank 200 associated with the lower number of reads. (B) Comparison of ranks when only mature miRNAs are compared in both preparations. The dashed lines show arbitrarily boundaries for defining a miRNA that show a significant deviation in ranking by at least 4 folds. The annotated miRNAs are representatives of those showing a significant bias in their rank order. Those that are biased towards a higher abundance in SF relative to CE are shown in the top-left section and the miRNAs with an opposite trend are in the lower-right section of the figure. Source data are in Supplementary Table S3 and Table S4.
Table 2. Top 20 miR-GAS ranked by their abundance in the SF relative to the CE. The full list includes 612 miR-GAS (Supplementary Table S3).

<table>
<thead>
<tr>
<th>miRNA</th>
<th>SF read count</th>
<th>CE read Count</th>
<th>Rank SF (re-mapping)</th>
<th>Rank CE</th>
<th>% in the SF</th>
<th>% in the CE</th>
<th>% SF mature miRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-6087</td>
<td>23686</td>
<td>0</td>
<td>1</td>
<td>239</td>
<td>17.04</td>
<td>0.00</td>
<td>0.19</td>
</tr>
<tr>
<td>hsa-let-7i</td>
<td>16229</td>
<td>243109</td>
<td>2</td>
<td>2</td>
<td>11.67</td>
<td>7.49</td>
<td>99.75</td>
</tr>
<tr>
<td>hsa-let-7g</td>
<td>11704</td>
<td>68575</td>
<td>3</td>
<td>13</td>
<td>8.42</td>
<td>2.11</td>
<td>99.97</td>
</tr>
<tr>
<td>miR-21</td>
<td>11085</td>
<td>1003460</td>
<td>4</td>
<td>1</td>
<td>7.97</td>
<td>30.91</td>
<td>98.70</td>
</tr>
<tr>
<td>hsa-let-7f-2</td>
<td>10257</td>
<td>137382</td>
<td>5</td>
<td>3</td>
<td>7.38</td>
<td>4.23</td>
<td>99.91</td>
</tr>
<tr>
<td>hsa-let-7f-1</td>
<td>10169</td>
<td>133159</td>
<td>6</td>
<td>4</td>
<td>7.31</td>
<td>4.10</td>
<td>99.18</td>
</tr>
<tr>
<td>miR-1246</td>
<td>3365</td>
<td>1025</td>
<td>7 (8)</td>
<td>96</td>
<td>2.42</td>
<td>0.03</td>
<td>100.00</td>
</tr>
<tr>
<td>miR-92a-1</td>
<td>3364</td>
<td>87008</td>
<td>8 (7)</td>
<td>7</td>
<td>2.42</td>
<td>2.68</td>
<td>99.11</td>
</tr>
<tr>
<td>miR-92a-2</td>
<td>3173</td>
<td>78660</td>
<td>9</td>
<td>12</td>
<td>2.28</td>
<td>2.42</td>
<td>99.97</td>
</tr>
<tr>
<td>miR-663a</td>
<td>2853</td>
<td>0</td>
<td>10 (11)</td>
<td>237</td>
<td>2.05</td>
<td>0.00</td>
<td>0.88</td>
</tr>
<tr>
<td>miR-7704</td>
<td>2798</td>
<td>0</td>
<td>11 (10)</td>
<td>246</td>
<td>2.01</td>
<td>0.00</td>
<td>100.00</td>
</tr>
<tr>
<td>miR-622</td>
<td>2436</td>
<td>0</td>
<td>12 (14)</td>
<td>264</td>
<td>1.75</td>
<td>0.00</td>
<td>0.08</td>
</tr>
<tr>
<td>miR-30a</td>
<td>2311</td>
<td>79239</td>
<td>13 (12)</td>
<td>11</td>
<td>1.66</td>
<td>2.44</td>
<td>98.92</td>
</tr>
<tr>
<td>hsa-let-7b</td>
<td>2063</td>
<td>17074</td>
<td>14 (13)</td>
<td>36</td>
<td>1.48</td>
<td>0.53</td>
<td>99.47</td>
</tr>
<tr>
<td>miR-20a</td>
<td>1446</td>
<td>8216</td>
<td>15</td>
<td>43</td>
<td>1.04</td>
<td>0.25</td>
<td>18.95</td>
</tr>
<tr>
<td>miR-3687</td>
<td>1415</td>
<td>4</td>
<td>16</td>
<td>219</td>
<td>1.02</td>
<td>0.00</td>
<td>4.31</td>
</tr>
<tr>
<td>miR-24-2</td>
<td>1410</td>
<td>21682</td>
<td>17</td>
<td>27</td>
<td>1.01</td>
<td>0.67</td>
<td>98.94</td>
</tr>
<tr>
<td>miR-24-1</td>
<td>1370</td>
<td>20909</td>
<td>18</td>
<td>29</td>
<td>0.99</td>
<td>0.64</td>
<td>99.93</td>
</tr>
<tr>
<td>miR-99a</td>
<td>1368</td>
<td>63965</td>
<td>19</td>
<td>14</td>
<td>0.98</td>
<td>1.97</td>
<td>100.00</td>
</tr>
<tr>
<td>miR-100</td>
<td>1342</td>
<td>107344</td>
<td>20</td>
<td>5</td>
<td>0.97</td>
<td>3.31</td>
<td>99.40</td>
</tr>
</tbody>
</table>

^a^ Ranking according to mapping vs human transcriptome (in parenthesis). The rank for all other miRNAs remained identical in both mappings schemes.
Over a third of the supraspliceosome-associated miR-GAS are not mature miRNAs

We classified the miR-GAS found in the SF according to the pre-miRNA region to which they align: miRNA tail, mature miRNA (combining the mature sequences that are at the 5p and 3p), undefined complement (i.e., no evidence for a mature miRNA in the complementary sequence in miRBase), loop, and compete pre-miRNA. In addition, reads that do not reside within a single region, but rather overlap two or more regions, are categorized as overlapping region. Finally, miR-GAS that were aligned just before or after the pre-miRNA tails are categorized as pre-miRNA margins (Figure 4A).

Figures 4B and 4C show the distribution of reads among these seven categories, as well as the read length distribution within each category. Mature miRNAs form the majority of miR-GAS (63.7% of normalized counts), and their length is, as expected, centered around 22 nt. The remaining 36.3% miR-GAS mostly belong to overlapping regions and to margins. Notably, the length distribution in these two categories is wide (Figure 4B).

Figure 4. Partition of the miR-GAS to seven pre-miRNA regions. (A) A schematic view on miRNA prototype, where the 7 different regions of the pre-miRNA are listed according to their positions on the pre-miRNA and their structural and functional features (see text). (B) Length histogram of reads mapped to each region according to the 7 regions defined in A. (C) The relative counts of reads to each of the pre-miRNA region. Total counts of reads (left) and normalized counts (right, the relative abundance of reads aligned to a
specific miRNA is summed to 1) are shown. All miRNAs that are analyzed have >=3 reads and a minimal read length is =>17.

Most miR-GAS associated with a high number of reads are assigned entirely to mature miRNAs (Table 2). However, there are some exceptions. For miR-6087, miR-663a and miR-622, mature miRNA sequences account for <1% of the aligned reads, suggesting an abundant yet-uncharacterized pre-miRNAs fragments. Additionally, for miR-20a and miR-3687, mature miRNAs account for 4.3% and 19% of the reads, respectively. In both instances, we identified a large number of overlapping reads, suggesting non-conventional processing.

We further analyzed all miR-GAS with >=10 reads. For only 94 of them (47%) we observed that the mature miRNA is supported by > 99% of the reads (Figure 5A, Supplementary Table S4). These mature miRNAs might be explained as shuttled back from the cytoplasm. In 77 of the miR-GAS (38.5%), at least 10% of the reads are not assigned to mature miRNAs. Among them, for 33 miRNAs (16.5%), the amounts of mature miRNAs is <1%. For example, the vast majority of the reads aligned to miR-1291 pre-miRNA are sequences of the complete pre-miRNA rather than the mature sequence. Another example, 216 out of 233 reads of miR-4436 are assigned to overlapping regions (Supplementary Table S4). The results in Figure 5A strongly suggest an overlooked processing mode that differs from the canonical one, that generates the bulk of cytoplasmic miRNAs.

Figure 5B displays the overall position and counts of reads along the pre-miRNA, showing the unified, normalized data from 304 miR-GAS from SF (with >=3 reads each, Supplementary Table S4). It is clear that many reads are associated with miRNAs at the margins in both ends of their pre-miRNAs. Some proximal expressed pre-miRNAs (separated along the genome by less than a read length, 76 nt) may share reads of neighboring miRNA genes that are in a cluster (e.g., miR-92a1 and miR-92b1 are only 36 nt apart, and mir-20A are only 57 nt apart from miR-92b1).
Figure 5. Read-length and relative position of aligned fragments. (A) Analysis of the 200 miR-GAS that are supported by ≥10 reads sorted by their fraction of mature miRNA versus the total reads. (B) Sum of reads mapped to all analyzed miRNAs. Read-length of the different miRNAs are aligned relative to the pre-miRNA start position. (C) Sum of reads mapped to miR-21. (D) Sum of reads mapped to miR-6087. (E) Sum of reads mapped to miR-663a. (F) Sum of reads mapped to miR-3687. Blocks of reads from the same starting points reflect non-randomized cleavage sites. The 5p and 3p mature miRNAs are colored in red, and the width of the bars is proportional to the number of reads. The coordinates of the miRNA are according to their genomics location as defined by the forward and reverse strands. Dashed block marks the position of the mature miRNAs as annotated by miRBase.
A varied distribution of reads along the pre-mRNA regions characterizes SF pre-miRNA collection

Figure 5C shows the profile and length distribution of miR-21. This miRNA is among the most abundant miRNAs identified in the SF. In this case, the mature miRNA dominates, and its fraction is >98.8% of all reads. It is likely that miR-21 was shuttled back from the cytosol and as such it is a prototype for other miRNAs that are processed in the cytosol but are still identified at the SF. Furthermore, most reads in Figures 5C have the same sharp and well-defined edges, supporting a well-defined processing pathway. In miR-6087, the most abundant miR-GAS expressed in SF (17%, Table 1, Figure 5E), 44% of the reads are aligned to the margins and the rest are aligned to overlapping regions, with only minute fraction aligned to the mature miRNA sequence. Another illuminating example is miR-663a (2853 reads, Figure 5D), where most of its reads map to overlapping regions and not to the classical mature 5p. Figures 5F illustrates miR-3687 that exhibit a complex, non-conventional mapping pattern and provides evidence for a full miRNA hairpin. The set of examples described above shows the unique characteristics that specify the pre-miRNAs that we have identified in the SF (Supplementary Table S4). This unique pattern supports a non-conventional processing, presumably at the supraspliceosome.

We tested the possibility that miR-GAS that are longer than classical mature miRNAs (22 nt) are characterized by stable secondary structures. A structure-based blockage in the processivity of a nuclease at the supraspliceosome may lead to accumulation of such reads. The folding energy that is associated with non-canonical miR-GAS fragments was calculated. Among the long miR-GAS that are abundant in the SF, we observed miR-6087, miR-3687, miR-622 and miR-639, as candidates for structural-dependent processing at the supraspliceosome (Supplementary Table S5).

5p and 3p miRNAs strand preference in the SF in comparison to the CE

To further search for characteristics of supraspliceosomal miRNAs, we tested the relative abundance of the strands identified by the mature miRNAs (5p and 3p). As both strands may originate from a single pre-miRNA, we defined the percentage of the 5p population using the fraction of 5p/(5p+3p). Supplementary Figure S3 presents these values for 127 mature miRNAs that are found in both the SF and the CE (Supplementary Table S6, a threshold >=10 reads). We note that as expected from a canonical processing of the duplex miRNAs, mature miRNAs originate from either the 5p arm or the 3p arm, and rarely from both. Indeed, mature miRNAs from
the CE have a balanced representation, on average (Supplementary Figure S3). Mature miRNAs from the SF show a balanced representation as well, with a slight, insignificant, preference to the 5p strand.

While the overall distribution of the 5p versus 3p is similar between the SF and CE, some individual miRNAs exhibit marked differences. Supplementary Figure S3 shows miRNAs where the difference in 5p percentage between the SF and CE is >20% (e.g., miR-30e and miR-27a). A complete list is provided in Supplementary Table S6.

Table 3 lists the miRNAs with the most significant difference between the prevalence of the 5p and 3p strands between the SF and the CE (only miRNAs with >20 reads in the SF are considered, to increase the reliability of detection). Table 3 also includes miR-GAS that seem to have preference towards a strand that is not documented in miRBase (50), thus labeled by us as ‘undefined complementary’. For example, miR-622 is considered to be produced entirely of the 3p arm. However, we identified 1513 reads that presumably emerge from the complementary 5p arm.

**Table 3. Percentage of mature miRNAs coming from the 5p arm.**

*Bottom, reads assigned to “undefined complement” within the pre-miRNA.*

<table>
<thead>
<tr>
<th>miRNA Intersect</th>
<th>SF</th>
<th>SF 5p (%)</th>
<th>CE</th>
<th>CE 5p (%)</th>
<th>Difference (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-mir-374a</td>
<td>36</td>
<td>8.3</td>
<td>2070</td>
<td>49.9</td>
<td>41.6</td>
</tr>
<tr>
<td>hsa-mir-27a</td>
<td>431</td>
<td>2.1</td>
<td>24594</td>
<td>43.6</td>
<td>41.5</td>
</tr>
<tr>
<td>hsa-mir-365a</td>
<td>23</td>
<td>56.5</td>
<td>762</td>
<td>20.3</td>
<td>36.2</td>
</tr>
<tr>
<td>hsa-mir-181a-1</td>
<td>81</td>
<td>100.0</td>
<td>6513</td>
<td>69.0</td>
<td>31.0</td>
</tr>
<tr>
<td>hsa-mir-30e</td>
<td>173</td>
<td>8.7</td>
<td>7258</td>
<td>34.6</td>
<td>25.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>miRNA</th>
<th>SF</th>
<th>SF 5p (%)</th>
<th>miRBase annotation</th>
<th>miRBase 5p (%)</th>
<th>Difference (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-mir-492</td>
<td>26</td>
<td>0.0</td>
<td>only 5’p</td>
<td>100</td>
<td>100.00</td>
</tr>
<tr>
<td>hsa-mir-4449</td>
<td>52</td>
<td>100.0</td>
<td>only 3’p</td>
<td>0</td>
<td>100.00</td>
</tr>
<tr>
<td>hsa-mir-622</td>
<td>1513</td>
<td>100.0</td>
<td>only 3’p</td>
<td>0</td>
<td>100.00</td>
</tr>
<tr>
<td>hsa-mir-6087</td>
<td>110</td>
<td>41.8</td>
<td>only 5’p</td>
<td>100</td>
<td>58.20</td>
</tr>
</tbody>
</table>
Potential functions of the pre-miRNAs from SF

To obtain insights regarding possible functions of miR-GAS, we focused on the 23 miR-GAS that were observed exclusively at the SF (Table 1). The number of reads per miRNA varies among them by two orders of magnitudes (Supplementary Figure S4A). We focused on seven of these miRNAs that reside within exons, as their possible function may be more readily understood (Supplementary Figure S4A, orange bars).

Supplementary Figure S4B shows a genome browser view for these miRNAs. Whereas miR-7705 and miR-4709 are located at the 3’UTR of their host gene (framed by light blue), the other five are located at the first 5’UTR exon of their host gene (framed by light red). Their location near the start of the transcription hints at a possible role of these five miRNAs in recruiting components of the transcriptional machinery and consequently are likely to attenuate the expression of their host genes.

We identified 2798 reads for miR-7704-5p (Supplementary Table S4) that resides within the HOXD1 genes covering the coding region of the first exon. This high number of reads suggests that this miRNA has the potential to indirectly attenuate the expression of HOXD1, a key nuclear transcription factor that is involved in limb differentiation. It is also possible that this mature miRNA prevents transcription by base-pairing with the HAGLR gene (called HOXD Antisense Growth-Associated Long Non-Coding RNA). As its name implies, it is complementary to miR-7704-5p and to HOXD1 (Figure 6A).

To validate potential function of SF miR-GAS in gene expression we tested the effect of changes in expression of SF miR-7704 on gene expression of HOXD1 and HAGLR. To this end, we compare the expression of HOXD1 and HAGLR transcripts in HeLa and in mammary epithelial cells MCF10A with the level of SF miR-7704 in these two cell types (Figure 6B). The same protocols for isolating SF, sequencing the small RNA (<200 nt) and NGS analysis were applied for the MCF10A cells. We find an inverse correlation between the expression of SF miR-7704 and that of HAGLR that is expressed in the antisense direction. The level of SF miR-7704 is 2.3 folds lower in HeLa cells relative to MCF10A cells. The RT-PCR analysis revealed that the expression of HAGLR is higher in the HeLa cells compared to MCF-10A cells (11.9 folds).
HOXD1 is likely to be indirectly affected by the abundance of miR-7704, presumably via the attenuation of HAGLR transcript (an antisense of HOXD1). We show that HOX1D expression is in accordance with the abundance of miR-7704 and in an inverse relation to the HAGLR levels, for each of the tested cell-lines (Figure 6C).

**Figure 6.** Analysis of miR-7704 from SF for potential functional role. (A) Genome browser view for miR-7704 and the overlap with HOXD1 and HAGLR, a long non-coding RNA gene. (B) Average expression levels of miR-7704 from the SF based on HeLa (6 samples) and MCF10A cells (3 samples). The expression values are based on normalizing each sample to miRNA read counts (100k per sample). The statistical bars are the calculated standard errors. (C) Validation of the effect of SF miR-7704 by RT-PCR analysis of HOXD1 and HAGLR expression in HeLa and mammary epithelial MCF10A cells. A quantification of the products is assessed with respect to β-actin expression (arbitrary units), error bars are the standard error for 3 experiments.
**Figure 7.** The chromosomal context of miR-99a portrays potential function in splicing. (A) The sequence of the pre-miRNA is shown and the 5p and the 3p are colored in red and blue, respectively. The image focuses on isoforms of SPACA6 gene (Sperm Acrosome Associated 6). The isoform with an extended upstream 5’ non-coding exons. (B) Several resolutions illustrating the ability of miR-99b-5p to completely complement, by base-pairing, to the 5’ splice junction of the intron of SPACA6P-AS, which is transcribed in the antisense direction of SPACA6. (C) Sum of reads mapped to miR-99b. (D) Detailed base-paring at the intron-exon junction. See text for details.

**Supraspliceosomal miRNAs may affect gene expression and splicing**

The presence of miR-GAS in supraspliceosomes suggests that at least some of these sequences might compete with splicing. An interesting example is miR-99b (Figure 7A). Almost all of the reads (96%) originates from the 5p arm of the miRNA (miR-99b-5p, Figure 7C) and the rest of the reads are from the 3p arm. Figure 7B shows that miR-99b is harbored in intron 1 of an isoform of the SPACA6 gene (Sperm Acrosome Associated 6), which has an extended upstream 5’ non-
coding exons. This isoform is one of several non-coding, but experimentally validated, variants of the SPACA6 gene. Interestingly, miR-99b is fully complementary to the 5’ splice junction of intron 1 of SPACA6P-AS, which is transcribed in the antisense direction of SPACA6 (Figure 7B), spanning positions -6 to +16 with respect to the splice junction. SPACA6P-AS, also termed LINC01129, is a long non-coding RNA for which very little is known. The abundant miR-99b-5p (Figure 7C) is complementary to the splice junction of SPACA6P-AS, 5’- AAG/gtcggt -3’ (exon is in capital letters) (Figure 7D), possibly forming a duplex of 22 nt.

DISCUSSION

miR-GAS from the SF differs from miRNAs in CE

Although the most studied role of miRNAs is the post-transcriptional regulation in the cytoplasm, information on nuclear functions of miRNAs is accumulating (reviewed in (7-9)). In this study, we focus on sequences aligned to pre-miRNAs found in supraspliceosomes, the nuclear pre-mRNA processing machine. We identified a large number of sequences aligned to 304 pre-miRNAs in supraspliceosomes, supporting our earlier finding of pre-miRNAs of intron 2 of HTR2C (40), and intron 13 of MCM7 in supraspliceosomes (39). It is also in accordance with our finding of the presence of DGCR8 and Drosha in supraspliceosomes, and the cross-talk between the microprocessor and the splicing machine (39). Taken together, it is possible that miRNA sequences also play a nuclear role in the supraspliceosome, possibly regulating gene expression and processing, through a yet undetermined mechanism.

Moreover, comparing HeLa miR-GAS profile in SF with their profile in CE, we identified sequences that are exclusive to the SF (Table 1, Figure 2). Some of the most abundant pre-miRNA sequences in the SF (e.g., miR-6087, miR-663a, miR-7704, miR-622, Table 1) are hardly detected in the CE even though the cellular extract coverage is 23.5-fold larger. This suggests the presence of a unique repertoire of pre-miRNA derived sequences at the supraspliceosome.

Many supraspliceosomal pre-miRNA aligned sequences result from a non-canonical RNA processing

As many miRNAs are harbored in introns, we expected to find sequences derived from intronic miRNAs in SF. Whereas the SF is indeed enriched with intronic miRNAs, we nevertheless find many sequences derived from miRNAs that are not embedded in introns (e.g., miR-1246, let-7i
and miR-92a-2, Supplementary Figure S2). Furthermore, the canonical pathway of miRNA biogenesis includes cleavage of the pri-miRNA in the nucleus by the microprocessor to generate the pre-miRNA stem-loop molecule of about 65-80 nt long. The pre-miRNA is then exported to the cytoplasm, where it is further processed by Dicer to generate the mature miRNAs (2,3,51,52). Thus, the expectation is to find mostly complete pre-miRNA in the SF. Yet, we observe sequences coming from various parts of the pre-miRNA – sequences that align to the complete pre-miRNA sequences, to mature miRNAs, and to other regions along the pre-miRNA (Supplementary Table S4). Atypical biogenesis pathways of pri-miRNAs in the nucleus have been documented (3), including the mirtron pathway which is DGCR8 and Drosha independent (53,54), the simtron pathway, which is DGCR8 independent (55), and the agotron pathway that is both DGCR8/Drosha and Dicer independent (3,56). Nevertheless, all these alternative processing pathways involve transport of a 65-80 nt stem-loop pre-miRNA from the nucleus to the cytoplasm. Thus, these pathways cannot explain the discovery of predominantly short fragments from the SF miR-GAS.

Many of the fragments of miR-GAS that we identified in supraspliceosomes are of a defined length and are characterized by sharp edges (Figure 5), suggesting that they are not byproducts of unspecific degradation. Furthermore, these defined fragments cannot be a result of the failure of the DNA polymerase to proceed due to modifications in the RNA along the process of library preparation. The library was prepared directly from the extracted small RNA fraction, and the 5’ and 3’ adaptors were ligated prior to any amplification step (see Materials and Methods).

A key finding in our study concerns the high percentage of mature miRNAs in the SFs, representing 64% of all normalized reads (Figures 4, Supplementary Table S6). The presence of mature miRNAs in nuclei has been reported as a result of an active shuttling from the cytoplasm to the nucleus (10,11). Notably, many of the exclusively identified miRNAs in the SF belong to the novel set of miRNAs that were recently identified from large-scale bioinformatics surveys [e.g., (57)].

Finding of reads which are aligned to overlap regions (>20%, Figure 4), and the fact that many of the reads are short and cover regions other than that of the mature miRNAs (Figures 4-5), suggest a local, nuclear processing through an unknown mechanism. We can therefore conclude that the defined sequences that we find represent genuine sequences, whose biogenesis and localization in the supraspliceosome calls for further research.
Supraspliceosomal pre-miRNA-like sequences originate from genomic strategic locations

Currently miRBase (version 21) includes 1881 miRNA precursors (2588 mature miRNAs) (50). Many of the newly reported miRNAs listed in miRBase may not comply with the canonical biogenesis and a genuine cytoplasmic cellular location [e.g., (48)]. Tools such as novoMIRank (58) are testing the confidence of novel candidate miRNAs from NGS experiments. Revised definitions for authentic miRNAs versus other miRNA-like sequences are proposed [see discussion in (59)].

A careful analysis of sequences that are abundant in SF and are not detected in CE (or detected in minute amounts, Supplementary Table S3) reveals about 10 cases where the sequences might also come from highly transcribed regions such as the rRNA gene clusters (60). Specifically, several of the exclusive miR-GAS are identical to sequences derived from rRNA clusters, and are often ignored at the NGS mapping step. Yet, they are also identical to sequences of the relevant pre-miRNA. For example, miR-6087 (within 28S), miR-3687 and miR-663a (within 5’-external transcribed spacers of rRNA, 5’ETS) (61). These fragments were renamed as rRNA-hosted miRNA analogs (rmiRNAs), and were shown to be involved in cancer (61). Whether these stable miRNA-like fragments are processing byproducts of abundant RNA species calls for future validations.

Potential functions of supraspliceosomal miR-GAS in gene expression and splicing

The supraspliceosome is a dynamic complex that packages pre-mRNA transcripts of different sizes and number of introns into complexes of a unique structure, that enables coordination of all pre-mRNA processing in the nucleus, including splicing, AS, 5’/3’ end processing, and RNA editing. It is not clear yet where and how the processing of the SF miR-GAS occurs, and further studies are required to elucidate this point. However, our results reveal that miR-GAS represent a defined population that varies from the cellular population in several aspects. These results suggest that at least some of the SF miR-GAS may function in regulating gene expression.

Recent studies revealed the involvement of nuclear miRNAs in a number of functions including transcription silencing and activation (7-11). An interesting instance is miR-320, a tumor suppressor that targets the BCR/ABL oncogene (62). Kim et al (19) have studied miR-320 that is encoded within the promoter region of POLR3D (RNA Polymerase III Subunit D) in the anti-sense orientation, and thus fully complementary to it. An inverse correlation between the expression of
miR-320 and POLR3D in a number of cell lines was shown. Furthermore, down-regulation of miR-320 resulted in an increased expression of POLR3D, and overexpression of miR-320 led to down-regulation of POLR3D. Also, transfection of HEK293 cells with miR-320 lead to heterochromatization and transcriptional gene silencing of POLR3D (19).

It should be noted that we identified 898 reads of miRNA-320a in supraspliceosomes, 99.5% of them were mature miRNA-3p that likely affect expression of POLR3D (Supplementary Tables S4, S6). Kim et al (19) also point at additional miRNAs that show full complementarity to promoter regions. Interestingly, among these miRNAs is miR-639, also identified in our study, which is expressed from the same strand as the protein coding gene TECR. This miRNA overlaps with the 5’UTR of TECR (Supplementary Figure 4B), and is thus likely to affect TECR transcription by sequestering factors. We identified four additional supraspliceosomal miRNA-derived sequences that overlap the 5’UTR regions of genes. Similarly, we noted the abundance of miRNAs that overlap genes at their 3’UTR as well. These include miR-7705 that overlaps the gene PABPC1, and miR-4709, located at the 3’UTR of the NPC2 gene (Supplementary Figure S4B). Most of these genes play a role in key cellular processes, and are thought to function in the nucleus. It is likely that such miRNAs may affect the stability and/or localization of their respective premRNAs. Three of the SF miRNAs, miR-7704, miR-7705 and miR-7706 (Supplementary Figure S4), were recently identified together with a small set of unknown miRNAs that are activated in macrophages by interferon-27 (63).

We validated the potential of miR-7704 to alter the expression of the gene product at its vicinity. This miRNA overlaps with the 5’ end of the HOXD1 gene, and also displays a perfect complementary to the 5’ UTR of the HAGLR gene (Figure 6A). Therefore, it can play a role in coordinating their expression patterns. Our results show an inverse correlation between the expression of HAGLR and miR-7704 in two different cell lines, as expected from a direct competition on transcription. It is likely that whereas HAGLR acts as a negative regulator of HOXD1, miR-7704 help enhancing HOXD1 transcription by neutralizing the expression of its antisense repressor.

Transcriptional gene silencing and other nuclear functions were proposed to occur by mature miRNA (19,20). Importantly, in the case of SF miR-GAS with potential nuclear function in gene expression, only the sequences derived from mir-7704 are mature miRNA sequences. In five of the other six cases, the majority of the reads are derived from overlap regions (miR-7705, miR-
4709, miR-639, miR-3652, miR-4444-1). The question how sequences derived from overlap regions of pre-miRNA were produced, and how they remain associated with the supraspliceosome is beyond the scope of this study. Irrespective of that, the identification of SF miR-GAS that can target or compete transcripts by base pairing, suggests that sequences beyond the mature miRNAs should be considered in playing a role in the nucleus.

The interaction of small RNAs with the supraspliceosome is not limited to miRNAs. Indeed, supraspliceosome-associated SNORD27, lacking the methylase fibrillarin, was demonstrated as regulating the AS of E2F7 transcription factor, through base-pairing, and also affecting AS of several other genes (47). Among the miR-GAS found in supraspliceosomes we identified miR-99b that might play a similar role in splicing. This miRNA (Figure 7), which is harbored in the intron of the SPACA6 gene, is fully complementary to the 5’ splice junction of the first intron of SPACA6P-AS, transcribed in the antisense direction of SPACA6 (Figure 7B). This complementarity suggests that SF miR-99b can compete with the binding of U1 and U6 snRNPs required for splicing of LINC01129 (Figure 7D). Thus, miR-99b can play a role in determining the ratio between spliced and unspliced LINC01129 RNA. Furthermore, the supraspliceosomal miR-99b might be a new player in the balance of expression of SPACA6 and LINC01129.

CONCLUSIONS

The supraspliceosome is a dynamic machine involved in all pre-mRNA processing activities. Because both the canonical and atypical biogenesis steps of miRNAs involve the export pre-miRNA to the cytoplasm, we do not expect to find mature miRNAs, or processed fragments that originate from pre-miRNAs in the supraspliceosome. Our findings illustrate the rich spectrum of miRNAs-aligned sequences in the supraspliceosome, suggesting a novel potential function for some of these miRNAs within the supraspliceosome, in addition to their known function of attenuating gene expression and suppressing translation in the cytoplasm. Here we presented a body of evidence that strongly suggest a role for specific miRNA-aligned sequences in regulating transcription, RNA processing and splicing.
ACKNOWLEDGMENTS

We would like to thank Aviva Pecho for excellent technical assistance, and Tsiona Eliyahu for preparation of the small RNAs used for the HeLa cell extract. We thank Prof. Stefan Stamm (University of Kentucky) for the anti-hnRNP G antibodies, and the CS system at the Hebrew University for the support in processing and storage the data. This study is partially supported by Israel Cancer Research Fund (ICRF) Acceleration Grant (R.S and L.C), and an ERC grant 339096 on ‘High dimensional combinatorics’ (S.M-V), and the ELIXIR Excelerate # 676559 (M.L).

FUNDING

This study is partially supported by Israel Cancer Research Fund (ICRF) Acceleration Grant (R.S and L.C), and an ERC grant 339096 on ‘High dimensional combinatorics’ (S.M-V), and the ELIXIR Excelerate # 676559 (M.L).

Availability of data and materials

The sequencing data of this manuscript have been deposited in GEO and the accession number is GSE100803.

Competing interests

The authors declare that they have no competing interests relevant to the work and data reported here.

Ethics approval

The authors declare that the study did not raise ethical issues in the context of the current legislation.
2.2 Supplementary Material

2.2.1. Supplementary Figures

**Figure S1.** (A) The number of reads (RPM) that were mapped to pre-miRNA in each of the HeLa 6 samples. (B) Number of different miR-GAS that were found in each sample. (C) The heatmap of the correlations between the 6 samples of HeLa (marked 1-6).

**Figure S2.** Partition of total miRNA genes according to their position: miRNAs within genes (blue) and intergenic miRNAs (no overlap with any other gene, orange). (A) All miRNAs in the SF and (B) All miRNAs in CE.
Figure S3. Comparison of mature miRNA expression in SF and CE. (A) Percentage of 5p mature miRNAs in both the SF and the CE (0 and 100 denote only 3p and only 5p, respectively). (B) Scatter plot of the percentage of 5p mature miRNAs in SF (x-axis) and the CE (y-axis).
Figure S4. Analysis of miRNAs from SF for potential functional role. (A) Read counts for miRNAs identified exclusively in SF and not in CE (log10). Marked in orange are miRNAs that reside within a gene and overlap terminal exons. (B) Genome browser view for the 7 miRNAs that are identified only in the SF and were undetected in the CE. For 2 miRNAs (framed, light blue) the miRNAs are located at the 3' UTR exon of their host gene: (i) miR-7705 has the potential to bind to the 3' UTR tail and compete with PABPC1 stability. PABPC1 is a poly(A) binding protein that shuttles between the nucleus and cytoplasm and binds
to the 3′ poly(A) tail of mRNAs. (ii) miR-4709 is located at the 3′UTR of the NPC intracellular cholesterol transporter 2 (NPC2). It is possible that miR-4709 affects NPC2 expression and/ or its localization.

The 5 mother miRNAs (framed, light red) have a unique location at the first exon at the 5′UTR of their host gene: (i) miR-4444-1, located on the 5′-UTR of heterogeneous nuclear RNP A3 (HNRNPA3), which plays a role in cytoplasmic trafficking of RNA and may be involved in pre-mRNA splicing (64). (ii) miR-639, which is expressed from the same strand as the protein coding gene TECR (Trans-2,3-Enoyl-CoA Reductase), which is involved in cholesterol transporting and fatty acid metabolism; (iii) miR-3652 overlaps the first exon (covering the 5′UTR and translation start site) of the Heat Shock Protein 90 Beta Family member (HSF90B1), a stress dependent molecular chaperon; and (iv) miR-7706 overlaps the 5′UTR and part of the promoter of the A-Kinase Anchoring Protein 13 (AKAP13), a gene encoding for an adaptor protein of protein kinase A. (v) miR-7704 overlaps HOXD1 and HAGLR (see Figure 6 and text).
REFERENCES


42. Roth, M.B., Murphy, C. and Gall, J.G. (1990) A monoclonal antibody that recognizes a phosphorylated epitope stains lampbrush chromosome loops and small granules in the amphibian germinal vesicle. JCB, 111, 2217-2223.


Chapter 3

Stochastic Modeling of miRNA Regulation

3.1. The Translation Machinery Is Immune from miRNA Perturbations: A Cell-Based Probabilistic Approach

This chapter will include the submitted work:

The Translation Machinery Is Immune from miRNA Perturbations: A Cell-Based Probabilistic Approach. Shelly Mahlab-Aviv, Nathan Linial, Michal Linial, bioRxiv 298596; doi: https://doi.org/10.1101/298596
The Translation Machinery Is Immune from miRNA Perturbations: A Cell-Based Probabilistic Approach

Shelly Mahlab-Aviv¹, Nathan Linial¹ and Michal Linial²*

¹The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel; ²Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

*Corresponding author: ML

The Hebrew University of Jerusalem,

Edmond J. Safra Campus, Givat Ram, Jerusalem 91904, ISRAEL

Telephone: +972-2-6585448; +972-54-8820035; FAX: 972-2-65623429

LM: michall@cc.huji.ac.il

Running title: Modeling cellular miRNA regulation

Figures: 1-7

Expanded View:

Appendix Text EV1, Appendix Figures S1-S6; Table EV1, Dataset EV1-EV10
Abstract

Mature microRNAs (miRNAs) are non-coding RNA that regulate most human genes through base-pairing with their targets. Under a condition of transcriptional arrest, cells were manipulated by overexpressing miRNAs. We observed global time-dependent changes in mRNA retention which are not restricted to the overexpressed miRNA targets. We developed COMICS (Competition of MiRNA Interactions in Cell Systems), a stochastic computational iterative framework for identifying general principles in miRNA regulation. We show that altering the composition of miRNAs governs cell identity. We identified gene sets that exhibit a coordinated behavior with respect to an exhaustive overexpression of all miRNAs. Among the stable genes that exhibit high mRNA retention levels, many participate in translation and belong to the translation machinery. The stable genes are shared among all tested cells, in contrast to the sensitive, low retention genes that are cell-type specific. We conclude that the stochastic nature of miRNA action imparts an unexpected robustness to living cells. The use of a systematic probabilistic approach exposes design principles of miRNAs regulation toward cell states, cell identity, and the translational machinery.

**Keywords:** Actinomycin D/ CLIP-Seq / miRNA-target prediction/ Stochastic model/ TargetScan
Synopsis

This study models the miRNA regulatory network in different cell-lines under a transcriptional arrest paradigm. The probabilistic model is implemented using a stochastic computational framework called COMICS. The molecular outcome under miRNA regulation scheme is revealed from thousands of simulations under exhaustive miRNA manipulations.

- Transcription arrest emphasizes the impact of miRNA manipulations on gene expression. The composition of miRNA, and primarily the most abundant ones dominate mRNA attenuation in living cells.
- Cell identity can be shifted by the cell-specific composition of the miRNAs but not the mRNAs.
- An exceptional immunity of genes vis-a-vis any manipulations of miRNAs is a property of the translational machinery. It suggests a common signature with respect to the nature of the miRNA binding sites.
- Changing the parameters of the miRNAs probabilistic model affects the dynamics of gene expression but not its steady state or the cell identity.
Introduction

Mature microRNAs (miRNAs) are small, non-coding RNA molecules (∼22 nucleotides) that regulate genes through base-pairing with their cognate mRNAs, mostly at the 3’ untranslated region (3’-UTR) (Ameres & Zamore, 2013; Moore et al, 2015; Pasquinelli, 2012). In multicellular organisms, miRNAs act post-transcriptionally by affecting the destabilization and degradation of mRNAs, as well as interfering with the translation machinery (Chekulaeva & Filipowicz, 2009; Eichhorn et al, 2014; Filipowicz et al, 2008). Transition between cell states are accompanied by alterations in the profile of miRNAs (Pelaez & Carthew, 2012). This applies, e.g., to cells that undergo quiescence (Cheung et al, 2012), differentiation (Yang et al, 2013), viral infection (Zhang et al, 2012) and cancer transformation (Bertoli et al, 2015; Lu et al, 2005).

In humans there are ~2500 mature miRNAs that are derived from ~1900 genes (Ameres & Zamore, 2013). The corresponding miRNA post-transcriptional regulatory network can be modeled as an edge-weighted bipartite graph with nodes corresponding to ~20,000 coding gene transcripts, and ~2500 mature miRNAs. Edge weights represent the efficiency of the relevant miRNA-mRNA interactions. Studies of the miRNA-mRNA regulatory networks reveal that almost all coding genes have multiple putative miRNA binding sites (MBS) at their 3’UTR (Landgraf et al, 2007; Liang et al, 2007; Stark et al, 2005), and many miRNAs can possibly target hundreds of transcripts (Balaga et al, 2012; Rajewsky, 2006). The abundance of miRNAs in cells is extremely unbalanced, with a few highly expressed miRNAs, and hundreds of low abundance (Gaur et al, 2007). This miRNA-mRNA “many to many” relation underlies the capacity for noise reduction (Ebert & Sharp, 2012; Herranz & Cohen, 2010; Schmiedel et al, 2015) and robustness against environmental fluctuation (Li et al, 2009).

Current knowledge on the specificity of miRNA-mRNA regulatory network is based on experimental and different computational prediction tools (Peterson et al, 2014). However, available miRNA-target prediction tools suffer from a large number of false positives (Pinzon et al, 2017). In-vitro studies in which gene expression is monitored following individual miRNA overexpression or knockdown (Hausser & Zavolan, 2014) validated hundreds of miRNA-mRNA pairs. Furthermore, large-scale experimental settings such as PAR-CLIP (Hafner et al, 2010), AGO-CLIP (Wen et al, 2011), IMPACT-seq (Tan et al, 2014) and other CLIP-based protocols provide more direct measures for the miRNAs-mRNAs interactions in living cells (Li et al, 2014a). The generation of miRNA-mRNA chimeras by CLASH (Helwak et al, 2013) has exposed the abundance of non-canonical binding sites and a large number of coding-region interactions. Further work that involves miRNA target identification by measuring the effect of miRNA on protein synthesis (Selbach
et al., 2008). Unfortunately, many of the above protocols suffer from low coverage and poor consistency (discussed in (Chi et al, 2012)). Hundreds of novel miRNAs (Kozomara & Griffiths-Jones, 2013) have led to further expansion in the number of miRNA-targets predicted pairs. Current estimates suggest that in reality only ~60% of the human coding genes are regulated by miRNAs in a cellular context (Ha & Kim, 2014; Jonas & Izaurralde, 2015). Thus, the potential of many predicted miRNA-mRNA pairs to effectively regulate gene expression remains questionable (Betel et al, 2010; Helwak et al, 2013; Seok et al, 2016). A picture emerges in which poorly expressed miRNAs are not part of the post-translational regulation of gene expression (Hausser & Zavolan, 2014).

The specifics of the miRNA-mRNA regulation in a particular cell type depend on the amounts and concentration of miRNAs and their stoichiometry. A key role in this regulatory mechanism is due to the availability of AGO protein, the catalytic component of the RNA silencing complex (RISC) (Janas et al, 2012; Wen et al, 2011). From the mRNA perspective, the number of molecules, the specific expressed variants and the positions of MBS along the relevant transcript (Jens & Rajewsky, 2015) dictate the potential of miRNA interaction, but not necessarily the potential for successful gene expression attenuation (Agarwal et al, 2015). In vivo, the miRNA regulatory network and its dynamics is part of larger circuits that include master regulators and transcription factors (Bisognin et al, 2012; Friard et al, 2010). The properties of the network call for an unbiased probabilistic model to the end of improving predictions (Nam et al, 2005).

A quantitative view of the miRNAs network in living cells challenges the level of competition, as formulated by the ceRNA hypothesis (Denzler et al, 2014; Salmena et al, 2011). Accordingly, an overexpression of MBS-rich molecules of RNA may displace miRNAs from their primary authentic targets (Denzler et al, 2016; Tay et al, 2014), resulting in an attenuation relief of specific mRNAs. As such, the availability of free and occupied MBS (Arvey et al, 2010), concentrations and binding affinities are major determinants in governing the new steady state in a cells. The result of such a competition is an interplay between direct and indirect effects on gene expression (Yuan et al, 2015a). Under this scheme, the functional potency of pairing of miRNAs to their targets is driven by the degree of complementarity (Seok et al, 2016). It was furthermore postulated that many weak sites contribute to target-site competition without imparting repression (Denzler et al, 2016). The dynamics of the miRNA-target regulatory network in view of direct and distal ceRNA regulation had been modeled (Nitzan et al, 2014).

In this paper, we describe a quantitative stochastic model that challenges the cell steady-state in view of alteration in miRNAs’ abundance. The model operates at the cellular level and compares the overall trend of miRNA regulation in various cell lines. We have tested the validity of an iterative cell simulator to
correlate with results extracted from experimental data, under the condition of transcriptional arrest. Based on a probabilistic exhaustive list of all miRNA-target pairings, we configure an iterative stochastic model and test it in view of an exhaustive set of miRNA overexpression manipulations. We show that genes that belong to the translational machinery and function in the translation process are in general unaffected by overexpression of the majority of miRNAs, while a small group of cell-specific gene set is extremely sensitive to the regulation across most miRNAs. We systematically analyzed the design principles of the miRNA-mRNA regulation in various cell types. We confirm that the stochastic nature of miRNA regulation imparts an unexpected robustness of the miRNA regulation in living cells.

Results

miRNAs stability and decay rate of mRNAs upon transcriptional arrest

The nature and extent of miRNA regulation in living cells is depicted by the absolute quantities, composition and stoichiometry of the main players of the network, i.e., the miRNAs and mRNAs (Arvey et al, 2010). The goal of this study is to model the outcome of the miRNA-mRNA network under a simplified setting of transcriptional arrest, where synthesis of new transcripts (miRNA and mRNAs) is prevented. To this end, we first tested the relative changes in the quantities of miRNAs and mRNAs in HeLa and HEK293 cell-lines in the presence of the transcriptional inhibitor Actinomycin D (ActD, Fig 1A). Overall, we mapped 539 and 594 different miRNAs expressed at time 0 in HeLa and HEK293 cells, respectively (Fig 1A). In addition, prior to ActD treatment, 16,236 and 16,463 different expressed mRNAs (not including miRNAs) were mapped from HeLa (Dataset EV1) and HEK293 cells (Dataset EV2), respectively.

We tested the composition of miRNAs and mRNAs along 24 hrs post ActD treatment. The amount of miRNAs after 24 hrs from the application of the drug is practically unchanged. High correlations between miRNA expression of the untreated (t = 0 hr) and ActD treated for 24 hrs (Fig 1B) are evident for HeLa (Spearman rank correlation, r= 0.94) and HEK293 cells (Spearman rank correlation, r= 0.97). In contrast to the stability of miRNAs, the number of mRNAs types monotonically declines, in accordance with the effect of ActD on the bulk of short lived mRNAs (Fig 1A). A maximal variability in the profile of mRNAs is measured between 0 hr and 24 hrs for HeLa (Fig 1B, Spearman rank correlation, r= 0.84, top right) and HEK293 cells (Spearman rank correlation, r= 0.88, bottom right, right). The correlations of the pairs of all time points for HeLa are shown in Appendix Fig S1, and for HEK293 in Appendix Fig S2.
Fig 1C follows the change in the expression level of individual genes in HeLa cells along 24 hrs from ActD treatment. All expressed genes were normalized according to the basal condition (i.e., 100% expression at 0 hr). The change in the abundance of each mRNA, at each time point was quantified relative to the abundance at the starting time point. As > 50% of the identified mRNAs are expressed at a very low level (Dataset EV1), we only report on the percentage of retention for genes that are expressed above a predetermined threshold (total 860 genes, Fig 1C). We illustrate how the distribution of the retention level (in %) of these 860 genes vary between two different time points (Fig 1D). The calculated average retention rate 8 hrs after ActD treatment is ~83%, and decreases to 53% after 24 hrs. These results validate that the decay rate for most mRNAs is a gradual process that continues for 24 hrs. However, the trend for the attenuation in individual gene is already achieved by 8 hrs from the initiation of transcription inhibition.

**Figure 1. Expression profiles of miRNA and mRNA under transcription arrest.** A. Counting of miRNA miRNAs (left) and mRNAs (right) for the 4 different time points for HeLa (top) and HEK293 (bottom). The samples were collected at 0 hr, 2 hrs, 8 hrs and 24 hrs following transcription inhibition by ActD. Source data is available in Dataset EV1 (HeLa) and Dataset EV2 (HEK293). B. Expression of miRNAs (left) and mRNAs (right) in pairs of 4 different time points for HeLa (top) and HEK293 (bottom). The samples were
collected at 0 hr, 2 hrs, 8 hrs and 24 hrs following transcription inhibition by ActD. The expression is presented by logarithmic scale (log_{10}). Spearman correlation (r) is listed for each pair along with the p-value of the significance. Source data is available in Dataset EV1 (HeLa) and Dataset EV2 (HEK293). C. Relative abundance of each expressed mRNAs at four different time points for HeLa cells. At time 0, the relative abundance is set to 100%, and at each proceeding time points the abundance relative to time 0 is reported. Each line in the graph represent one gene (mRNA). Only genes with a minimal expression level of 0.02% expression are listed (equivalent to 97 fpkm, total of 860 genes). The bolded blue line represents the average of all reported genes at each time point. D. Compilation of mRNA retention distribution (PDF) of all the reported genes after 8 hrs and 24 hrs from initiation of transcription inhibition by ActD. All genes with a retention level ≥100 are combined (at 100% retention).

**Overexpression of specific miRNAs impacts targets and non-target mRNAs**

Fig 2 shows the results of direct and indirect effects of overexpressing hsa-mir-155 under the setting of transcriptional arrest. HeLa cells were transfected with the individual miRNA, and the effect of miRNA and mRNA composition was tested 24 hrs after cell exposure to ActD. We quantified the effect of hsa-mir-155 by considering its predicted targets. Specifically, for each miRNA, we split the list of all expressed genes to targets and non-targets using the high-quality TargetScan 7.1 prediction table (Agarwal et al, 2015) (see Materials and Methods).

Retention rates of all genes relative to their starting point (i.e., overexpressed miRNAs prior to the transcriptional arrest) are shown (Fig. 2A). The average decay rates of the direct target genes (Fig 2A, pink thick line) and non-target genes (Fig 2A, blue thick line) for hsa-mir-155 shows that the decay of direct targets is slightly faster compared with the rest of the non-target set (Fig 2A, right panels).

The significance of the differences in the decay rate after 24 hrs of ActD treatment on HeLa cells for cells overexpressing hsa-mir-155 indicates faster degradation and an overall lower retention for hsa-mir-155 targets in transfected vs. naïve cells (Fig 2B, upper panel). The observed shift in the relative mean statistics (Fig 2B) of direct targets is below significance (p-value = 0.12). More significant is the shift in the higher retention rates for the non-target genes (p-value 0.002, Fig 2B, compare solid and dashed line). This implies a certain degree of indirect stabilization of hsa-mir-155 non-target genes as a result of overexpression of hsa-mir-155. The same trend in the retention profiles was observed by analyzing HeLa cells overexpressing hsa-mir-124a in HeLa cell (Appendix Fig S3).
These results argue that under the described experimental settings, indirect effects are anticipated presumably due to a competition on MBS and the continuous changes in the stoichiometry of the key players which affect the probabilities of miRNA-target interactions.

Figure 2: Retention profile of mRNAs following overexpressing miRNAs in HeLa cells. A. Relative mRNA retention in HeLa cells that were transfected and overexpressed with hsa-mir-155. Measurement were taken at 4 time points as indicated. The retention plots are partitioned to target genes (i) (pink, left panels) and (ii) non-target genes (blue, middle panels). In (iii), the average retention patterns for of hsa-mir-155 targets (pink line) and non-targets (blue line) are shown. B. Distribution of genes retention after 24 hrs from ActD treatment, according to their labels as targets (upper panel, pink) and no-n-targets (lower panel, blue). The plots compare the retention of genes from the control (smooth line), and from hsa-mir-155 overexpressed condition (dashed line). The number of genes that are included in the analyses are shown in parentheses. Target genes are marked by pink lines (top) and the non-target genes by blue lines (bottom). Note the shift in the distribution in the non-target genes towards the genes with higher retention level. All genes with a retention level ≥100 are shown as 100% retention.
Assessing the probabilistic approach for miRNA - mRNA interactions

The experimental results (Figs 1-2) emphasize the need for a systematic analysis of the miRNA-mRNA interaction network acting under quantitative and stoichiometric constrains in living cells. It is evident that even under transcriptional arrest paradigm the high levels of overexpressed miRNA do not solely affect their direct target genes.

Evidently, the molecular interactions of miRNA and mRNA within the cell is a stochastic process. The specific composition of miRNA and mRNA in cells, and their binding probabilities dictate the effectiveness of attenuation of gene expression. The miRNA-target prediction of TargetScan provides a sparse table of miRNA-MBS interactions and reports on 1,183,166 such pairs (see Materials and Methods). Each miRNA-MBS interaction is associated with a probabilistic score that is a proxy for the level of confidence for that interaction, and can be considered the probability of effective binding for any specific pair.

To further investigate the properties and the design principles of the miRNA-mRNA interaction network, we developed an iterative simulator called COMICS (Competition of miRNAs Interactions in Cell Systems). Fig 3A illustrates the main flow in a single iterative cycle in COMICS. The probabilistic framework relies on a constant update of the cell state which is defined by the amounts and composition of miRNAs and mRNA types, and the balance of occupied and free molecules. COMICS iterations capture the stochastic process that takes place in living cells that are subjected to miRNA regulation.

Fig 3B is a breakdown of COMICS process according to the fate of the regulated mRNAs. Specifically, it shows the sampling process driven by the composition of miRNAs and mRNAs and their measured amounts (Fig 3B, pink frames). Recall that the measured expression profiles of miRNAs and mRNAs is cell-type specific as evident from the list of mRNAs and miRNA from HeLa (Dataset EV1) and HEK293 (Dataset EV2). Each mRNA is characterized by the types and positioning of its MBS at the 3'-UTS of the transcript. The interaction prediction table is associated with a probability-based scores for any specific pairs of miRNA and MBS in the context of a specific mRNA. In each iteration, a miRNA is sampled randomly, according to the cell’s miRNA abundance and composition. Next, one of its target genes is chosen randomly according to the measured expressed mRNAs distribution. In the following randomized step, the chosen miRNA and its target may possibly bind according to the reported probability. Following an actual binding event, the distribution of the miRNAs and the mRNAs are updated accordingly (Figs 3A-3B). The status of the mRNA following a successful pairing is changed (i.e., marked as prone to degradation).
The new status of mRNA as ‘occupied’ does not prevent it from engaging in additional subsequent miRNA binding. For MBS that are in close proximity to each other, a minimal spacing is required, otherwise, the interaction will be prevented. The occupied mRNA is marked for degradation with some delay that mimics the likely instance of a cooperative binding on a target by multiple miRNAs prior to its degradation. Based on the validated stability of miRNAs (Fig 1B), once the occupied mRNA is removed, all miRNAs that were bound to it are relieved and return to the free miRNA pool. As a result, the stoichiometry of miRNA to mRNA is gradually changing with an increase in the apparent ratio of miRNAs to free mRNAs in the cell.

Fig 3C shows the result of COMICS following one million iterations on HeLa cells. Note that following 1M iterations, the average retention of mRNAs is 43.5%, similar to the decay rate observed in living cells (Figs 1-2). Fig 3C shows the decay rate of 755 genes (with expression threshold (>0.02%). The output of the mRNAs along the 1M iteration run is reported in Dataset EV3.

We challenged the validity of COMICS to capture the pattern of mRNA downregulation upon transcriptional arrest using the results from living cells. We applied two complementary tests: (i) Detecting a correlation between genes that were unoccupied by miRNAs (i.e., all genes that remain available following 24 hrs of ActD treatment, Fig 3D). (ii) Scoring the correspondence of genes that were occupied along the COMICS iterative run (100k iterations, Fig 3E) and their correspondence with published results from CLASH performed on HEK293 (Helwak et al, 2013). For the first test, we compare the COMICS performance with the experimental results considering a large set of genes that display high level of retention (>85%). This high retention set includes 122 genes in HeLa and 158 genes in HEK293, respectively. We found that overall, the corresponding score (Jaccard score) is higher for HEK293 cells (Fig 3D). However, the statistical significance is maximal for 100k iteration in the case of HeLa cells (hypergeometric p-value of 0.00064) with high significance for matching results from COMICS simulator and the experimental results (Fig 3D).

Using the cell state data for HEK293 along the simulation process (up to 100k iterations), the overlap with the results presented by CLASH (Helwak et al, 2013) is highly significant. We compared all the pairs that were identified in CLASH experiment (Helwak et al, 2013) and expressed above a pre-determined threshold in HEK293. The analysis on HEK293 indicates that as the number of the iterations increases, the overlap of the validated set of pairs increases, and remain highly significant throughout the process (Fig 3E, hypergeometric p-value 0.0014). Most importantly, the correspondence of the results of COMICS to the pairing observed by the CLASH methodology, is strongly dependent of the use of TargetScan miRNA-MBS probabilistic interaction table. Applying two randomization modes (see Materials and Methods and

59
Appendix Text EV1) for the miRNA-MBS interaction table, has led to a drop in the correspondence of the simulator and CLASH reported results (Fig 3E).

We further tested COMICS for mimicking the competition of miRNAs and the trend of mRNAs regulation in living cells. This was performed by testing the sensitivity and the robustness of the simulation results with respect to changes in a broad range of parameters (Appendix Text EV1). Importantly, changing the binding probabilities from the TargetScan interaction table (see Materials and Methods) by two randomization procedures has drastically reduced the correlations at the end of the simulation run. For details see Appendix Text EV1. Additionally, extensive changes in the sets of the operational parameters used by COMICS were tested. Overall, the output of the different simulation runs was consistent and implies the robustness of the system to a wide range of parameters (see Appendix Text EV1).

To account for the variation in the result due to the stochastic nature of the samplings (Fig 3B) along the simulation, we repeated 3 simulation runs. We show that without changing any of the parameters, the final retentions of the genes after 100k iterations in each simulation run are highly correlated (Pearson correlations 0.86-0.88, Appendix Fig. S4).
**Figure 3: Scheme of the COMICS platform and performance of the simulation process.**

A. A schematic view for a single iteration of COMICS simulation. The iteration step is repeated 100k times (unless indicated differently). After each successful interaction step (i.e. a valid binding of miRNA with MBS on mRNA molecule), the entire distributions of the miRNA and mRNA in the cells are updated. Therefore, the next iteration is slightly changed due to the refinement of mRNA composition and the availability of the free pool of miRNAs. On average, only 3% successful iterations occur from a total of 100k sampling events. The input for the simulator varies according to the profiles of the cell-type under study.

B. The outline of the major steps of COMICS operation from the mRNA perspective. The composition of the miRNA in the cells is obtained from the experimental measurement (at time point of 0 hr) but normalized to have 50k miRNAs and total of 25k mRNA. For HeLa cells, these are 3666 types of mRNAs that are included in the analysis (i.e. above a minimal threshold, see Materials and Methods). Sampling of the miRNA and mRNAs are done...
according to their distribution and the probability of the interaction is driven from the score of TargetScan MBS interaction scores (with 1.2 M values). A dashed mRNA shown after N iterations signifies an occupied transcript that is still halted (e.g., delay of 1k iterations) prior to its degradation, and releasing the bound miRNAs to the pool. In the end of the simulation, two sets of mRNAs are considered, an occupied set that is validated against miRNA-MBS pairs from CLASH data and the unoccupied mRNA set that is tested in view of the experimental results under transcriptional inhibition scheme (Fig 1-2). C. The retention of HeLa expressed genes along COMICS simulation run. In this setting, 1M iteration steps were performed. For this input of COMICS simulation on HeLa cells, 3666 types of mRNA and 110 of miRNAs are included in the input. These numbers account for the 50k and 25k molecules of miRNAs and mRNAs, respectively. Each grey line represents the retention profile of a single type on mRNA. The blue line shows the average retention profile. After 1M iterations, the average retention level is 43.5%, Altogether, for 1M iterations, 11k successful events had been reported. For clarity of presentation, only mRNA above a predefined expression are presented (see Materials and Methods). Total of 755 genes with a minimal expression of 0.02% are shown. D. Validation of COMICS performance in view of the results from transcription arrest in HeLa (grey) and HEK293 cells (orange). At each of the indicated steps of the COMICS simulation run, the overlap in gene retention for the set of genes that remain stable (defined as >85% retention) was measured by calculating the Jaccard score. The statistical significance associated with the correspondence of the results (measured by p-value of the fisher exact test) are indicated by asterisks * <0.05 and **, <0.005. E. Testing COMICS performance and dependency on the information in TargetScan interaction matrix. COMICS simulation performance in HEK293 was compared to the bounded pairs as reported from CLASH data on HEK293 (see Materials and Methods). The histogram shows the performance in term of the significant of the overlap of the reported COMICS results (100k iterations) using TargetScan probabilistic converted matrix (grey), and two versions of randomization for the interaction table (see Materials and Methods). The statistical test was based on the 251 genes that are reported as pairs miRNA-mRNA pairs by CLASH and expressed above the minimal expression threshold used for COMICS simulation protocol. The use of the TargetScan matrix shows significant results versus CLASH data (at the significant range p-value of e-4 to e-6). Applying any of the randomization for the miRNA-MBS interaction table, caused a drop in the performance to non-significant values.

Alteration in miRNA profiles but not the mRNAs determines cell identity

The gene expression profiles for miRNAs and mRNAs of the 3 tested cell-lines are quite different (Fig. 4A and Dataset EV5). HeLa, HEK293 and MCF-7 cells are representatives for fibroblasts, cells from kidney and breast origin, respectively. We compared the dominant miRNA profiles as processed by COMICS for these cell-types (Fig 4A). The difference in evident in comparing the fraction and composition of the miRNAs that occupy 90% of the total amounts of miRNA in each cell (i.e. 45k out of 50k molecules of miRNA per cell). The unified set for all three cell line miRNAs is shown (total 28 miRNAs). Differences among cells is evident from the presence or absence of specific miRNAs (hsa-let-7, hsa-mir-16, Fig 4A). More significant
is the large deviation in expression of miRNAs in a specific cell (e.g., hsa-mir-21), and the fraction that is occupied by a small set of dominating miRNAs.

To further test the impact of the unique miRNA profile for the cell identity, we applied COMICS but activated the simulation process using input extracted from the miRNA profile of one cell type and tested it in the context of mRNA composition of the other cell. We repeated such artificial settings for all 9 combinations (of which 6 are artificial) as shown in Fig 4B. In all the settings, the outcome of the retention profile fully resembles the profile of the cell providing the miRNA profile. For example, by including in COMICS input the data of miRNA profile from HeLa (according to Dataset EV5), the correlation of the retention profile in the end of the simulation run is in the range of 0.9 for input of the mRNA of any of the other cell types. On the other hand, the input of miRNA of MCF-7 or HEK293 using the authentic mRNA profile resulted is a Pearson correlation $r=0.53$ and $r=0.46$, respectively. A similar trend is observed for the HEK293 that shows a high correlation of $r=0.93$ and $r=0.9$ when applied on the background of mRNA from HeLa or MCF-7, respectively (Fig 4B).

**Figure 4. Abundant miRNA and shuffling cell-specific miRNA profiles.** A. Heatmap of miRNA from HeLa, HEK293 and MCF-7 cell-lines in view of the abundance of each miRNA (color in log scale) in each cell-type. The joint list of miRNAs includes the most abundant miRNAs that occupies 90% of the miRNA molecules (i.e. 45k out of 50k) in each cell-type. The fraction occupied by each of the listed miRNA, for each of the cell-type is available in Database EV5. B. Pairs of miRNA and mRNA are shown according to their origin. Pearson correlation of the endpoint of the genes following 100k iterations and testing any of the genes that are above the minimal expression (5 molecules, 0.02% of mRNA). The number of mRNAs that are considered in the analyses are the 516 to the pair of HeLa and HEK293; 285 for the pair of HeLa and MCF-7 and 305 to the pair of HEK293 and MCF-7. The p-value of the correlations are very significant for all pairs. All p-values correlation values are $<1\times10^{-15}$.  

63
Simulating miRNA overexpression by COMICS reveals stabilization of non-target genes

The COMICS system was used to simulate exhaustive overexpression experiments while testing explicitly the effect of miRNA-mRNA probabilities on the mRNA decay. We activated COMICS by manipulating the abundance of hsa-mir-155 from the naive cell state (x1, no overexpression) to another 7 levels for hsa-mir-155 amounts (x0.5, x3, x9, x18, x90, x300 and x1000). The addition of miRNA molecules is considered in a probabilistic framework. Thus, in practical terms, overexpression of a single specific miRNA changes the distribution of all other expressed miRNAs (Fig 5A). The probability of miRNA and target to engage in a successful interaction is dependent on the interaction strength converted to the miRNA-MBS scoring (Fig 3B). The scheme in Fig 5A shows that at an overexpressed factor x300 for hsa-mir-155, almost 20% of all miRNAs are occupied by this miRNA and it reaches almost 50% at the extreme overexpression level (x1000). It is important to note that the calculated fraction of miRNAs in the cell following overexpression is an immediate reflection of its original abundance in the naïve cell. These amounts are characteristic for the different cell types.

Fig 5B shows the gradual change in the retention of each gene (above a predetermined threshold) along 100k iterations of COMICS simulation process. Each panel shows the results of the simulation that starts from the abundance of hsa-mir-155 in the naïve cell (control). Fig 5B show the gradual alteration in the dynamic of the gene retention following the increase in different overexpression levels (determined by the multiplication factor).

We found that the final retention level is sensitive to the overexpression factor (Fig 5C). In the case of hsa-mir-155 in HeLa cells, elevating the miRNA from x18 to x90 caused a drop in the average retention of its targets followed by a steep drop in retentions at overexpressing factor of x1000. A minor but consistent increase in the retention of non-target genes is observed along the elevation in the overexpression level for of hsa-mir-155.
Figure 5. miRNA overexpression paradigm using COMICS platform. A. Schematic The relative percentage of each miRNA abundance is separated by vertical line. The miRNAs percentage is sorted from lowest (left side) to highest (right side). Seven different hsa-mir-155 over expression simulations are shown (x1, x3, x9, x18, x90, x300 and x1000) from bottom to top. The percentage of hsa-mir-155 is marked in pink. B. Over expression simulation of hsa-mir-155 in eight different factors, where hsa-mir-155 target genes are marked in pink and all other non-target genes are marked in blue. The average retentions of both gene groups are plotted in bold lines. C. Average final retention of the different simulation runs, using different overexpression factors of hsa-mir-155 as shown in B, of hsa-mir-155 target and non-target genes.

A unified pattern of mRNA retention is associated with overexpression of miRNAs

To determine whether the composition and stoichiometry of miRNAs and mRNA are dominant features in determining the design principles of many cell states, we performed an exhaustive and systematic manipulations of cellular miRNAs. We applied COMICS simulations on overexpression experiments for miRNA families. For HeLa cells, 248 miRNAs were compiled and match their representation in the miRNA-MBS TargetScan prediction table (some miRNA are identified as miRNA family). We multiply the basal abundance (x1) of each of these miRNA families by several factors: x3, x9, x18, x90, x300 and x1000. For each such factor (f), a final retention table was computed. The result for HeLa cells is a table with 773 genes (rows) whose initial expression level exceeds a pre-determined threshold, and 248 miRNA (columns). Each of these miRNAs was overexpressed by the tested factor. Therefore, each cell in the
matrix $M_{fi}$ is the final retention of gene $i$ after 100k iterations of COMICS for the overexpressed experiment of miRNA $j$ (Fig 6A, Dataset EV6). For unexpressed miRNA, a standard minimal level of expression is defined as $x_1$ level (see Materials and Methods).

Inspecting $M_{fi}$ for each overexpression condition reveals the presence of a substantial set of genes that are characterized by final retention above 85% for $\geq 90\%$ of the tested miRNAs in the overexpression experiments (i.e., high retention criterion satisfied by at least 225 or 248 tested miRNAs). We refer to them as cross-miRNAs stable genes. We found such stable genes for Hela (185 genes), HEK293 (176 genes) and MCF-7 (124 genes). For a full detailed analysis of cross-miRNAs stable genes see Dataset EV7. These results imply that a set of genes in each cell type are resistant to gene expression attenuation regardless of which miRNA is overexpressed.

The matrix $M_f$ also reveals a small defined gene set that is globally sensitive to miRNA regulation. Specifically, these are genes with retention rate below 50% for $\geq 90\%$ of the tested miRNAs among all overexpression experiments. These genes are referred to as cross-miRNAs sensitive genes. We report on these sensitive genes for Hela (23 genes), HEK293 (34 genes) and MCF-7 (22 genes). For a full detailed analysis of cross-miRNAs sensitive genes see Dataset EV8. These results imply that a set of genes in each cell type are sensitive to gene expression attenuation regardless of which miRNA is overexpressed.

The retention matrices $[M_{fij}(x_{300})]$ for HeLa (Fig 6B) and HEK293 (Fig 6C) are colored by high retention (red) and low retention (blue) levels. The matrices represent clustering by genes and miRNAs. Analyzing the matrices by the clustering dendrogram shows that the set of sensitive genes (cluster of blue rows) in both cell types is distinguished from all other genes. However, the richness of the retention pattern is demonstrated by zooming at any section of the $M_{fij}(x_{300})$. It is also evident that some miRNAs are naturally clustered by a similarly profile of their columns. We tested whether the type of the cell dictates the characteristic of the retention profiles, and whether cells that were manipulated by overexpressing miRNAs established a new cell states that are different from their naïve states. Fig 6D and 6E compare the average retention observed for each of the shared genes in the HeLa and HEK293. Large difference is observed in the distribution of genes for HeLa and HEK29 when the profile in $M_{fij}(x_{1})$ and $M_{fij}(x_{300})$ are compared. This global view suggests that each manipulated cell converges to it unique cell state and potentially reflects the different sensitivity of the analyzed genes to the unique cell specific miRNA composition.
We analyzed the specificity of the unified pattern for Mf$_i$ that was found for HeLa, HEK-293 and MCF-7. We tested the overlap in the resulting sets of cross-miRNAs stable genes and cross-miRNAs sensitive genes from these cells (Fig. 7A). In the present analysis we consider only genes that are common to all three cell types and are expressed at >0.04%. Of those, the numbers of cross-miRNAs stable genes that appear in all three cells are 78 (MCF-7), 102 (HeLa) and 110 genes (HEK-293). Of these, 48 genes are common to all three types. The overlap of this number of genes is statistically very significant (chi-square test p-value 1.35e-08, Fig 7A). It argues that these stable genes are designed to resist miRNA regulation under a wide range of overexpression settings and across cell types. The list of these 48 genes that are shared by all three cell lines is shown in Dataset EV9.

**Cross-miRNAs sensitive genes are cell specific**

We next carry out a similar analysis for cross-miRNAs sensitive (Dataset EV8). These sets are much smaller with 6 (MCF-7), 12 (HeLa) and 17 genes (HEK293) and no sensitive genes are common to all three cell types (Fig 7B). We conclude that these cell-specific sensitive genes are exhibit an exceptional sensitivity to miRNA regulation.

We analyzed the cross-miRNA sensitive lists from the different cell lines using annotation statistics tools (see Materials and Methods). This analysis shows only a moderate enrichment for terms associated with mRNA regulation and processing. For example, among the 23 sensitive genes from HeLa (Fig. 7B) numerous genes are related to transcription processing and modifications, other annotations are for RNA poly-adenylation, and RNA-related metabolic processes (Dataset EV10). Among the sensitive genes are PCBP1 and PCBP2, single-stranded nucleic acid binding proteins that were identified in an mRNP complex, and play a role in viral internal ribosome entry site (IRES). Another small set of proteins include major pre-mRNA-binding proteins from the heterogeneous nuclear ribonucleoproteins family (HNRNPK and HNRNPA1). HNRNPA1, involved in the packaging and transport of poly(A) mRNA from the nucleus to the cytoplasm. It was recently proposed that HNRNPA1 binds specific miRNA hairpins and acts in miRNA biogenesis (Treiber et al, 2017). Inspection of the sensitive gene set from MCF-7 (22 genes) shows borderline statistical significance for regulation of transcription from RNA polymerase II promoter. HEK293 (34 genes) resulted in enrichment for translational processes (including mitochondria and viral translation). We conclude that in each cell type a small set (22-34 genes) that is extremely sensitive, but there is no unified functional coherence among these cell specific sets.
**Cross-miRNAs stable genes are enriched in the translation machinery**

We applied annotation enrichment tools (see Materials and Methods) to the set of cross-miRNA stable genes from HeLa cells (185 stable genes, Dataset EV7). This analysis reveals that they are extremely enriched by terms associated with many aspects of translation, including translational elongation (GO:0006414), mitochondrial translation (GO:0032543), SRP-dependent cotranslational protein targeting to membrane, translational termination (GO:0006415) and more. These annotations are associated with highly significant statistics (corrected FDR p-value = ~1*e-77). Clusters of enriched annotations were compiled using the DAVID gene expression tool. The score of the top clusters is 53, depicting e-53 as the average enrichment for all annotations in that cluster. The top clusters are scored by tens of coherent annotations, all related to the structure of the ribosome, elongation machinery and fidelity of translation. For example, the annotation of translational elongation (GO:0006414) shows an enrichment of 43.5 with FDR p-value of 5.53e-87. The results of the most significant annotation clusters are shown in Dataset EV10.

A similar significant enrichment remains as we apply the statistical tests using more specific background lists (e.g., the set of HeLa expressing genes from Dataset EV1 and the set of genes that appears in the analyzed Mfij (Dataset EV6). Significant enrichment for clusters of annotation for protein translation and the translation machinery was duplicated for any of the cross-miRNA stable gene lists derived from HEK293 (176 genes) and MCF-7 (123 genes), Dataset EV8), with a DAVID clustering annotation scores of the top cluster of 56.5 and 39.9, respectively. The strongest enrichment of the annotations terms for tested cell-types, extends to genes participating in translation initiation, elongation and co-translation to the ER membranes.

Fig 7C shows the partition of 48 shared genes that are common to all three cell lines (Dataset EV9) by their protein function. The dominant role of translational machinery (DAVID clustering enrichment score 49.4) is shown by the enrichment corrected score for translation elongation and cytosolic ribosome (FDR p-value of 1.18e-67 and 9.36e-60, respectively, Dataset EV9). Translational machinery component with small and large subunits (35 genes), elongation factors (EIF4A1, EEF1D and EEF1B2) and nucleolin (NCL) account for 79% of this list and are key factors in ribosomal production and its function. The list includes also actin, myosin and tubulin (total 4 genes) that are major cytoskeletal components essential to cell shape and physiology.
Figure 6: A scheme of the over-expression matrix by a single factor. A. The different columns stand for the different pre-miRNA over-expressed by factor $f$, and the rows stand for the different genes. B. Heatmap of the range of the retention for genes that were overexpressed at a factor $x_{300}$. Each row is associated with a miRNA. The clustering is performed by the row (i.e. genes). The matrix includes 250 expressed miRNAs in HeLa cells. C. Zoom in of a small section of the heatmap of the range of the retention for genes that were overexpressed at a factor $x_{300}$. Each row is associated with a miRNA. The clustering is performed by the row (i.e. genes). The matrix includes 250 expressed miRNAs in HeLa cells. D. HeLa and HEK293 average final retention comparison in control simulation (no over expression). Each point stand for each gene in all 248 overexpression condition (each row in the heatmap presented in B, C) using over expression factor $x_{300}$. E. HeLa and HEK293 average final retention comparison. Each point stand for each gene in all 248 overexpression condition (each row in the heatmap presented in B, C) using over expression factor $x_{300}$.

We conclude that the cross-miRNA stable gene set signify the translational machinery. Thus, the translational machinery highlights a functional gene set that are immune to the regulatory layer of miRNAs. This observation applies to all tested cells.

Finally, we tested the properties that characterize genes associated with the cross-miRNA stable and sensitive genes in view of all genes that participate in the simulation process (Dataset EV7 and Dataset EV8). Four properties were tested: (i) the number of targeting miRNAs (Fig 7D), (ii) the number of MBS (Fig 7E). The other two features are: (iii) the initial expression level (Fig 7F) and (iv) the binding potential
according to the expression of the most dominant miRNAs (Fig 7G). Features (iii-iv) are cell-type specific. The detailed t-test statistics of the results for all three cells are provided in Table EV10. The genes in the stable set are characterized by having fewer MBS and fewer targeting miRNA relative to the other (i.e., not stable and not sensitive) genes (t-test 6.48E-21 and 5.64E-15, respectively). The significance of the statistics for the initial expression of these genes is marginal in all cell types. However, the most significant feature that differentiate the stable from the sensitive set (t-test 2.52E-22, Fig. 7G) is associated with the immunity of the stable set to be targeted by the most abundant miRNA (based on the miRNA list in Fig 4A). Therefore, the stable genes are unlikely to be effectively regulated by any of the most abundant miRNAs. This trend is evident from the statistics for the three analyzed cell lines (Table EV1).

The analysis for the 4 quantitative features was replicated using the sensitive and stable dataset from HEK293 and MCF-7 (Appendix Fig. S6). For example, the average number of MBS for a stable gene is 24.2 (vs. 62.5 for the sensitive gene set) in HEK293 cells. The most abundant miRNA expressed in HEK293 is hsa-mir-7 (25% of all miRNAs) targets only 3.5% of the stable genes but 94% of the genes among the cross-miRNA sensitive gene set have a MBS.

We conclude that a comparison between the stable and sensitive genes reveals a signal for MBS evolution. Despite the great difference in overall miRNA composition of cell lines (Fig 4A), several of the miRNAs are shared across many cell-types (e.g., hsa-mir-21, hsa-let-7 and hsa-mir-92). The stable genes are critical gene of the translational machinery, are mostly excluded from having MBS that could engage in binding of a dominant miRNA players in multiple cells.
Figure 7: Comparison of sensitive and stable gene sets in different cell type. A. Overlap of the cross-miRNA stable genes in HeLa, HEK293 and MCF7 cells. Only genes that are expressed in at least two cells are listed. The gene list of the stable genes is available in Dataset EV6. B. Overlap of the cross-miRNA sensitive genes in HeLa, HEK293 and MCF7 cells. Only genes that are expressed in at least two cells are listed. The gene list of the stable genes is available in Dataset EV7. C. Partition of the stable genes expressed in HeLa, HEK293 and MCF-7 cells to their functional annotations: (i) small ribosomal subunit (18 genes), (ii) large ribosomal subunit (17 genes), (iii) cytoskeleton (5 genes), (iv) translation elongation (3
genes) (v) 4 additional genes – NCL, ATP5J2, CALR and MIF. For detailed list see Dataset EV8. D. Comparison of the number of targeting miRNA of sensitive genes, stable genes, and other (not sensitive and not stable in HeLa cells. Statistics of the comparisons are significant for the comparison of stable genes set and both sensitive and other gene sets (ttest p-values of 7.53e-11 and 6.48e-21, respectively), and no significant difference between sensitive and other gene sets. Full statistics are shown in Table EV1. E. Comparison of the number of MBS of sensitive genes, stable genes, and other (not sensitive and not stable in HeLa cells. Statistics of the comparisons are significant for the comparison of stable genes set and both sensitive and other gene sets (ttest p-values of 2.07e-9 and 5.6e-15, respectively), and no significant difference between sensitive and other gene sets. Full statistics are shown in Table EV1. F. Comparison of the initial abundance of sensitive genes, stable genes, and other (not sensitive and not stable in HeLa cells. Statistics of the comparisons are significant for the comparison of stable genes set and both sensitive and other gene sets (ttest p-values of 0.017 and 0.015, respectively), and no significant difference between sensitive and other gene sets. Full statistics are shown in Table EV1. G. Comparison of the of the average expression of the targeting miRNA of each gene, of sensitive genes, stable genes, and other (not sensitive and not stable in HeLa cells. Significant differences between all three gene sets were found (ttest p-values 2.52e-22, 3.07e-9 and 8.7e-11 for the comparison of stable-sensitive, stable-other and sensitive-other, respectively.

Discussion

miRNAs stability as a major determinant in cell regulation

Cells’ behavior cannot be trivially predicted from direct measurement of their composition of miRNAs and mRNAs (Arvey et al, 2010; Landgraf et al, 2007). Most insights on the regulation of gene expression by miRNAs are based on global observations (e.g., CLIP and CLASH (Li et al, 2013)), or knockdown or overexpression of a specific miRNA, in a specific cell type or tissue (Thomas et al, 2010). Based on many such studies, it was concluded that detailed quantitative considerations of miRNA and mRNA govern the dynamics and the steady state of gene expressed (Bosson et al, 2014; Hauser & Zavolan, 2014). Nevertheless, the underlying rules for post-transcriptional regulation by miRNAs are still missing (Erhard et al, 2014).

We studied cellular outcome following miRNA regulation under a simplified condition of transcriptional arrest, using ActD and focusing on mRNA retention profiles. The post-translation regulation of miRNAs is not limited to mRNA attenuation. However, it was shown that under such condition of transcriptional arrest the dominant effect of miRNAs is via their influence on mRNA stability and not on translation repression (Bethune et al, 2012). As most miRNAs are transcribed by RNA Polymerase II, it was essential
to assess the effect of transcription arrest on the abundance of miRNAs. The results shown in Fig 1B substantiate the notion of an extreme stability of miRNAs at least during 24 hrs from cells’ exposure to the drug. In the time frame of our experiments, AGO-2 is insensitive to transcriptional or translational arrest (Olejniczak et al, 2013).

It is expected that miRNA stability is mostly attributed to AGO-2 that stabilizes the leading strand in the cytoplasm (Winter & Diederichs, 2011). The number of miRNA (and AGO proteins) relative to the number of MBS had been estimated from experimental data (Bosson et al, 2014; Janas et al, 2012). These estimates suggest that the number of AGOs proteins limit the pool of RISC loaded miRNAs which plays a role in gene expression attenuation. AGO-2 is estimated to account for 60% of the AGOs in cells. While the number of AGO-2 molecules is unknown, it was estimated at around ~15k in HeLa cells (Janas et al, 2012) and ~100k in skin tissue (Wang et al, 2012). We consider the miRNA pool that is constrained by the amount of AGOs to be at the order of 50k molecules in a cell. COMICS simulations are started with a molecular ratio of 2:1 ratio between miRNA to mRNA, accounting for the AGO-bound active miRNAs. For the probabilistic formulation of cells under varying levels of miRNA overexpression, the sampling is governed by the recalculated composition of miRNAs, assuming that loading of the leading strand of miRNA on AGO proteins is driven by such distribution of that miRNAs (Fig 5A).

**miRNA composition is a major determinant in establishing cell identity**

We developed the COMICS platform to handle cells that undergo a wide range of miRNA manipulations, and to monitor mRNA retention profiles at the simulation’s endpoint (100k iterations, Fig 2B, Fig 3A). Altogether, thousands of simulation processes were completed to test the impact of altering the expression of hundreds of miRNA families. For example, in HeLa cells, 250 miRNAs (Fig 5) were considered, with each miRNA being altered by increasing factors of overexpression, from the base default level (x1) to x1000 (Fig 5, 8 factors). The same protocol was applied to test miRNA regulation in the other cell types (HEK293 and MCF-7). General trends could be drawn from the results of thousands of COMICS simulations. The main observations hold across most (>90%) of the expressed miRNAs. Specifically, we show that there is a relatively larger set (about 20% of the reported genes) that are exceptionally stable (Fig 7, Dataset EV7). The overlap of these sets of genes in the three cell types is statistically very significant, thus suggesting that the identity of these genes reflects a strong evolutionary signal of stability in the face of extreme increase in almost all miRNAs.

The other phenomenon that COMICS revealed is the presence of a small set of genes (about 3% of the
reported genes) that are extreme sensitivity to down-regulation by almost every miRNA. However, the identity of these genes is cell-type specific. Fig 7A shows the minimal overlap that is found between any pair of cells for the sensitive genes. The strong signal of cross-miRNA behavior suggests that it is the miRNA composition of each cell that dominates the identity of the most sensitive genes, many of which participate in transcription regulation. To further test the impact of the actual miRNA profile on cell identity using COMICS, we ran the simulation on artificial combinations of miRNA and mRNA profiles that come from different cell types. In all these artificial settings, the outcome of the retention profile was more similar to the cell providing the miRNA profile (Fig 4B) and not dictated by the mRNA composition. Our results are in accord with the notion of miRNA profiles as a major determinant for cell identity (He et al, 2012). Indeed, specific profiles of miRNAs are associated with varying malignancy states (Bockmeyer et al, 2011; Volinia et al, 2006). A transition among cell types is attributed to the expression of a specific miRNAs (e.g. miR-34a (Bu et al, 2013)). It was shown that miRNA profile is carefully regulated to promote and stabilize cell fate choices (Shenoy & Blelloch, 2014). In this study we discuss genes that belong to cross-miRNA stable or sensitive sets, however the retention profiles for most genes and along the different overexpression levels provides a rich picture and is far more complex as illustrated (Figs 6A, 6B zoom-in panels).

**COMICS accounts for the effect of the most abundant miRNA**

The stable and sensitive gene sets differ in a statistically significant manner in almost every relevant quantitative feature (Fig 7). As shown for HeLa cells (Figs 6D-6G), the numbers of MBS at the 3'-UTR and the number of different targeting miRNAs are much higher in the sensitive set compared with stable genes. However, the most significant difference between the sensitive and stable sets (p-value = 1.57E-23) concerns the sensitivity to miRNA expression level (Fig. 6F). As shown in Fig. 4A, a small number of miRNAs together comprises 90% of all miRNA molecules in any cell-type. Based on the mass conservation principle, we can expect that highly expressed miRNAs would play a key role in miRNA-mRNA interactions. This principle is indeed implemented in COMICS sampling protocol, and dominant miRNAs are sampled more frequently to interact with their possible targets. Evidently, the probability of a successful interaction for mRNA molecule grows the more MBS that it carries to which some highly abundant miRNAs can bind. Remarkably, and perhaps counterintuitively, it is the stable set has higher initial counts (Fig 6E). This provides additional support to the claim that miRNAs dominate the system, whereas the initial abundance of mRNAs plays only a minor role. The statistical observations regarding the dominant role of the highly expressed miRNAs shown in Fig 7G apply to the other tested cell types. For example, targeting
by hsa-mir-21 is prevalent among the cross-miRNA sensitive genes. The hsa-mir-21 occupies 27% and 32% of total miRNAs in the naïve MCF-7 and HeLa cells, respectively (Dataset EV5). The hsa-mir-21 occupies 58 and 73 MBS among the 22 and 34 sensitive genes identified in MCF-7 and HEK293, respectively (Dataset EV8).

Even in settings with high levels of overexpression, the impact of miRNAs which are highly expressed in the naïve cells, prior to the overexpression manipulation (Fig 4A) remains substantial. Such miRNAs would still constitute a significant fraction of the miRNA cellular pool and are therefore likely to be selected by the probabilistic sampling protocol (see Fig 3A).

Fig 7 offers insights on the properties of the stable genes in the various cell lines (Appendix Fig. S6). Combining the statistical observations (Fig 7D-6G, Table EV1) with the functional annotation of the cross-miRNA stable genes suggests an evolutionary signature for these genes. This evolutionary signal could be detected thanks to the unbiased approach taken by COMICS and the exhaustive testing of different cells under many perturbations by all conserved miRNAs. In all the tested cells, a strong enrichment in translation annotations was shown (including elongation, rRNA metabolic process co-translation to the ER and related functions, Datasets EV9 and EV10). We propose that an evolutionary signal unifies many of the ribosomal proteins and the observed components of the translation machinery. The driving force of evolution acting on targets and their MBS underlying the structure of miRNA network in many organisms (Berezikov, 2011). This evolution signal combines high expression with low density of MBS at the 3’-UTR (Fig. 7E). Even more critically, it relies on the low sensitivity of these mRNAs to the effect of highly abundant miRNAs (Fig. 7G). In cancer cells, the most abundant miRNA genes (e.g., hsa-mir-21, hsa-mir-30, hsa-mir-15/16) are involved in cancer development and even minor changes in their expression govern transition in cell states (Volinia et al, 2010).

We propose that the translation machinery has evolved to maximal robustness vis-a-vis miRNA regulation. The stable genes, enriched with the translation components, are resistant to changes in the presence of abundant miRNAs. The immunity of the translation system to miRNA regulation suggests that it may be part of a global cell strategy (Lopez-Maury et al, 2008). Accordingly, there is a fundamental difference between transcription and translation processes. While transcription system can quickly respond to the needs dictated by changes in the environment (e.g., post-transcription regulation of the sensitive set), the translational machinery is stable and much less prone to variations. The immunity from miRNA regulations of the translational machinery is achieved by the relatively high expression levels (Fig 7F), but most significantly the insensitivity to most abundant miRNA in the cell.
**Alteration in COMICS simulation parameters only slightly change cells’ steady state**

We tested the reliability of the picture provided by COMICS of the miRNA-mRNA competition in living cells. This testing was done by rigorously varying the simulator’s operational parameters and checking the sensitivity and robustness of the results (Appendix Text EV1).

Our simulations operate with a total of 50k miRNAs and 25k mRNAs. These numbers are at the lower range of the amounts estimated in most living cells. We showed that changing the number of molecules, while keeping the stoichiometry only slows the dynamics, with a minimal impact on the endpoint (Appendix Fig S5).

Sampling of miRNAs and mRNA and assessing their pairing is done according to a table of probability scores that represents a rich computational-experimental body of knowledge (Agarwal et al, 2015). Once two molecules are successfully paired, our tables of probabilities get appropriately updated (Fig 3A). We show that the mRNA retention statistics at the endpoint of the simulation runs of COMICS is surprisingly robust under a wide range of parameters (Appendix Fig 4 and Fig 5). From computational considerations and based on numerous empiric observations, each pair of miRNA-target is assigned a score that reflects the calculated degree of downregulation (e.g., (Betel et al, 2010). Correlating the profile of overexpressed miRNAs from in-vitro experiments with the outcome of downregulated genes allows a refinement of the miRNA-target prediction scores (Bloom et al, 2014; Li et al, 2014b), and inferring a probabilistic measure for the effectiveness for each prediction. Along this line, TargetScan prediction tool reports on the probability of a given MBS and its combination to effectively interact with the miRNA at hand (Agarwal et al, 2015). The underlying scoring method takes into account 14 sequence features and integrates them into a probabilistic framework of repression effectiveness. We show that the values calculated in the TargetScan interaction table ((Agarwal et al, 2015)) are the most sensitive parameters for the implementation of COMICS (Appendix Fig S4). If we run the simulation after the MBS interaction values undergo a constrained randomization modification, then our self-correlations drop drastically (Fig 3E, Appendix Text EV1). In all our simulations we use the same TargetScan probabilities table. The approach offered by COMICS allows us to develop more refined models for different cell types.

**A dynamic view of miRNA regulation**

The dynamics of AGO-miRNA search complex is driven by the need to accelerate the search, while maintaining high specificity (Jo et al, 2015; Klein et al, 2017). We have implemented two formats to model an iterative cycle of the AGO-miRNA search process (Fig 3A). Under the default protocol, we consider
AGO-miRNA to be sampled according to its relative abundance, followed by a second random sampling, for any of the mRNA potential candidates (based on their probability). According to the sparse interaction affinity matrix from TargetScan (see Methods), most sampling interaction are futile, and on average only 3% of iterations end in a successful binding. This model mimics the search process of miRNAs is living cells as studied by single molecule microscopy analyses (Chandradoss et al, 2015). In order to accelerate the simulations, we have also activated an alternative implementation of COMICS, where an interacting pair of miRNA and mRNA are selected randomly at each step. The probability distribution of the selected interaction is derived from the nonzero entries of the TargetScan matrix. This implementation avoids futile interactions, thus leading to much faster computational dynamics. This computational process makes no assumptions on the mechanistic aspects of AGO-miRNA search process (Klein et al, 2017), leaving everything to the choice of the interaction probabilities matrix. The resulting retention profiles for these two modes of implementation models yield very similar views of both the cross-miRNAs stable set (Fisher exact test, p-value= 1.11e-41) and sensitive sets (Fisher exact test, p-value= 4.12e-29). The good agreement between these two probabilistic implementation schemes further supports the finding that the set of miRNA and their profile dictate the cell state (Fig. 4). The correlation of 1M runs (as in Fig 3B) and the implementation of the alternative COMICS implementation are highly correlated (Pearson correlation r= 0.91).

Another aspect of the miRNA-mRNA dynamics concerns the ceRNA paradigm (Denzler et al, 2016). Activation of ceRNA in cells is strongly dependent on an induction of specific transcripts (e.g. circular RNA, pseudogene (Thomson & Dinger, 2016)) that may alter the availability and composition of MBS (Tay et al, 2014). Accordingly, the MBS accessibility may result in redistribution of already occupied MBS by competition. In practice, the extent of de-repression due to MBS shuffling is limited (Bosson et al, 2014). A quantitative test on the direct and indirect effects of manipulating miR-122 in liver cells (Denzler et al, 2014) shows that the abundance of miRNAs and their binding sites precludes the proposed ceRNA effects under all physiological contexts (Broderick & Zamore, 2014). The situation in which miRNAs are competing by reshuffling their binding is supported in living cells only in extreme cases and under artificial extreme parameters. Based on this careful quantitation, it was proposed that low-abundance miRNAs, even with targets of highest-affinity sites are unlikely to experience any meaningful repression (Denzler et al, 2014). Indeed, the simulator explicitly guarantees that low-abundance miRNAs are sampled with low probability, and are thus unlikely to dictate the repression and the outcome of the mRNA retention level. The probabilistic nature of COMICS simulation makes it an ideal tool for testing this paradigm.
There are several improvements and extensions to COMICS that we intend to explore in the future. This includes a synergistic cooperativity of miRNA binding at different non-overlapping MBS (Balaga et al, 2012; Friedman et al, 2014), and the possible induction of preselected mRNA, for assessing the quantitative parameters associated with the ceRNA paradigm. An exhaustive application of the COMICS framework can be used for testing unresolved questions and emerging principles of miRNA regulation in vivo (Denzler et al, 2014; Hausser & Zavolan, 2014; Yuan et al, 2015b).

**Materials and Methods**

**Cell culture**

Human cell line of HeLa (cervix epithelial, # CCL-2) and HEK-293 (embryonic kidney, # CRL-1573) were purchased from the cell-line collection of ATCC. Cells were cultured at 37°C, 5% CO₂ in Dulbecco’s Modified Eagle Media (DMEM, Sigma), supplemented with 10% FBS (Life Technologies), and 1% antibiotics mixture (Sigma-Aldrich, Cat # P4333). Cells were maintained for 2 weeks and passing and splitting cells was carried out at 70-80% confluence.

**Transcription arrest and miRNA overexpression**

Overexpression of miRNAs was performed by transfected HeLa cells and HEK-293 with miRNA expression vectors that are based on the miR-Vec system, under the control of CMV promotor (Origene). Cell transfection was done using Lipofectamine 3000 (Invitrogen) as described by the manufacturer. Cells at 70% to 80% confluency were transfected with 1.5μg purified plasmid DNA containing hsa-mir-155 and hsa-mir-124a (kindly contributed by Noam Shomron, Tel Aviv University). Medium was changed 6 hrs post-transfection and fresh media was added 24 hrs post-transfection. Control empty vector expressing GFP (0.15μg) was mixed with the CMV-miR expressing vectors. Cells were monitored by fluorescent microscopy in a parallel culture at 36 and 48 hrs post transfection. The efficiency of cell transfection was >75% of the HeLa cells and ~100% of the HEK293 according to the GFP expression at 48 hours post transfection. Transcription inhibition was achieved by adding to cultured HeLa and HEK293 cells media containing Actinomycin D (ActD, 10 μg/ml in DMSO), or the appropriate control (i.e. DMSO). Cells were treated with ActD (10 μg/mL, Sigma) 24 hrs post-transfection. Cells were cultured in 6-well plates and
following treatment were lysed in 1 ml TRIzol (Invitrogen) at the indicated time points (0 hrs, 2 hrs, 8 hrs, 24 hrs).

**RNA library preparations**

At the indicated time points, HeLa cells and HEK293 cells were harvested using a cell-scaper. Purification of total RNA containing miRNA extracted from ~10^6 cells using QIAzol Lysis Reagent RNeasy plus Universal Mini Kit (QIAGEN, GmbH, Hilden, Germany). To ensure homogenization a QIAshredder (QIAGEN, GmbH, Hilden, Germany) mini-spin column has been used. To the upper aqueous phase 1.5 volumes of 100% ethanol added, and mix thoroughly. Sample has been transferred up to an RNeasy Mini spin column and centrifuge for 15s at ≥8000g at room temperature, and the mixture was processed according to the manufacturer’s standard protocol. Samples with an RNA Integrity Number (RIN) >8.5, as measured by Agilent 2100 Bioanalyzer, were considered for further analysis. mRNA libraries were generated using the Illumina Truseq RNA V2 library Seq protocols.

**miRNA library preparation**

High quality RNA was determined by Agilent 2100 bioAnalyzer. For small RNA library construction, ~1 µg of RNA was used. RNA was ethanol precipitated to enrich for small RNA. Small RNA libraries were prepared according to NEBNext Small RNA Library Prep Set for Illumina (Multiplex Compatible) Library Preparation Manual. Adaptors were then ligated to the 5’ and 3’ ends of the RNA, and cDNA was prepared from the ligated RNA and amplified to prepare the sequencing library. The amplified sequences were purified on 4% E-Gel Agarose gels (ThermoFisher # G401004), and sequences representing RNA <200 nt were extracted. Data used are derived from at least two biological duplicates. The average values of the two independent sets is reported.

**TargetScan probabilistic miRNA-mRNA pairing**

The probabilistic framework interaction table was adapted from the scores provided by TargetScan (Agarwal et al, 2015). Accordingly, high probability of successful interactions is calculated from a combination of strongly supported miRNA-mRNA pairs that comply with many features from sequence, secondary structure and evolution conservation. The complete miRNA-mRNA table include 8.22 M pairs that covers also poorly conserved interactions. We compiled the version of TargetScanHuman (Release 7.1) that reports on 19,475 genes (28,353 transcripts). We extracted the TargetScan mRNA CWCS scores (cumulative weighted context++ score), which is proxy for the predicted repression based on the different
properties of the MBS sites. The CWCW estimates the score by compiling the contribution of multiple MBS according to a miRNA family and the relative positioning at the 3’-UTR of the transcript. The predicted repression scores range from 0-1, and are identical for all representation of the relevant miRNA family members (Agarwal et al, 2015). We used a compressed version of the table that report only on pairs that are supported by conserved miRNAs with 1,183,166 pairs, covering 18953 genes and 289 miRNA families.

RNA deep sequencing analysis

RNA extracted from HeLa and HEK293 cells were taken from independent library preparations and were processed in the same sequencing slides according to standard Illumina Protocol (Trapnell et al, 2012). Next Generation sequencing was performed on small RNA (<200 nt) molecules and for mRNA by standard RNA-Seq Illumina Protocol. Each of the 48 RNA-sequencing samples covers the mRNA and miRNA sets (24 sets for the ActD treated on two cells types, at 4 times points with two sets of miRNA overexpression and a set for the control transfected by an empty vector). Each sample consisting of ~25M total reads of length 100 for each read for mRNA detection, and ~10M total reads for the miRNA detection.

The sequencing data, after removal the adaptors and filtering out low quality sequences, were aligned to mirBase (Release 21). In addition, the filtered high-quality fragments were mapped to the human transcriptome of hg19 gtf file from UCSC provided by Galaxy. Specifically, the sequenced small RNA were trimmed using Cutadapt ver. 1.13 and quality filtered using FASTX toolkit. Short reads (~30 nt long) were mapped to miRNA using mapped to miRNA genes using miRExpress 2.0 (Kim et al, 2009). Longer reads were aligned against human genome hg19 using TopHat 2.1.1, allowing 90% sequence identity and a maximum of two mismatches, with a limit of two genome alignment matches. For mRNA expression evaluation, mapped reads were submitted to Cufflinks toolkit version 2.2.1. Out of the mapped reads, only reads of length >= 17 were considered. miRNA sequences refer to mapped, high quality reads that are aligned to any of the pre-miRNA as defined by miRbase databases (ver. 21) (Kozomara & Griffiths-Jones, 2013)

Normalizations of mRNA expression and miRNA families

For analysis of all experimentally tested samples an estimation of mRNA molecules per cell was assigned to 25,000 molecules at time 0 (prior to activation of the transcription inhibition protocol). Ten of the highly expressed genes were selected from the top ranked list of mRNA. These genes were selected as being
stable throughout the 24 hrs of the ActD protocol, along all four time points. These genes were used as anchor genes. According to their quantification and the total quantity of the gene expression distribution a correction was implemented based on these anchored genes. For the rest of the analysis, reported genes are those with an overall expression which is above the threshold of 5 molecules after the quantification correction procedure (>0.02% expression). For miRNA normalization we estimated 50,000 molecules per cells and only miRNAs with more than 1 molecule after quantification were considered. The identified miRNAs were compiled to their families. Within a miRNA family, we combine the expression of miRNAs that are marked by either 5p or 3p, as well as duplicated miRNAs that are annotated according to their genomic positions. From original list of 303 miRNAs we compiled a list of 250 miRNA families. This transformation was applied to the TargetScan scoring tables and the most significant score of miRNA representative was assigned to its family.

**Probabilistic based miRNA-mRNA simulator**

The simulator input are the number of molecules for the expression profiles of miRNAs (total 50k molecules) and mRNAs (total 25k molecules) in the specific cell type, and a table of miRNA-mRNA interaction prediction extracted from TargetScan. In addition, the simulator, called COMICS (COmpetition of MiRNA Interactions in Cell Systems) supports a wide set of configurable parameters: (i) the number of total miRNA; (ii) the number of mRNA molecules in the cell; (iii) the number of iterations for completing the run; (iv) the number of iteration interval between miRNA-mRNA binding event and the mRNA removal; (v) a random removal of unbounded mRNAs according to predetermined decay rate of the mRNA as extracted and extrapolated from experimental data of mRNA half-life; (vi) addition of newly transcribed mRNAs during a configurable number of iterations interval; (vii) miRNAs or genes overexpression according to a selected multiplication factor for the degree of overexpression. (vii) incorporation of alternative miRNA-target mapping. It is also possible to activate the simulator by a set of random genes as an initial state of pre-existing iterations prior to the simulation run.

In each run, a random miRNA is chosen from the predetermined available miRNAs distribution. Next, a target is chosen randomly according to the available targets distribution. mRNA that is already bounded by other miRNA molecules can be a putative target for the chosen miRNA, if the relevant binding site is not overlapping other occupied MBS on the same molecule. Overlapping binding sites are considered for neighboring MBS that are <50 nucleotides apart. Note that MBS that physically overlap in their sequence are already removed by TargetScan with the notion that overlapping sites cannot be occupied at the same time. A binding event will occur according to the miRNA-mRNA binding probability as extracted from
TargetScan interaction table (or other prediction tables). The conversion of the interaction scores to the binding probabilities was done according to TargetScan score: \( p = 1 - 2^{-\text{score}} \). Upon a binding event, the free miRNA and mRNA distributions are updated, and the bounded mRNA molecules are marked as being occupied. An occupied molecule is removed after 1000 iterations following a successful binding event (a tunable parameter for halting an instant mRNA degradation). For mRNA to be eliminated, at least one MBS must be reported as occupied. During those iterations it is still available to bind other miRNAs in any of its non-overlapping binding sites. After mRNA removal, the bound miRNAs are released and return to the free miRNA pool and becomes eligible to engage in further binding events.

Overexpression scheme is based on multiplication of the available miRNA amount by all 7 factors (from x1 to x1000). This addition of miRNA molecules calls for calculating a new miRNA distribution while remaining with the same amount of miRNA in the cell. In case the miRNA had not detected in naïve cell, an arbitrary starting minimal amount of 0.01% (equivalent of 5 molecule/ cell).

**Statistics and Bioinformatics**

P-values were calculated using a paired and unpaired t-test, Fisher exact test, Kolmogorov Smirnov (KS) test or Chi-square tests. For testing the correspondence of two sets of different sizes, we have used the Jaccard score (J-score) that is the size of the intersection divided by the size of the union of the sample sets and it is range from 0 to 1 (no correspondence to a complete overlap, respectively). Statistical values are that are based on correlations were performed using standard Python statistical package. For annotation enrichment statistics and visualization Enrich (Kuleshov et al, 2016) was used. For testing the effect of different background gene lists for the enrichments statistics we applied DAVID (Huang et al, 2007) clustering enrichment score is based on one tail Fisher exact corrected for the number of gene ontology annotation that are used. GOrilla ranked list enrichment score that is based on Hyper Geometric statistics for any selected background (Eden et al, 2009). Enrichment was performed in view of genes that are potential candidates for our analysis and against the set of genes that express with a minimum of 0.02% of the mRNA overall expression.
Acknowledgements

The study was supported in part by Supported by ERC grant 339096 on High Dimensional Combinatorics (N.L) and ELIXIR-Excelerate grant (M.L). We thank Tsiona Eliyau for supporting the experimental data for HeLa and HEK293. We thank Noam Shomron (Tel Aviv University) for sharing the miRNA expression plasmids used in this study. We thank the members of the Linial’s lab for useful comments and discussions.

Author contributions

SMA, NL and ML contributed to the experimental design and leading the research for this study. All authors provided critical review. SMA developed the cell simulator, executed the results and conducted the data modeling. SMA wrote the computational code for the implementation of COMICS platform, SMA, NL and ML wrote the paper.

Conflict of interest

None. The authors declare that they have no conflict of interest.

Expanded View

Data availability

The mapped mRNAs and miRNAs for HeLa and HEK293 cells are listed in Datasets EV1 and EV2, respectively. The output of the % retention along 1M COMICS iterations at a 10k resolution from HeLa as input is shown in Datasets EV3. The output of the % retention along 100k COMICS iterations at a 1k resolution is shown in Datasets EV4. List of the cross miRNA stable genes for three cell types is in Datasets EV4. The lists of miRNA expression levels for 3 cells as input for COMICS are in Dataset EV5. The Matrix of the miRNA retention levels for 248 mirNa and 773 genes are in Dataset EV6. Lists of the stable genes for the three cells are in Dataset EV7. Lists of the stable genes for the three cells are in Dataset EV8. The results from enrichment tests for functional annotation for stable genes that are shared among 3 cell lines (48 genes) are in Dataset EV9. The results from enrichment tests for functional annotation for stable genes for 3 cell lines are in Dataset EV10.

Appendix Text EV1 presents the alteration of the operational parameters of COMICS.
Table EV1 summarizes the statistics of characteristics for sensitive and stable sets for 3 cell lines. Appendix Fig S1 and Fig S2 show the time point correlations of the expression of miRNAs and mRNAs from HeLa and HEK293, respectively. Appendix Fig S3 shows the retention profile for hsa-mir-124a in HeLa cells. Appendix Fig S4 shows the results of repeated runs and randomization protocol for the probabilistic miRNA-mRNA interaction table. Appendix Fig S5 shows the results of alteration of COMICS parameters. Appendix Fig S5 shows the boxplots of four quantitative characteristics of the sensitive and stable sets for HEK293 and MCF-7 cell lines.

References

Agarwal V, Bell GW, Nam JW, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. Elife 4


3.2. Sensitivity of Gene Sets to miRNA Regulation: A Cell-Based Probabilistic Approach

This chapter presents the submitted work: Sensitivity of Gene Sets to miRNA Regulation: A Cell-Based Probabilistic Approach
Sensitivity of Gene Sets to miRNA Regulation: A Cell-Based Probabilistic Approach

Shelly Mahlab-Aviv¹, Nathan Linial¹ and Michal Linial ²*

¹The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel; ²Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

SMA (Shelly.Mahlab@mail.huji.ac.il) ; NL (nati@cs.huji.ac.il), ML: (michall@cc.huji.ac.il)

Abstract

Mature microRNAs (miRNAs) are small, non-coding RNA molecules that function by base-pairing with mRNAs. In multicellular organisms, miRNAs act post-transcriptionally, leading to mRNA destabilization. The quantities of miRNAs and mRNAs, as well as their stoichiometry, govern the effectiveness of the regulation. In this study, we characterize genes by their sensitivity and robustness to miRNA regulation. We used COMICS (Competition of miRNA Interactions in Cell Systems), a stochastic computational iterative framework which simulates miRNA regulation. We monitor the cell state by quantifying the retention level of all mRNAs, at the end of the simulation run. We show that genes can be naturally classified by their mRNA decay rate following miRNA regulation. In HeLa cells, we present 5 gene classes according to their retention rates along the run. We found that the major class of genes (69%) are apparently immune to miRNA regulation. We produced thousands of cell states by overexpressing 248 miRNAs, at 7 levels relative to the basal level (x1 to x1000) by COMICS. We classified genes according to their sensitivity to changes in miRNA expression profiles, based on the end-point retention ratio for any pairs of overexpression settings (OXR, OvereXpression Ratio). For each of the 248 miRNA tested, a total of 750 genes were classified into a small number of gene sets according to their OXR. We found that while most genes are indifferent to OXR, about 30% are differently classified according to a specific miRNA that undergoes manipulation. Our results expose the impact of an overlooked quantitative view of miRNA regulation on cell states.

Keywords: miRNA-target prediction/ Stochastic model/ TargetScan/ CeRNA

Introduction

Mature microRNAs (miRNAs) are non-coding RNA molecules (~22 nucleotides) that regulate genes through base complementarity with their cognate mRNAs, at the 3’ untranslated region (3’-UTR) (Moore et al., 2015). In multicellular organisms, miRNAs act by destabilization of mRNAs and interfering with the translation machinery (Chekulaeva and Filipowicz, 2009; Eichhorn et al., 2014). Alteration in the relative abundance of miRNAs may lead to transition between states, and the establishment of cell identity (Pelaez and Carthew, 2012). Changes in miRNA cell composition is associated with viral infection (Zhang et al., 2012), differentiation, and cancer transformation (Bertoli et al., 2015; Lu et al., 2005).

The current catalogue of miRNAs (Kozomara and Griffiths-Jones, 2013) contains ~2500 mature miRNAs from humans. However, the number of miRNA types that are expressed at
a substantial amount in most cells is much limited. The miRNA regulatory network can be formulated as a bipartite graph with nodes at one side correspond to coding gene transcripts, and on the other side to mature miRNAs. Many miRNAs carry the potential for targeting hundreds of transcripts (Balaga et al., 2012; Rajewsky, 2006), and tens of miRNA binding sites (MBS) are predicted at the 3’UTR of mRNAs (Landgraf et al., 2007). Most of our current knowledge on the specificity of miRNA-mRNA regulatory network is based on computational prediction tools (Peterson et al., 2014). Most such tools suffer from a large number of false positives (Pinzon et al., 2017). Recent results from CLIP-based protocols provide rich, unpaired collection of miRNAs and mRNAs from living cells (Li et al., 2014). Unfortunately, many of the deep sequencing-based protocols suffer from low coverage and poor consistency (discussed in (Lu and Leslie, 2016)). It is expected that poorly expressed miRNAs do not contribute to downregulation of gene expression (Hausser and Zavolan, 2014). In reality, only ~60% of the human coding genes are regulated by miRNAs in a cellular context (Ha and Kim, 2014; Jonas and Izaurralde, 2015).

The properties of the miRNA-mRNA regulation in a specific cell type depend on the amounts and concentration of miRNAs and their exact stoichiometry. A key player in the regulation is the availability of AGO protein, the catalytic component of the RNA silencing complex (RISC) (Jandas et al., 2012; Wen et al., 2011). In view of mRNAs, the abundance and positions of MBS along the relevant transcript (Jens and Rajewsky, 2015) dictate the potential of miRNA interaction, but not necessarily the potency for gene expression attenuation (Agarwal et al., 2015). The properties of the miRNA-mRNA complex network call for an unbiased probabilistic model for defining the principle design of miRNA regulation.

In this paper, we develop a quantitative view on miRNA regulation. We provide a stochastic -probabilistic model that operates at the cellular level. We configure an iterative simulator on HeLa cell-line by an exhaustive set of miRNA overexpression manipulations. In the end of the simulation run, the mRNA retention levels are monitored and a new steady state is defined. We created two unbiased gene sets - the first one concerns the dynamic behavior along the simulation run (dynamic-classes). The second classification concerns gene class by their sensitivity to changes in miRNA quantities, based on the end-point retention ratio for any pairs of overexpression settings, these are called OXR-classes (OverExpression Ratio-classes). The rich representations by the dynamics- and OXR-classes expose overlooked properties of miRNA regulation.

Methods

Probabilistic map for miRNA-mRNA pairing

The probabilistic framework interaction table was adapted from the scores provided by TargetScan (Agarwal et al., 2015). Accordingly, high probability of successful interactions is calculated from a combination of strongly supported miRNA-mRNA pairs that comply with many features from sequence, secondary structure and evolution conservation. The unfiltered complete miRNA-mRNA interaction table includes 8.22 M pairs covering the highly and poorly conserved interactions. The refined version used from TargetScanHuman (Release 7.1) includes 19,475 genes. We used a compressed version of the table that reports only on pairs that are supported by conserved miRNAs with total of 1,183,166 pairs, covering 18,953 genes and 289 miRNA families. The predicted repression scores range from 0-1, and are identical for all representation of the relevant miRNA family members (Agarwal et al., 2015). The
conversion of the interaction scores to the binding probabilities was done according to TargetScan score: $p = 1 - 2^{-\text{score}}$.

**Normalizations of mRNA expression and miRNA families**

For mRNA expression profile, we used experimental data that we have produced (HeLa cell, expression by mapped reads (Mahlab-Aviv, 2018)). We reliably mapped 16,355 mRNAs and 539 miRNAs that were expressed at a minimal level of $\geq 1$ reads (from total mapped reads $\sim 3$ M). For miRNA normalization we estimated 50,000 molecules per cell and miRNAs with more than 1 molecule after normalization was considered (total 110 miRNAs). Similarly, we estimated 25,000 mRNA molecules and with that threshold (total 3666 mRNAs). For the generic miRNA list, miRNAs are compiled to match the definitions provided by the TargetScan interaction matrix, a total of 248 miRNAs according to TargetScan table. The analysis that are based on % retention is bounded by genes having a minimal number of molecules (i.e. 0.02% expression, total 753 genes).

**Probabilistic based miRNA-mRNA simulator**

The simulator inputs include the number of molecules from the expression profiles of miRNAs (50k molecules) and mRNAs (25k molecules), and a table of miRNA-mRNA interaction prediction extracted from TargetScan. In each run, a random miRNA is chosen from the predetermined available miRNAs distribution. Next, a target is chosen randomly according to the available targets distribution. mRNA that is already bounded by other miRNA molecules can still be a putative target for the chosen miRNA, if the relevant binding site is not overlapping an occupied MBS on the same molecule. Overlapping binding sites are considered for neighboring MBS which are $<50$ nucleotides apart. A binding event will occur according to the miRNA-mRNA binding probability as extracted from TargetScan interaction table and converted interaction scores to binding probabilities. Upon a binding event, the free miRNA and mRNA distributions are updated, and the bounded mRNA molecules are marked as being occupied. An occupied molecule is removed after 1k iterations following a successful binding event. For mRNA to be eliminated, at least one MBS must be reported as occupied. After mRNA removal, the bound miRNAs are released and return to the free miRNA pool, and therefore become eligible to engage in further binding events.

COMICS supports a wide set of configurable parameters: (i) the number of total miRNA; (ii) the number of mRNA molecules in the cell; (iii) the number of iterations for completing the run; (iv) the number of iteration interval between miRNA-mRNA binding event and the mRNA removal; (v) a random removal of unbounded mRNAs according to predetermined decay rate of the mRNA as extrapolated from experimental data of mRNA half-life; (vi) addition of newly transcribed mRNAs during a configurable number of iterations interval; (vii) miRNAs or genes overexpression according to a selected multiplication factor for the degree of overexpression. (vii) incorporation of alternative miRNA-target mapping. It is also possible to activate the simulator by a set of random genes as an initial state of pre-existing iterations prior to the simulation run.

Overexpression scheme is based on multiplication of the available miRNA amount by all 7 factors (from x1 to x1000). The addition of miRNA molecules calls for calculating a new miRNA distribution. In case the miRNA had not been detected in naive cell, an arbitrary starting minimal amount of 0.02% (equivalent of 10 molecule/ cells) are added to the naive cell (x1).

**Analytical methods**

Statistical values are that are based on correlations were performed using standard Phyton statistical package. For annotation enrichment statistics and visualization Enrich (Kuleshov et al., 2016) was used. Clustering was performed by the k-mean classification. We used the unsupervised Elbow method to test the consistency within clusters by the percentage of variance
explained. (i.e. the ratio of the between-group variance to the total variance). A change in the
slop is indicative to the optimal number of clusters in that dataset.

**Results**

**Assessing the probabilistic approach for miRNA - mRNA interactions**

The goal of this study is to model the outcome of the miRNA-mRNA network under simplified
conditions of translational arrest, mimicking the stochastic nature of miRNA regulation in
living cells. Classifying genes into different sets reduces the complexity of the analyzed system.
Moreover, it provides new insights on shared sensitivity to miRNA regulation.

The nature and extent of miRNA regulation in living cells is depicted by the absolute quantities,
composition and stoichiometry of the main players of the network, i.e., the miRNAs and
mRNAs (Arvey et al., 2010). Evidently, the molecular interactions of miRNA and mRNA
within a cell is a stochastic process. The specific composition in cells, and binding probability
dictate the effectiveness of attenuation of gene expression. Systematic analysis of the miRNA-
mRNA interaction network shows that the miRNA regulation operates under tight
stoichiometric constrains in living cells. The experimental data from HeLa for miRNAs and
mRNAs are extracted from repeated NGS experiments. Total of 539 were mapped and 16,236
expressed mRNAs (not including miRNAs). We developed an iterative simulator called
COMICS (Competition of miRNAs Interactions in Cell Systems) which was designed to
capture the properties and quantitative consideration of miRNA-mRNA interaction in living
cells. Fig. 1 illustrates the scheme from a cellular perspective while focusing on the probabilistic
framework.

![Fig. 1 Schematic procedure of the probabilistic nature of COMICS.](image)

COMICS iterations capture the stochastic process that occurs in cells according to the quantities
and the ratio of miRNA to mRNAs. The sampling process (Fig. 1, [1]) driven by the distribution
miRNAs and mRNAs according to experimental measurements (Fig 1, pink frames). Each
mRNA is characterized by the types and positioning of its miRNA binding sites (MBS) at the
3’UTS of the transcript. The interaction prediction table is associated with a probability-based
scores for any specific pairs of miRNA and MBS in the context of a specific mRNA. Recall
that the expression profiles of miRNAs and mRNAs are cell-type specific.

In each iteration, a miRNA is sampled randomly, according to the cell’s miRNA abundance
and composition. Next, one of its target gene is chosen randomly according to the measured
expressed mRNAs distribution. In the following stochastic step, the randomly chosen miRNA
and its target are expected to interact according to the probability of such pairing to be
successful. The binding probabilities are based on the miRNA-target prediction of TargetScan
(Fig. 1 [2]). It provides a sparse table of miRNA-MBS interactions and reports on 1.2M pairs

---

4
(see Methods). Each miRNA-MBS interaction is associated with a probabilistic score that is a proxy for the level of confidence for that interaction, and can be considered the probability of effective binding for any specific pair. Following a successful binding event, the distribution of the miRNAs and the mRNAs is updated accordingly (Fig. 1 [3]).

The status of the mRNA following a successful pairing is changed (i.e. ready to degradation), and it is marked as ‘occupied’. Upon binding, it may be engaged in additional binding of miRNAs. For MBS that are in a physical proximity to each other, the overlapped interaction is being excluded. The occupied mRNA is marked for degradation with some delay that mimics the likely instance of a cooperative binding on a target by multiple miRNAs prior to its degradation. Once the occupied mRNA is removed from the system, all miRNAs that were bound to it are relieved to the free miRNA pool. As a result, the stoichiometry of miRNA to mRNA is gradually changing with an increase in the apparent ratio of miRNAs to free mRNAs in the cell.

Results of such simulation is illustrated in Fig. 1[4]. In this case, hsa-mir-155 was overexpressed. The decay rate for all genes is shown. Target genes of hsa-mir-155 are extracted from the TargetScan table. The target and non-target genes are differently colored (pink for targets, blue for non-targets). Cell state is defined as the retention levels (%) of the unbounded genes at the end of the simulation run. COMICS routinely runs for 100,000 iterations.

COMICS supports a wide set of configurable parameters (see Methods), and was tested against the paired interactions as presented by CLASH (Helwak et al., 2013). CLASH data reports on HEK293 cells mRNA. By including the input from HEK293 we validated that the overlap with CLASH results is statistically significant (10k to 100k iterations, stable p-value <0.002) (Mahlab-Aviv, 2018). These tests confirm the relevance of the simulator protocol to successfully mimic the experimental results. In this study, we operating the system for a status of ‘transcriptional arrest’ with no production of additional mRNA, and ignoring other sources for mRNA degradation beside the regulation by miRNAs.

**Gene classes according to mRNA retention: The dynamic-classes**

![Fig. 2. Dynamic classification and analysis. (A) Dynamic classes of Hela cells along 100k iterations of a single simulation run. (B) The distribution of number of MBS per gene in the 5 dynamic classes.](image)

K-means classification was performed on all genes according to their retention profiles throughout one simulation run (using the Elbow method for 100k COMICS iterations). Fig. 2A presents the clustering result for k=5 (ensuring a minimal class size). The average behavior of the dynamic classes of all genes is shown (total 750 genes). The number of genes associated with the different dynamic class are very skewed, with cluster #1 occupies 69% of the genes and <4% are associated with the fast decay cluster (cluster #5). Inspecting the decay rate of the
low retention clusters (cluster #4 and #5, 7% of total genes) shows that while the rates are very different the endpoint converged to a similar low retention levels.

Cluster #5, the lowest retention rate, contains genes that are enriched in annotation of transcription regulators and splicing factors. On the other hand, cluster #1, the most stable one, contains genes of the translation machinery, ribosomal subunits, chaperones, and cytoskeletal components. Genes that are associate with specific functional groups provide additional support for the benefit of classifying genes for reducing the system’s complexity.

Fig. 2B shows a wide distribution of the number of MBS per gene for clusters #2-#5 (irrespectively to their size), with the exception of cluster #1. The number of MBS in the mRNAs in cluster #1 is skewed towards low number of MBS, with an average of 29.7 MBS per gene, while for the rest of the clusters the average is 54.1 MBS per gene. The statistical analysis on average amount of molecules in the cells, length of 3'-UTR are insignificantly different among all dynamic classes. We conclude that class #1 is characterized by relatively low potential to be engaged in successful interactions due to the low binding potential of its genes.

**Exhaustive perturbations of miRNAs expose an underlying gene specific classes**

To determine the underlying partition between properties that can be attribute to the specific cell type (i.e. profile of its miRNA and mRNAs) and features that are evolutionary driven (i.e. number of MBS in the 3'-UTR, length of 3'-UTR), we activated a systematic analysis on hundreds of cell states derived from a broad range of manipulations of all expressed miRNAs. We applied COMICS simulations using overexpression paradigm in HeLa cells for 248 miRNAs that were compiled according to their representation in the miRNA-MBS TargetScan prediction table (see Methods). We multiply the basal abundance (x1) of each of these miRNA families by several factors: x3, x9, x18, x90, x300 and x1000. For each such factor (f), a final retention table was computed and the cell states in the end of the COMICS run were monitored.

![Fig. 3](image_url)

**Fig. 3.** Retention values following several factors of overexpression from HeLa. Rainbow color code: red=0 to purple=100% retention level at the end of the simulation run ). Each Table consists of 248 miRNAs (columns) and 750 genes (rows). The overexpression factors are indicated. The matrices are clustered by genes.

Fig. 3 shows the pattern of the retention (%) for a matrix of miRNAs (columns) and genes (same gene list, Fig. 2). The 4 panels are associated with Mfij at factors x3, x9, x90 and x1000. A cell
state results are presented by a color-coded table with 750 genes (rows) whose initial expression level exceeds a pre-determined threshold, and 248 miRNAs (columns, organized alphabetically). Each of the listed miRNA was overexpressed by the marked factor. Therefore, each cell in the matrix Mfij is the final retention of gene i after 100k iterations of COMICS for the overexpressed experiment of miRNA j (Fig. 3).

Inspecting the Mfij (4 factors of overexpression conditions) reveals that genes are naturally clustered by their final retention level. For example, the top ‘red’ patch represents a set of genes that were very sensitive (red=0% retention; purple=100% retention) to almost all miRNA manipulations. Clear observations can be drawn from inspecting these matrices (Fig. 3): (i) Large number of coordinated behavior is evident for each set of genes. It is reflected by observing a similar color in the entire row across most miRNA columns. A large number of rows display this distinctive property. (ii) As the overexpression factor increases (i.e. x1000) the pattern of the columns (i.e. specific miRNAs) becomes evident. It is reflected by observing a similar color across the entire columns across most rows.

**Perturbation pair ratio classes expose varying sensitivity to miRNA quantitative changes**

In view of the observation that many genes behave similarly with respect to their retention levels at a wide range of Mfij, we tested whether genes can be characterizing by their degree of sensitivity to an alteration in abundance of miRNAs (as studied by COMICS exhaustive overexpression tests).

To follow the relative changes of each gene retention in each miRNA overexpression, and in each overexpression factor, we computed the retention ratio between any tested overexpression factors. Formally, we computed the value of Mfij/ Mkij, which is the ratio of the retention of a specific gene (simulation at 100k iterations) in a specific miRNA overexpression of factor f, and its retention in the same miRNA overexpression factor k (Fig. 4A). For visualization purpose, a discretization was applied for which ratio that is >2 folds implies that the retention of genes i in the overexpression of miRNA j by factor f is higher than its retention where miRNA j was overexpressed by factor k (Fig 4B, blue cells). However, ratio that is <0.5 implies that in factor f the gene is more prone to degradation with respect to factor k (Fig 4B, red cells).

![Fig. 4. OXR classes.](image-url)
Fig 4C illustrates the retention ratios of selected three genes (for illustrative purposes). It is shown that following overexpression of hsa-mir-155 the gene TPI1 remains stable throughout any of the tested retention ratios. As expected, TPI1 belongs to the dynamic class of genes that are extremely stable in the system. A different behavior is observed for ITGB1 whose expression is very unstable and sensitive to a minor change in overexpression factors (e.g. x18/x9). The non-monotonic behavior of ITGB1 and DSTN are evident.

Fig 4D illustrates the gene sensitivity as measured by the retention rate of different genes in the case of hsa-mir-155 for 6 different pairs of factors: (x1, x3), (x3, x9), (x9, x18), (x18, x90), (x90, x300), and (x300, x1000). Same analysis (as in Fig. 4C) was performed for all expressed genes. The results for all genes for a specific miRNA overexpression were clustered by K-means clustering algorithm (a cluster must contain >5 genes; OXR-classes). The analysis reveals that OXR-classes display different sensitivity pattern with respect to the pair-overexpression retention ratio. Fig. 4D shows the partition of all genes to 5 clusters (marked #1 to #5). The majority of the genes (~71%, cluster #1, colored grey) are indifferent to the levels of overexpression factors. However, the rest (~29%) of the genes are sensitive to some extend for the overexpression factors that was used. For example, cluster #3 (Fig 4D, yellow) contains genes that their retention rate is drastically decreased by OXR, as the overexpression factor increases (f). It is satisfying to note that most hsa-mir-155 expressed target genes (52%) belong to cluster #3. However, other target genes are associated with additional clusters with most of them belong to cluster #1.

The behavior of 4 selected miRNAs according to their OXR-classes, for 6 matrix ratio combinations is shown (Fig. 4D and Fig. 5). The represented miRNAs are expressed at different order of magnitudes. miRNAs that are highly expressed (e.g., hsa-mir-7, 4.2%) were analyzed as well as miRNAs that are expressed at a low level (hsa-mir-320, 0.15%; hsa-mir-155, 0.02%). The expression levels of 110 miRNAs that are express in HeLa cells are listed in Supplemental Table S1.

Fig. 5 shows several behaviors associated with the OXR-classes: (i) In almost all instances, OXR-class that includes most targets of the subjected miRNA, decreases monotonically with a maximal effect seen for ratio of the the highest overexpression pair (i.e. x1000/x300). This observation is generalized, irrespectively to the initial expression level of the miRNAs. This strongest effect is associated with OXR-classes that include only targets (e.g., cluster #2 in hsa-mir-7). (ii) Some OXR-classes show a characteristic behavior that cannot be trivially anticipated. For example, a switch in the trend of cluster #4 of hsa-mir-99 shows that for the ratio of x18 and x90 the effect of retention does not follow the behavior at any of the other expression ratio pairs. (iii) For some miRNAs there are no target detected in the list of analyzed genes (e.g., hsa-mir-92). In such instance, we do not detect a ‘target’ OXR-class and the trend of the OXR-classes are more indicative for a gene stabilized pattern (e.g., hsa-mir-99 and hsa-mir-92). (iv) For all miRNAs, the largest OXR-classes includes 82% to 91% of the analyzed genes. This general observation implies that most genes are insensitive to the perturbation according to the pair-ratio. (v) Some of the clusters show extreme increase of decrease in the retention rates (e.g., hsa-mir-320 for cluster #3, at x90/x18). This implies the high sensitivity of these gene sets to a specific miRNA abundance.

We illustrated OXR-classes for 6 ratio-matrices. The comparison and analyses are valid for each miRNA (total 248), and for 21 pairs the the 7 overexpress factors tested.

Discussion
Cells’ behavior cannot be extracted from direct measurement of the composition of miRNAs and mRNAs (Arvey et al., 2010; Landgraf et al., 2007). Most insights on regulation of gene expression by miRNAs in the complexity of the cells are based on CLIP-Seq and CLASH methodologies (Li et al., 2014). Based on many such studies, it was concluded that detailed quantitative considerations of miRNA and mRNA govern the dynamics and the steady state of gene expressed in cells (Bosson et al., 2014; Hausser and Zavolan, 2014). Nevertheless, the underlying rules for post-transcriptional regulation by miRNAs are still fragmented (Erhard et al., 2014).

![Fig. 5. Example of OXR classes for different miRNAs, following manipulations of hsa-mir-7, hsa-mir-99, has-mir-24, and hsa-mir-320. The numbers of the targets of each tested miRNA are indicated (see color legend). miRNA expression (%) is marked. The values at y-axis are expressed by log2.](image)

In all the presented COMICS results we consider miRNA pool to be constrained by the amount of AGOs in the cells. Several studies estimate in each cell ~50k AGO molecules (but huge variation is reported). COSMIC is insensitive to such quantitation debates as the sampling protocol (Fig 1) relies on probabilistic formulation. Under varying levels of miRNA overexpression, the loading of miRNA molecule on AGO is driven by the actual distribution of that miRNAs.

In this study we consider two sets of gene classes: dynamic-class (Fig. 2) and OXR-classes (Figs 4-5). These two complementary types of classes capture different aspects of miRNA regulation dynamics. Results from the dynamic-class show than genes that are likely to be successfully targeted are those having large number of MBS at the 3’-UTR (Fig. 2). However, the dynamic-classes #2 to #5 are not distinguishable by such feature. Specifically, cluster #2, #3, #4, #5 are associated with 50.3, 71.6, 63.3 and 47.6 average MBS per gene. In the future, and using analysis from different additional cell lines, we will extend the analysis to include evolutionary based features such as the distribution of MBS along the 3’-UTR, the number of alternative spliced variants that affect miRNA regulation, the presence of MBS that binds any of the most abundant miRNAs in the cell. Addition of cell specific features to the clustering protocol (i.e. cell specific genes variants) will benefit the refinement of the classification protocol.
The OXR-classes aim to capture the system dynamics rather than the dynamics of the gene expression downregulation by miRNAs. We were able to cluster genes to their OXR-classes by performing hundreds of simulations allowing a robust assessment of cell states. For most instances, under all conditions, the majority or the expressed genes are not sensitive to the matrix-ratio measures. Namely, the final retention that is achieved in all conditions of overexpressed miRNAs is unchanged (Fig. 5, y-axis = 0). In a smaller set of genes, miRNAs strongly regulate their targets when a switch in the abundance occurs from one level to another.

It was shown that miRNA profile is carefully regulated to promote and stabilize cell fate choices (Shenoy and Blelloch, 2014). In accord with this notion, we show a general trend for the highly expressed miRNAs. Actually, many large-scale miRNA overexpression experiments (e.g., (Lim et al., 2005; Wang and Wang, 2006)) overlooked and ignored the gene sensitivity to the actual amounts of miRNAs. For the majority of studies, the degree of overexpression of miRNA is not even reported. We suggest that it may contribute to the inconsistency that is often observed among miRNA-target experimental results. Using COMICS dynamic comparison benefits an accurate prediction considering the exact levels of miRNA overexpression.

Despite two decades of research in the miRNA field, some basic principles remained to be discovered. Current miRNA-mRNA prediction tools suffer from a large number of false positive. The experimental methodologies (e.g. CLASH and CLIP-Seq) that are based on capturing the interactions followed by sequencing are not always consistent, show poorly reproducibility (Lu and Leslie, 2016). We suggest that a representation of genes by their dynamic and OXR-classes can yield an accurate yet simplistic model for miRNA regulation. This study illustrates the use of COMICS results for gene classification. It further stresses the importance of quantitative view for miRNA regulation modeling.

In pathological cells, such as in cancer, a quantitative change in the amounts of miRNAs is often the most significant molecular change observed in early phase of cancer development. Assessing changes in behavior of representative genes from OXR-classes could benefit cancer diagnosis. Some OXR-classes may serve as indicator for a shift in cell states and identity. The available of a very small number of coherent gene classes, argues that cells display an unexpected robustness with respect to the miRNA regulation. The ability to classify genes according to dynamic overlooked features carries its potential to improve cell modeling, and improving the understanding of cellular miRNA regulation.

**Funding**

The study was supported in part by ERC grant 339096 on High Dimensional Combinatorics and ELIXIR-Excelerate, Horizon 2020, grant 676559.

**References**


3.3. Appendix

Supplemental file S1

**The Translation Machinery Is Immune to miRNA Perturbations: A Cell-Based Probabilistic Approach**

Mahlab-Aviv Shelly¹, Nathan Linial¹ and Linial Michal²*

¹The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel; ²Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

(include legends of Supplemental Figures S1-S6)

**COMICS simulator sensitivity and robustness with respect to parameters’ change**

We further tested the results of COMICS for reliably mimicking the competition of miRNAs in living cells, and the outcome on attenuation in mRNAs expression. This was performed by rigorous testing the sensitivity and robustness of the of results from the COMICS simulator following various operational parameters.

Supplemental file S1 provides a set of tests on the key parameters used. As the protocol we apply is probabilistic with a stochastic component for each iteration, we tested the variation of the simulator for independent runs. We compared the simulator output of three different runs using
the same parameter set. We show a very high correlation (Pearson correlations are between 0.86 to 0.93) (Supplemental Figure S4).

Next, we tested the effect of the binding probability table. TargetScan score table (Agarwal et al, 2015) was shuffled using two randomization methods: (i) Constraint randomizing - the total score for each gene and for each miRNA was randomized while fixing the total probability values for each gene and each miRNA as presented in the original scoring table; (ii) Naïve randomization - In this case, we kept the interaction scores as provided by TargetScan matrix but randomized the MBS-miRNA pairs (see Materials and Methods). The correlation of the original TargetScan table and the randomized versions drops significantly. The correlations measured after 100k iteration steps is 0.48-0.51 and 0.06-0.01 for the constraint and naïve randomization schemes, respectively (Supplemental Figure S4). This comparison shows the critical information that is provided by the miRNA-mRNA interaction prediction table for the output of the simulator.

We tested the effect of varying the initial quantities of miRNAs and mRNAs. Evidence of experimental data of total mRNA molecules from single cell estimates the number of transcripts to vary from 25k to 1M (Marinov et al, 2014; Ramskold et al, 2012). The number of miRNAs was estimated from direct measurements in various cells to range between 120k to 200k molecules (Denzler et al, 2014; Janas et al, 2012; Mukherji et al, 2011). However, the number of AGO proteins remained limiting (~20k-100k molecules). Those estimates greatly vary among cell types and developmental stages. The sensitivity of these absolute amounts to the expected effectiveness of the regulation was assessed experimentally and via a mathematical model (Bosson et al, 2014).

The effect of the varying the stoichiometry of miRNAs relative to mRNAs was tested by changing the absolute and relative ratio of the two types of molecules. The final analyses are based on testing
the retention of 754, 755 and 694 expressed genes in HeLa, HEK-293, and MCF-7, respectively.

We found very high and significant Pearson correlations between different sets of miRNA and mRNA quantities (Supplemental Figure S5). In all tests, the number of molecules in the cells were bounded by a minimum of 25k molecules for miRNA or mRNAs. The default setting was 50k and 25k molecules for miRNA and mRNA, respectively. The ratio between miRNAs and mRNAs was tested from 1:1 to 8:1 ratio. Similarly, the ratio of mRNA to miRNA was elevated from the default value of 0.5:1 to 2:1. In all instances, the effect on the endpoint was minimal and a change in the ratio up to 8 folds had an effect on the kinetics but not on the endpoint of the simulation. The notion of miRNA regulating the gene expression by thresholds rather than by simple buffering was proposed (Mukherji et al, 2011). Changing the cell miRNA concentration without affecting the stoichiometry of 2:1 resulted in a minor change in the decay rate. Changes in miRNA quantity result in no significant changes in the final retention distribution (KS test p-values range between 0.80 to 1.0, Pearson correlation between 0.68 to 0.89, Supplemental Figure S5A).

The robustness of COMICS to the sampling procedure was tested. We have implemented two formats to model an iterative cycle of the AGO-miRNA sampling process (Figure 3A). In order to accelerate the simulations, we have also activated an alternative implementation of COMICS, where an interacting pair of miRNA and mRNA are selected randomly at each step. The probability distribution of the selected interaction is derived from the nonzero entries of the TargetScan matrix. This implementation avoids futile interactions, thus leading to much faster computational dynamics. This computational process makes no assumptions on the mechanistic aspects of AGO-miRNA search process (Klein et al, 2017), leaving everything to the choice of the interaction probabilities matrix. The resulting retention profiles for these two modes of implementation models yield very similar views of both the cross-miRNAs stable set (Fisher exact test, p-value=
1.11e-41) and sensitive sets (Fisher exact test, p-value= 4.12e-29). The good agreement between these two probabilistic implementation schemes further supports the finding that the set of miRNA and their profile dictate the cell state (Figure 4). The correlation of 1M runs (as in Figure 3B) and the implementation of the alternative COMICS implementation are highly correlated (Pearson correlation r= 0.91).

An additional parameter that was tested concerns the interval of iteration between the pairing of miRNA-mRNA and the removal of the mRNA (mimicking gene expression attenuation by AGO-2). We change the setting of the mRNA elimination interval by changing the parameter from 1000, 5000 and 10,000 iterations (i.e. each 1%, 5% and 10% of the simulation run). We found no significant changes in the simulation results using any of those values (Supplemental Figure S5C).

References


### Table S1: Statistical analysis for sensitive and stable gene sets for quantitative features in HEK293, HeLa and MCF-7 cells

### Supplemental Figure Legends

#### Supplemental Figure S1. The expression profile of miRNA and mRNA in HeLa cells under transcription arrest by ActD. (A) Expression of miRNAs in pairs of 4 different time points. RNA samples were collected at 0 hr, 2 hrs, 8 hrs and 24 hrs following transcription inhibition by ActD. The scale for the expression levels is in log scale. Spearman correlation (R) is listed along the p-value of the significance. (B) Expression of mRNAs in pairs of 4 different time points. RNA samples were collected at 0 hr, 2 hrs, 8 hrs and 24 hrs following transcription inhibition by ActD. The scale for the expression levels is in log scale. Spearman correlation (R) is listed along the p-value of the significance. Source data is compiled in Dataset EV1.

#### Supplemental Figure S2. The expression profile of miRNA and mRNA in HEK293 cells under transcription arrest by ActD. (A) Expression of miRNAs in pairs of 4 different time points. RNA
samples were collected at 0 hr, 2 hrs, 8 hrs and 24 hrs following transcription inhibition by ActD. The scale for the expression levels is in log scale. Spearman correlation (R) is listed along the p-value of the significance. (B) Expression of mRNAs in pairs of 4 different time points. RNA samples were collected at 0 hr, 2 hrs, 8 hrs and 24 hrs following transcription inhibition by ActD. The scale for the expression levels is in log scale. Spearman correlation (R) is listed along the p-value of the significance. Source data is compiled in Dataset EV2.

Supplemental Figure S3. Retention profile of mRNAs following miRNA overexpressing in HeLa cells (A) Relative retention of (i) hsa-mir-124 targets genes, (ii) hsa-mir-124 non-target genes and (iii) the average of targets and non-target genes (pink and blue lines respectively). (B) Retention distribution of the genes according to their labels as targets (upper panel, pink) and non-targets (lower panel, blue) as was measured at 24 hrs. The plots compare the partition of genes from the control (smooth line), and from hsa-mir-124 overexpressed condition (dashed line). The number of genes that are included in the analyses are marked in parentheses. Target genes are shown in pink lines (top, t-test p-value 0.53) and the non-target genes are shown in blue lines (bottom, t-test p-value 0.004). Note the shift in the distribution in the non-target genes towards the genes with higher retention level. All genes with a retention level ≥100 are shown as 100% retention.
Supplemental Figure S4. Comparison of different simulator runs using different scoring tables. Pearson correlation of different simulation runs using three miRNA-mRNA interaction scoring tables: (i) TargetScan (marked as TS), (ii) randomized table of TargetScan. The randomization was done by fixing the total scores of each gene (row) and each miRNA (column) in the original TargetScan table. (iii) Naïve random table. In this case the original scores of TargetScan table were assigned to random pairs of miRNA–mRNA. Three independent simulation repetitions were performed using each of the above tables.
Appendix Figure S5. Comparison of different parameter settings of COMICS simulator. (A) Pearson correlation coefficients of the final retention after different simulation runs. Each run was done using different set of parameters: different quantification and stoichiometry of miRNA mRNA ratio; different iteration interval between mRNA binding events and the removal of the mRNA from the system. (B) The retention distribution at the end of different runs. The quantity of total mRNAs was fixed to 25,000 molecules, while its ratio with miRNA total quantity was set to 1:1, 1:2, 1:4 and 1:8. (C) The retention distribution at the end of different runs. The quantity of total miRNAs was fixed to 50,000 molecules, while its ratio with mRNA total quantity was set to 2:1, 1:1, 1:2 and 1:4. (D) The retention distribution at the end of different runs. The removal interval was set to 1000, 5000 and 10,000 iterations.
Appendix Figure S6. Boxplot statistics for sensitive and stable gene sets quantitative features. (A) Quantitative analysis of four features associated with group of genes in HEK293 cells. The statistics is applied for 4 different group of genes: Cross miRNA sensitive genes (sensitive); Cross miRNA stable genes (stable); Others, all gene excluding the sensitive and stable gene sets, and all analyzed genes (753 genes). (B) Quantitative analysis of four features associated with group of genes in MCF-7 cells. The statistics is applied for 4 different group of genes: Cross miRNA sensitive genes (sensitive); Cross miRNA stable genes (stable); Others, all gene excluding the sensitive and stable gene sets, and all analyzed genes (753 genes).

Datasets:

Dataset S1: HeLa miRNA and mRNA expression profiles in four different time points after transcription inhibition, in control and hsa-mir-155 over expression experiments

Dataset S2: HEK293 miRNA and mRNA expression profiles in four different time points after transcription inhibition

Dataset S3: Genes retention (in %) during a simulation run of 1,000,000 iterations using HeLa gene and miRNA expression profile as input. Reporting each 10k iterations.

Dataset S4: Genes retention (in %) during one simulation run of 100,000 iterations using HeLa and HEK293 gene and miRNA expression profile as input. Reporting each 1k iterations.
**Dataset S5:** miRNA expression profiles after quantification and removal of miRNAs that are below a minimal expression threshold in HeLa, HEK293 and MCF7 cells.

**Dataset S6:** Retention map of 248 miRNAs and 773 genes from HeLa. Each cell shows the results in the end of the simulation (100k iterations, factor X300).

**Dataset S7:** List of gene names and symbols of shared cross miRNA-stable genes in 3 cell types

**Dataset S8:** List of gene names and symbols of shared cross miRNA-sensitive genes in 3 cell types

**Dataset S9:** A list of 48 stable genes that are shared by Hela, KEK293 and MCF-7. Results from the annotation enrichment test (along with the clustering of annotations and statistics).

**Dataset S10:** Results of the annotation enrichment test and statistics from GO annotation enrichment tools (Enrichr and DAVID, see Materials and Methods). Annotations for sensitive (HeLa) and stable gene set from Hela, KEK293 and MCF-7.
Chapter 4

Computational Analysis of Translation Elongation Rate

4.1. Conservation of the relative tRNA composition in healthy and cancerous tissues

This chapter includes the following publication:

Conservation of the relative tRNA composition in healthy and cancerous tissues

SHELLY MAHLAB,1,5 TAMIR TULLER,2,4,5 and MICHAL LINIAL3,4,5

1School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 91904, Israel
2Iby and Aladar Fleischman Faculty of Engineering, Department of Biomedical Engineering, Tel Aviv University, Tel Aviv 69978, Israel
3Department of Biological Chemistry, Institute of Life Sciences, Sudarsky Center for Computational Biology, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

ABSTRACT

Elongation in protein translation is strongly dependent on the availability of mature transfer RNAs (tRNAs). The relative concentrations of the tRNA isoacceptors determine the translation efficiency in unicellular organisms. However, the degree of correspondence of codons and the relevant tRNA isoacceptors serves as an estimator for translation efficiency in all organisms. In this study, we focus on the translational capacity of the human proteome. We show that the correspondence between the codon usage and tRNAs can be improved by combining experimental measurements with the genomic copy number of isoacceptor groups. We show that there are technologies of tRNA measurements that are useful for our analysis. However, fragments of tRNAs do not agree with translational capacity. It was shown that there is a significant increase in the absolute levels of tRNA genes in cancerous cells in comparison to healthy cells. However, we find that the relative composition of tRNA isoacceptors in healthy, cancerous, or transformed cells remains almost identical. This result may indicate that maintaining the relative tRNA composition in cancerous cells is advantageous via its stabilizing of the effectiveness of translation.

Keywords: codon usage; ENCODE; RNA polymerase III; noncoding RNA; translation elongation

INTRODUCTION

Translation elongation efficiency

Translation must be tightly controlled for coping with the cell demands and its limited resources. Energetically, it is the most expensive process in dividing cells (Arava et al. 2003; Ingolia et al. 2009; Plotkin and Kudla 2010; Gingold and Pilpel 2011). Thus, an appropriate regulation of the rate of translation reduces the ribosomal drop-off and the translation errors and improves overall ribosomal allocation (Zhang et al. 2010; Gingold and Pilpel 2011). The relative genomic abundance of the synonymous codons varies in all organisms from bacteria to mammals (Sharp and Matassi 1994; Stenico et al. 1994). Furthermore, codon usage in different genes tends to be related to their expression levels (Marais and Duret 2001; dos Reis et al. 2004; Plotkin and Kudla 2010; Tuller et al. 2010b). Specifically, highly expressed genes (e.g., ribosomal proteins) usually include codons that are recognized by more abundant tRNA molecules, suggesting that the control of the translation process is under selective pressure (Anderson 1969; Bulmer 1987).

In all organisms, less than 61 tRNA types carry out the decoding of all codons. For example, there are only 40 tRNA types (called “tRNA isoacceptors”) in Escherichia coli K12, 44 tRNA types in Drosophila melanogaster, 48 tRNA types in Caenorhabditis elegans, and 51 tRNA types in humans. The decoding of mRNA molecules to proteins in most organisms is therefore based on the presence of some tRNAs that use the same anticodon for recognizing more than one codon (according to the wobble restricted rules) (Percudani 2001; Duret 2002).

In unicellular organisms such as bacteria and fungi, the genomic tRNA copy number correlates with the intracellular tRNA levels (Ikemura 1981; Sorensen and Pedersen 1991; Dong et al. 1996; Percudani et al. 1997; Kanaya et al. 1999; Dittmar et al. 2004). Thus, in general, translation efficiency can be analyzed at the level of amino acids, codon usage, tRNA isoacceptors, and genomic tRNA copy number.

Technologies for large-scale quantifying protein levels have lagged behind the methodologies for mRNA level
quantification such as microarray or deep sequencing. Thus, a common assumption in the field is that the transcriptome signature of a cell is an appropriate reflection of its proteome. However, in mouse and humans, it has been shown that the mRNA levels explain only 27%–40% of the protein level variation (Ghazalpour et al. 2011; Schwanhaussler et al. 2011). In addition, the level of tRNA molecules is the best-known approximation for the translational rate and the efficiency of codon usage. Unfortunately, measuring the intracellular levels of tRNA molecules remains technologically challenging (Dittmar et al. 2006). Specifically, conventional technologies such as DNA microarray, tiling platform, and PCR-based sequencing methods fail to determine the expression of tRNA molecules. Because tRNAs are short and extensively modified molecules, the routine molecular manipulations (e.g., preparation of RNA) may not be straightforward (Juhling et al. 2009). Thus, currently, the genomic tRNA dosage has been used as a proxy for the actual tRNA cellular abundance (dos Reis et al. 2004; Man and Pilpel 2007; Tuller et al. 2011).

Measurements of short RNAs from deep sequencing platforms are archived in deepBase (Yang et al. 2009). Among these sequences, tRNAs occupied a substantial amount. These data provide an opportunity to study the relations between the actual tRNA measurements and translational-related properties such as the transcribed amino acids, the codon usage, and the tRNA genomic copy number.

The tRNA adaptation index (tAI) (dos Reis et al. 2004; Man and Pilpel 2007; Tuller et al. 2010b) is a measure of the adaptation of a gene (or a codon) to the cellular pool of tRNAs. In practice, to calculate this measure, the genomic tRNA copy number is combined with thermodynamic considerations of the codon–anticodon interaction (Man and Pilpel 2007; Tuller et al. 2010b). The tAI is based on the assumption that the concentrations of the tRNA molecules that recognize a codon have a strong effect on the efficiency and speed of translation.

In this study, we analyzed large-scale genomic and transcriptomic data that were generated by technologies that provide accurate measurements of the relative quantities of tRNA molecules. Based on these data, we found that the relative concentrations of tRNA molecules in different cell types and pathological states remain remarkably stable. We conclude that, for a wide variety of healthy, transformed, and cancerous cells, the tRNA molecules act as stabilizers by providing balanced tRNA pools that resemble the pools of healthy cells.

RESULTS

In the following sections, we report various analyses that we performed with the tRNA measurements. We have focused on human cells for which there are measurements of the intracellular tRNA levels and gene expression, under identical conditions.

Figure 1A illustrates the levels of resolution that were addressed in this study. There are 21 amino acids (including selanocysteine, which is encoded by the codon UAG). Each amino acid is decoded using the tRNA codon–anticodon hybridization and according to the restricted wobble rules (Percudani 2001). In human, there are 51 different isoacceptor types. When considering transcription of genes, each of the codons (64, including stop codons) is represented. In human, the number of tRNA genes for an individual tRNA isoacceptor is between one gene and 32 genes. The total number of tRNA genes in the genome is referred to as the “genomic tRNA copy number.” In human (according to hg19 version), there are potentially 512 functional tRNA genes that cover the 21 amino acids. An additional 100 pseudogenes are also found, but they will not be further discussed. Each tRNA is identified by its anticodon and a numeric value (e.g., tRNA13–AlaCGC). There are tRNA genes that share the same sequences throughout their length. According to this definition, the total number of unique tRNA genes is reduced to 434.

Our research flow is shown in Figure 1B. We start with the comparison of the experimental tRNA measurements that are based on several deep-sequencing resources and tRNA hybridization arrays. The experimental codon usage is calculated from the gene expression transcriptomic data of the analyzed cells. These data are collected from different human cells and tissues. Then, a comparison to the genomic tRNA copy number is performed. A refined estimation of the effect of the tRNA abundance on the efficiency of the translation rate of codons is achieved from the tRNA adaptation index (tAI; see Materials and Methods) (dos Reis et al. 2004). We further compare the tRNA genomic copy number and the codons that are used according to the transcriptomic data and according to various accepted measures of codon preference (Sharp and Li 1987; Duret 2000) (Materials and Methods). We investigate the correlation of the tRNA level measurements based on hybridization arrays and the deep-sequencing reads with the tRNA copy numbers (CN). All analyses were performed for healthy and transformed breast cell lines and for healthy and cancerous tissues (Fig. 1B).

tRNA abundance from RNA-seq correlates with the amino acid frequency

We have focused on cells for which we have different experimental measurements including the direct measurement of tRNA levels and the gene expression profiles that were collected under identical conditions. Figure 2 includes the correlation between the amino acid frequencies (as determined from the transcriptome data) and the number of tRNA genes that recognize the codons of each amino acid (e.g., 43 tRNA genes for Ala) (Fig. 1A). The data were collected from the MCF-10A epithelial breast cell line. A high correlation ($r = 0.765$, $P$-value $= 3.3 	imes 10^{-5}$) indicates that the number of tRNA genes that decode the codons of an amino acid is in accordance with their genomic abundance (Fig. 2).
tRNA abundance from RNA-seq correlates with the genomic tRNA copy number

We further tested whether such a trend occurs between the experimental measurements of the tRNA molecules and their genomic copy number. Data sets that include measurements of tRNA genes were generated by the main sequencing platforms. For such analyses, we include data from deepBase (Yang et al. 2009) that includes about 60 tRNA experiments (only 20 of them are of a substantial coverage). The entire collection of reads apparently covers all known tRNA genes (510 of the 512 genomic tRNAs). The correlation of tRNA reads and their genomic copy number was moderate but significant ($r = 0.34$, $P$-value = 0.012) (Supplemental Fig. S1). This unbiased analysis emphasizes the need for high-quality and high-coverage deep-sequencing data for accurate tRNA abundance measurements.

At the next stage, we focused on individual high-coverage experiments. Such an unbiased data set was extracted from the ENCODE short RNA-seq project (Washietl et al. 2007). Millions of reads for short RNA molecules (e.g., tRNA and miRNA) are reported. The data include short RNA sequences (20–200 nt) without poly(A) that were extracted from several human cell types.

We analyzed the 52,893 and 28,959 reads that were associated with the erythrocytic leukemia cells (K562) and the B-lymphoblastoid cells (GM12878), respectively. The two cell lines differ in their origins as well as in their chromatic state (Ernst et al. 2011). For each tRNA gene that was uniquely defined, we compared the number of reads that were recorded and its genomic copy number. The reads for the GM12878 cells according to their partitioning to tRNA isoacceptor groups (a total of 51) (Fig. 1A) are shown in Figure 3A. As can be seen in the figure, some tRNA isoacceptor groups are assigned to only a few reads, while others are assigned to >10% of the reads.

Comparison of the relative abundance of the 51 groups of tRNA reads to their genomic copy number is shown in Figure 3B. The two distributions are far from being identical. Specifically, the tRNA levels of the anticodons CAG and TTT are over-represented in the GM12878 data set, while many tRNAs are under-represented in these cells. Among the under-represented anticodons are GAA and GCA but also the anticodons for Ser...
(T/C)GA, Arg (T/C)CT, and more. Nevertheless, the correlation between experimentally measured tRNA levels (according to the unique reads) and their genomic copy number for the 5 isoacceptor groups was \( r = 0.3813 \) (\( P \)-value = 0.0058) (Fig. 3C).

A similar analysis was performed for the K562 cells. However, for the K562 cells, a poor and not significant correlation was measured. In these cells, the number of the tRNA isoacceptor types represented is low (34/49) (Supplemental Fig. S2). We suggest that the tRNAs from deep sequencing are a useful source for tRNA sampling and quantification. However, the experiments differ in terms of their reproducibility, coverage, and quality.

In the next step, we seek an independent source for tRNA sequences. The immunoprecipitation experiment using polymerase III (Pol III) fulfills this criterion (Raha et al. 2010). The isolated sequences represent an enriched fraction of genes that were directly attached to Pol III (Figs. 3D–F). The experiments included the K562 and GM12878 cell lines. In both cell lines, a significant correlation of the tRNA measurements and the 51 groups of the tRNA types clustered from the genomic copy number was obtained (Fig. 3F; Supplemental Fig. S2). Figure 3E shows the relative abundance of tRNA reads in comparison to their genomic copy number. The reported correlation for the GM12878 data set was high (\( r = 0.548 \), \( P \)-value = \( 3.13 \times 10^{-5} \)). A similar analysis was performed for the K562 cell line (Supplemental Fig. S2).

A direct correlation between the tRNA assigned reads obtained from the two experimental settings (ENCOD and Pol III) for the GM12878 cells is \( r = 0.475 \) (\( P \)-value = \( 4.28 \times 10^{-4} \)). This relatively high correlation suggests a sufficient consistency in the tRNA reads from the deep-sequencing methodology, despite substantial variations among the experiments (Supplemental Fig. S3).

**Fragmented tRNAs do not correlate significantly with the tRNA genomic copy number**

The RNA-seq methodology using deep sequencing can estimate the abundance of sequences that are related to the several steps in the life cycle of tRNA genes. For example, the 3′-tRNA trailers are subsequences from the 3′ end of the pre-tRNA. These tRNAs are cleaved by a specific endonuclease (RNaseZ) in the next step of the life cycle.

**FIGURE 2.** Correlation between the genomic copy number and the amino acid usage from epithelial normal cell line MCF-10A. Each codon from the cell transcriptome (10,132 identified expressed genes) was multiplied by the relative gene expression signal. The number of tRNA genes grouped by tRNA isoacceptor groups specifies each amino acid. The 21 amino acids are abbreviated according to standard convention, with SeCys denoting selanocysteine. The maximal value in the x-axis is for alanine (A) with a total of 43 tRNA genes.

**FIGURE 3.** Expression levels of tRNA types for GM12878 cells by RNA-seq sequencing technology. The total reads that match tRNA genes from GM12878 samples are indicated. (A–C) The ENCODE data set is based on short (20–200 nt) non-poly(A) RNA. (D–F) The data set was extracted from the RNA polymerase III (Pol III) immunoprecipitation experiment. The reads were normalized and presented as a sorted list according to the relative read values (A,D). (B,E) The list includes the relative log abundance above and below the expectation value according to the 51 available tRNA types (note that tRNAs that had no reads are not shown). The correlation of the absolute number of reads and the genomics tRNA copy numbers is plotted. The correlation coefficient \( r \) value and the \( P \)-value are indicated (C,F). Note that the correlations and the \( P \)-values were calculated from the original dot plots. The data are shown in a log scale for data compression graphical reasons.
during the process of tRNA maturation (Nashimoto 1997). In Liao et al. (2010), the presence of the 3′-tRNA trailers was quantified in the cytoplasm and the nucleus of the human nasopharyngeal carcinoma (NPC) 5-8F cell line. Interestingly, the estimated levels of the 3′-tRNA trailers showed no correlation with the genomic copy number of tRNA isoacceptors \( r = -0.1169, P\text{-value} = 0.4140 \). Importantly, the measurements of the 3′-tRNA trailers are of high quality, because the correlation between different samples (nuclear or cytosolic fractions) is extremely high \( r = 0.9461, P\text{-value} = 3.715 \times 10^{-24} \). Furthermore, the number of uniquely matched reads reaches nearly half a million, suggesting a high coverage of the experiment.

Another set of RNA-seq experiments was conducted on HeLa cells with the aim of tracing extremely short stable RNA species (<30 nt) (Cole et al. 2009). In this study, a surprisingly high number of tRNA fragments were detected. Despite the sequencing depth, only a few tRNAs dominate (e.g., tRNA isoacceptors for Lys, Val, Glu, and Arg). In this case, we did not get a significant correlation at the amino acid resolution \( r = 0.295, P\text{-value} = 0.18 \) (Supplemental Fig. S4).

**tRNAs detected by microarray hybridization strongly correlate with tRNA genomic copy number**

In a study published by Pavon-Eternod et al. (2009), a specific microarray for identifying tRNAs by the hybridization signal was designed. Each probe was designed to complement either a single tRNA type or a combination of some of the isoacceptor tRNAs. Consequently, all amino acids (except proline) were covered. Whereas in the study of Pavon-Eternod et al. the relative changes in the expression levels of tRNA (each gene compared to itself) were emphasized, in the present study we focused on the ranking between the expression levels of individual tRNAs. We have used the raw data (Pavon-Eternod et al. 2009) from the normal MCF-10A breast cell line to generalize our findings across technologies.

As mentioned above, some of the probes cross-react with several isoacceptor tRNAs. Consequently, the contribution of each specific tRNA cannot be deduced. Indeed, when the probe hybridization intensity was compared with its isoacceptor group according to the tRNA copy number, no correlation was found (MCF-10A cell line).

We therefore refined the data by applying a strict definition for the tRNAs and the probe sequences. We eliminated the data derived from the mitochondrial tRNAs and focused on the nuclear set. Furthermore, we filtered out the results by analyzing only the subset of probes that perfectly match only a single tRNA gene. Under such a criterion, we reduce the discussion to only 22 valid tRNA isoacceptors (see Materials and Methods). Following such probe–tRNA match selection, the correlation increases but remains insignificant \( r = 0.323, P\text{-value} = 0.1424 \).

Many tRNA genes share a remarkably high sequence similarity. Thus, at the next step, we considered only the 183 tRNA genes having an unequivocal hybridization potential by the 22 selected probes. The correlation of the hybridization intensity with these validated tRNA genes was considerably improved \( r = 0.4275, P\text{-value} = 0.047 \).

The strong dependency of the analysis on the selected probes encouraged us to increase the number of probes in the analyses. We thus used the definitions that were provided by Pavon-Eternod et al. (2009) for a uniquely matched probe. Thirty tRNA probes were considered based on this definition (some probes matching tRNAs within the same isoacceptor group).

The rest of the analysis is based on the uniquely defined 30 probes (associated with 25 different tRNA isoacceptor groups). Using the experimental hybridization intensities, we found that the correlation between these measured expression levels of the selected probes and the genomic copy number of the relevant tRNA isoacceptor groups was quite high \( r = 0.639, P\text{-value} = 1.4 \times 10^{-4} \) (Fig. 4).

**Correspondence of experimental data with various measures of codon usage**

In this section, we evaluate the comparison of the hybridization intensity to common scores of codon usage.

Specifically, we tested classical measures including the RGF (the relative tRNA gene frequency from the isoacceptor of a specific amino acid) and the partition between the synonymous codons as measured by RSCU (i.e., the relative usage of a codon among all codons for an amino acid). We included a measure of the dKL (Kullback–Leibler divergence) that provides an overall estimation for the similarities of the various measures (Prat et al. 2009). (For formal definitions, see Materials and Methods.)

Figure 5 shows that for most isoacceptor groups, the codons that had the highest RSCU within each amino acid were those that are decoded by the tRNA isoacceptor with the highest RGF values (the genomic tRNA copy numbers used are listed in Supplemental Table S1). The wobble-restricted rules do not explicitly include the affinity or specificity for the cognate codon–anticodon and the wobble codon. A naive view suggests an equal decoding capacity by a tRNA for the perfectly matched codon and the wobble-codon. While it is clearly an oversimplification, the results show that, for the 16/19 tRNA isoacceptor groups, a full correspondence of the top RGF and the top RSCU is achieved (note that the 6-based decoding of leucine, arginine, and serine was separated according to the genetic code table; see Materials and Methods). Some tRNA isoacceptors cannot be ranked either because the number of tRNA genes within a codon group is identical (Fig. 5B, green background) or because there is only one anticodon that corresponds to the amino acid (Fig. 5B, blue background).
The dKL calculates the difference in the distributions of the experimental tRNA measurements and the expressed codon as compiled from the gene expression transcriptomic data. The minimal dKL value for the distribution of genomic copy numbers and the codon usage is low (0.1717), supporting the relatedness between these distributions (Fig. 5C).

**Estimating the tRNA abundance using tAI-based measurements**

In this subsection, we report on the correlation of codon bias (based on gene expression microarray; see Materials and Methods) and tRNA levels or copy numbers. In the case of the MCF-10A cell line, the correlation between mRNA level codon bias and tRNA copy number was 0.38 ($P$-value = 0.0016). This correlation is based on the 51 tRNA type copy numbers that have at least one gene and an additional 12 codons that have no perfectly matched tRNA (for definitions, see Fig. 1A; Supplemental Table S2). The “missing” codons are decoded according to the restricted wobble rules (Fig. 5A).

We also tested the quality of the correlation of the mRNA codon usage (based on the transcriptomic data for MSF-10A cells) and the tAI values of the codons (based on tRNA copy number). The correlation coefficient of the tAI computed by genomic tRNA copy number and the codon usage according to the gene expression array was $r = 0.57$, $P$-value = 8.6 $\times$ 10$^{-7}$.

We repeated the calculation of the correlation between the partial set of experimental tRNA levels based on probe data (Pavon-Eternod et al. 2009) and the genomic copy number. However, in this stage, when values for the probe hybridization were missing (based on uniquely assigned 30 probes) (Fig. 4), we estimated them based on the relevant genomic tRNA copy number (see Materials and Methods). The combined tAI obtained a higher and more significant correlation of $r = 0.70$ ($P$-value = 1.4 $\times$ 10$^{-16}$; see Materials and Methods) with mRNA codon usage (cf. Fig. 6A,B).

The correlation between the two measures for the tAI is exceptionally high, supporting the validity of the approach used for the missing values ($r = 0.814$, $P$-value = 1.4 $\times$ 10$^{-16}$).

**The tRNAs and the codon usage vary in a coordinated way in different cell types**

Evidently, the human genomic tRNA copy number is identical in all cells (considering the normal karyotype). We have shown that the codon usage correlates with the hybridization intensity directly (Figs. 4, 5B,C) or through the refined tAI measure (Fig. 6A,B). We therefore analyzed the impact of the alteration in the expression levels in MCF-10A (mammary epithelial cells) and ZR-075 (a breast transformed cell line), which represent normal and cancerous breast cell lines, respectively. These cells were also used for measuring the tRNA abundance (see Materials and Methods). We have compared the mRNA codon bias with their genomic tRNA copy number and with the tRNA expression levels obtained based on hybridization (for the subset of 183 validated tRNA genes) (Fig. 7, indicated as tRNA probe 25).

Overexpressing genes are primarily composed of housekeeping genes that are optimally expressed. This optimization is monitored by the codon adaptation index (CAI), which is maximal for highly abundant genes (e.g., ribosomal proteins). We tested whether such support is also valid for the calculated tAI. We performed the analysis for three complementary data sets: (i) all genes in the gene expression array; (ii) the 200 most highly expressed genes; and (iii) the 200 most lowly expressed genes. The tests were performed for the normal breast cell line MCF-10A and the cancerous breast cell line ZR-075.

Figure 7 demonstrates that there is a strong correlation for the tRNA probes when we consider all expressed genes (more than 10,000 genes) with $r = 0.69$ ($P$-value = 1.3 $\times$ 10$^{-4}$) for the normal cells, and $r = 0.71$ for the cancerous cell lines ($P$-value = 7.5 $\times$ 10$^{-5}$).

When the 200 most highly expressed genes were analyzed separately, a slightly higher correlation between the mRNA codon bias and the genomic tRNA copy number ($r = 0.54$, $P$-value = 3.8 $\times$ 10$^{-3}$ for the normal cell line; $r = 0.53$, $P$-value = 0.05 for the cancerous cell line) was obtained. As shown (Fig. 6), when experimental tRNA levels and tRNA copy numbers are combined, the correlation becomes stronger (Fig. 7, labeled as tRNA probe 25).

Importantly, both cells show a similar correlation trend for multiple selections of tRNA measurements (cf. Fig. 7A).
We tested the direct correlation between the tRNA levels in these apparently different cell lines based on the experimental hybridization intensity (25 unique isoacceptors) (Fig. 7C). In a similar way, we applied the refined measure based on tAI that is based on tRNA hybridization experiments and tRNA copy numbers (when the data are missing) (Fig. 7D). In both instances, when the normal cells were compared with the cancerous cell line, an almost perfect correlation was revealed ($r = 0.98$, $P$-value $= 1.23 \times 10^{-3}$).

Human cell lines and tissues maintain a stable composition of tRNAs in pathological conditions

We studied whether the strong and consistent correlation found between tRNA levels and tRNA copy numbers in cell lines is comparable to these correlations in healthy and cancerous breast tissues (Table 1).

The analyzed cancerous tissues included three main subtypes of breast cancer (a total of nine samples): luminal (ER+, HER2$^-$), basal (ER-, HER2+), and ER$^-$/HER2$. When we computed the correlation of the tRNA probe hybridization intensity for the normal (average of three healthy tissues) and the cancerous tissues (average of nine breast cancer tissues), we found it to be significant ($r = 0.38$, $P$-value $= 0.037$).

At the next step, we analyzed the internal correlations between the 12 different tRNA measurements of the raw data in human tissue (Pavon-Eternod et al. 2009). Specifically, we tested the internal correlations, the mutual dKL, and the correspondence of the tRNA levels (measured by the hybridization intensity for the 25 uniquely defined tRNA isoacceptors) to the genomic tRNA copy number. The results according to the dKL calculations are shown in Figure 8. There is high similarity of the tRNA distributions among all tested tissues (Fig. 8; Supplemental Table S3). Unsupervised clustering supports two main clusters of the healthy (three samples) and the cancerous (nine samples) tissues. Interestingly, among the cancerous tissues, clustering of the results failed to indicate their cancer typing (in terms of the expression of ER and HER2).

The results (Fig. 8; Table 1; Supplemental Table S3) are consistent with the notion that while the absolute level of tRNAs had changed drastically (Pavon-Eternod et al. 2009), the relative abundance of each tRNA type is quite robust.
A stable tRNA composition is valid among cell lines and healthy and diseased tissues.

**DISCUSSION**

**Advances in tRNA measurements**

In this study, we analyzed tRNA measurements from RNA-seq that originated from all leading technological platforms (454, SoliD, and Illumina). Our analyses covered several cellular settings including healthy, transformed cell lines, and cancerous tissues. We focused on most of the reported experiments that measured tRNAs by different methodologies. These methodologies include the Pol III immunoprecipitation and a survey for short non-poly(A) RNA conducted by the ENCODE project. We had validated the feasibility of the deep-sequencing data to provide a reproducible source for tRNA concentrations. However, the high level of modifications in human tRNAs may lead to a failure in the required reverse transcription reaction. Consequently, all further steps in the sequencing protocol will be affected. Indeed, in the ENCODE data (Fig. 3; Raha et al. 2010), the tRNAs fraction occupies only 5% of all of the reads. We attribute this low recovery of tRNAs to an actual methodology limitation. The challenges and biases in identifying the small noncoding RNAs, including tRNA molecules by deep-sequencing technologies, were recently discussed (Beck et al. 2011).

**tRNA gene regulation**

The expression of a tRNA gene that is not subjected to a regulation is expected to be similar to its respective copy number. Genes with the highest skew from this rule may be candidates for some tissue-specific regulation. Figure 3, B and E, highlights such instances. For example, the tRNAs that recognize the anticodons CGA and TTT (Fig. 3B) are suspected to be the most up-regulated in the ENCODE project data of GM12878 cells, while the tRNA molecules that recognize the codons GAA and GCA are down-regulated in this tissue. Additional tRNA genes appear as outliers in the plot of tRNA levels versus tRNA copy number (for example, Fig. 6A,B). The results reported in this study suggest that individual tRNA genes are regulated more significantly than it was initially anticipated. The involvement of epigenetic signature and chromatin state is a plausible explanation for the observed difference in specific tRNA expression (Ernst et al. 2011).

**tRNA fragments do not exhibit a significant correlation with tRNA abundance**

We found that there is a poor correlation between tRNA copy numbers and tRNA fragments (Cole et al. 2009; Liao et al. 2010). Thus, in all of the reported studies, the relationship between measured fragments of tRNAs and translation elongation efficiency is not supported.

Recently, another study used deep sequencing (based on 454 sequencing) for estimating the expression levels of small RNAs from prostate cancer cell lines (Lee et al. 2009). Many of the reads were assigned to the processed tRNA fragments that were derived from the regions that overlap with the 3’-tRNA-trailers (Liao et al. 2010). Our results reject the hypothesis that the frequencies of these fragments reflect the expression levels of tRNA genes. Thus, these tRNA fragments may have cellular functions not related to translation. For example, it was suggested that tRNA fragments have a regulatory role in the apoptotic pathway (Mei...
et al. 2010) and in regulating cell proliferation (Lee et al. 2009). Furthermore, it was also suggested that the tRNA fragments act as regulators of miRNA by competing on the miRNA processing (Lee et al. 2009; Pederson 2010). Our analysis is in accord with the notion that tRNA fragments have some yet-unknown functionality (Okamura and Lai 2008).

A robust expression of tRNAs in various cell types

We have rigorously analyzed several types of cells including healthy, transformed, and cancerous tissues. The absolute tRNA level in the transformed cells is about 20-fold higher than the levels in normal cells (Pavon-Eternod et al. 2009). As we showed, the measured level of tRNA expression is in a strong accordance with the miRNA codon bias extracted from a global gene expression analysis.

We argue that the tRNAs that change their overall expression roughly maintain their relative concentrations upon a wide range of conditions. Moreover, the change in codon usage among cell lines of different identity is negligible (Fig. 7). The relatively constant ranking of the concentrations of tRNAs under a broad range of cells and conditions may indicate that fine-tuned tissue-specific changes in the gene translation rate are probably mostly a result of an additional layer of regulation such as epigenetic, transcriptional, and miRNAs, and not a result of a programmed change in the tRNA levels.

The first systematic analysis that was based only on a careful measure of the relative hybridization intensities (Pavon-Eternod et al. 2009) indicated that tRNAs carrying specific amino acids (such as S, T, and Y) are mostly overexpressed in breast cancer cell lines and breast tumors. In our study, we show that the ranked order of tRNAs (and not necessarily the total amount) is similar whether it is tested by direct experimental data (e.g., RNA-seq) or under a rich model that includes thermodynamic codon–anticodon parameters (e.g., tAI) (dos Reis et al. 2004). The same trends hold when we compare the RGF to the RSCU (Fig. 5). Thus, expression of a balanced, stable ranking of the isoacceptor tRNAs dominates our study.

A detailed study of tRNA relative expression in tissues and cell lines showed that specific tRNA isoacceptors have higher-than-expected variation in some tissues (Dittmar et al. 2006). Our results suggest that the variations in specific isoacceptors are insignificant relative to the overall trend showing a wide variation in the amount of the entire set of tRNAs for a number of tissues and cell lines (as measured in Fig. 8). In agreement with the results from this study, it was noted that the relative expression of tRNAs in HeLa and HEK293 cell lines is similar among the isoacceptors (Dittmar et al. 2006), even though they are derived from different tissues (cervix and embryonic kidney, respectively).

The high correlation between tRNA (or tAI) levels in cells with different transcriptomic profiles (Fig. 7C,D) supports

FIGURE 7. Correlation between codon usage and tRNA approximations. The correlation according to the tRNA copy number (CN) and the codon usage based on the MCF-10A transcriptome (A) and the same data analysis from the cancer ZR-75-1 cell line (B). The CN is based on those that were selected for the 25 tRNA isoacceptor groups. Note that these are identical to the 30 elected probes from the tRNA microarray experiments. The tRNA probes applied the actual measurements from the 30 probes that are associated with 25 tRNA isoacceptor groups. Note that few of the tRNA probes hybridize to the same isoacceptor tRNA type. The correlation with all genes (blue bar) concerns all expressed genes in the transcriptome (10,132 genes). (Red bar) The 200 most highly expressed genes in the array; (dark green bar) the 200 lowly expressed genes. (C) Correlation of tRNA probes hybridization intensity. (D) The tAI computed by the combination of the tRNA copy number with the hybridization intensity. The correlations performed for normal (MCF-10A) and cancerous (ZR-75-1) cell lines. The raw data were from Pavon-Eternod et al. (2009). The correlation and the P-values are reported.
fundamental robustness in the process of protein translation across a range of conditions and tissues. According to our findings, the global rate of translation is probably altered under pathological conditions, while the difference in the relative translation rates of specific genes is less likely to occur.

The correlation of the tRNA abundance and the codon usage is extremely robust in the healthy and transformed cells. Thus, we propose a model in which the tRNA levels change either by an excess of RNA Pol III, by instability of the karyotype, or other indirect cellular scenarios. However, the changes in tRNA gene expression are general and occur across the entire tRNA gene sets, maintaining the relative expression levels of tRNA genes. In support of this intuitively unexpected phenomena, we noted that also the correlation between the calculated tAIs of the normal cells and the transformed cells is extremely high ($r = 0.9784$, $P$-value = $3.879 \times 10^{-44}$). Thus, as a first approximation, one can order genes according to their translation efficiency in the same conditions by considering the genomic tRNA copy numbers and pre-calculated factorization of the codons (as in Fig. 6). However, because we expect global changes (and changes that do not affect the ranking of the expression levels of individual tRNAs) in tRNA levels across tissues or conditions, the tRNA copy number should not be used for a comparison of translation efficiency of a specific gene across different conditions/tissues.

Breast cancerous tissues of different origin display similar tRNA composition

Probably, cell lines were adapted for a stable, constant growth. This may lead to a loss of regulatory mechanisms of the tRNA genes while maximizing the expression levels of each tRNA isoacceptor to meet the constant need for cell divisions. However, we showed that the strong and significant resemblance of the tRNA expression levels also occurs in human healthy and cancerous tissues. All of the 12 samples (three healthy and nine cancerous) show a strong correlation. However, the correlations (and the dKL) of the tRNA levels with the tRNA copy number are considerably weaker for each of the tested samples (Fig. 8; Supplemental Table S3). The average value (Table 1) demonstrates that, despite a strong coherence in the results among all tissues, the genomic tRNA copy number is not a perfect approximation of the tRNA abundance. Several reasons may have reduced the correlations (Fig. 8; Table 1): (1) Data from the 30 tRNA probes (covering 25 tRNA isoacceptors) may not be an optimal sampling for the entire 64 codons and may include different sources of noise and bias. (2) The tRNA copy number may be inaccurate. For example, the identification of functional tRNAs is based on algorithmic arbitrary thresholds, and it is known that the functionality of some tRNAs remains uncertain (Lowe and Eddy 1997). In addition, some pseudogenes that have been excluded from the copy number calculation are expressed (based on the deepBase data) (Yang et al. 2009). (3) Different tRNA genes have different levels of regulations that reduce the correlation between copy number and tRNA levels.

Based on the results reported in this study, we conclude that, in human, the genomic tRNA copy number is a reliable and valid approximation for their expression levels. Thus, when performing a large-scale transcriptomic study, the tRNA copy number can be safely used for estimating global translation efficiency. However, we showed that data from deep sequencing or tRNA microarrays are useful be-

![Figure 8](image)

**FIGURE 8.** Calculation of the dKL of the tRNA expression levels in healthy and diseased tissues. The values of the dKL measures are shown by a color gradient (black to red). The calculations are based on the hybridization signals from the 30 unique tRNA probes for healthy (three samples) and cancerous tissue samples (nine samples). The symmetric matrix indicates the clustering of the 13 columns in the matrix. The diagonal is indicated as dKL = 0. (The left column and the bottom row of the matrix) The dKL for the tRNA hybridization intensity and the genomic tRNA copy number (CN). (Red) A weaker correspondence (higher dKL value). The columns are sorted based on the clustering. The correlations, the minimum dKL, and the $P$-values are listed in Supplemental Table S3. The samples are colored by their labels as ER$^+$/HER2$^+$: 59826, 60046, 62706, 62944 (blue); ER$^-$/HER2$^-$: 46258, 58955 (red); ER$^+$/HER2$^+$: 41299, 57731, 45163 (orange); and healthy breast tissues: A-01, A-06, and S-23 (green). Note that there is no clear separation between ER$^+$/HER2$^+$ and ER$^-$/HER2$^-$ by this measure.
cause they potentially improve the estimation of the tRNA levels. At present, accurate measurements of processed, functional tRNAs are still fragmented and mostly missing. Thus, high-quality data are urgently needed.

Our comprehensive study that is based on collecting most available experimental data led to new insights on translation efficiency in a wide range of cellular settings. Ample research studies showed that due to genomic instability and changes in chromatin structure, the expression of hundreds of genes is altered in cancer relative to healthy cells. In sharp contrast, at the level of translation, such alterations seem to be tamed and attenuated. Our findings argue that a regulation of tRNA expression is not at the gene level or the chromosomal level, but instead it is performed globally on the entire collection of tRNA genes. We raised the notion that translation, being the most energetically expensive process in dividing cells, acts as a stabilizer that maintains a balanced translation potential even under unstable cellular conditions.

MATERIALS AND METHODS

Genomic copy number

The data of genomic tRNA copy number, chromosomal locations, and the sequence identity tRNA genes were taken from the Genomic tRNA Database using human genome hg19 (NCBI Build 37.1, Feb 2009) (Lowe and Eddy 1997). For each tRNA gene, the number of copies was counted, ignoring pseudogenes but including selanocysteine (for a detailed list, see Supplemental Table S1).

tRNA gene counting

The convention is based on the algorithms described in Lowe and Eddy (1997). The tRNA probe reanalysis is based on a replacement of the degenerate base. The degenerative probes are indicated by one-letter codes (http://www.bioinformatics.org/sms/iupac.html). Each tRNA is designated by the anticodon that is depict by three bases. For consistency, all codons and anticodons are described with the base thymidine (T) instead of uridine (U). For example, the tRNA Met (CAT) decodes the codon ATG. We kept the notation of the tissues used by Pavon-Eternod et al. (2009).

RNA-seq sequencing data of tRNAs

tRNA 3’-trailers

The data for the tRNAs 3’-trailers were based on a cytoplasmic and a nuclear extraction from human nasopharyngeal carcinoma (NPC) 5-8F (Liao et al. 2010).

ENCODE

The data were downloaded from the small RNA [non-poly(A)] ENCODE project at http://genome.ucsc.edu/. The tRNA measurements were collected from the K562 and GM12878 cell lines. Each validated read was assigned to the appropriate genomic location (contigs), and the number of reads that were detected for each genomic sequence was recorded. All sequences that overlap (even partially) a tRNA gene were considered. The chromosomal location of each tRNA gene was based on the data of the UCSC Genome Browser (Raney et al. 2010).

Pol III immunoprecipitation

Reads of tRNA from the K562 and GM12878 cells were collected using an antibody against Pol III (Raha et al. 2010). The number of reads for each tRNA gene was used as a measure for the tRNA abundance.

deepBase

deepBase compiles 59 individual experiments; among them about 20 are of high coverage (more than 1 million reads). A total of 625 tRNAs are reported as expressed genes by deep sequencing. The list includes 99 genes that were assigned as pseudogenes. Out of the complete list of tRNAs, only 12 were unidentified. It is important to note that for sequence-identical tRNAs, no unique identification is possible, and therefore some of the reads should be considered as a sum of the multiple tRNAs that are identical in sequence.

Analyzing tRNA probe hybridization intensity data

The data of tRNA probe hybridization intensity were taken from Pavon-Eternod et al. (2009). The set of probes was used to measure the differences in tRNA levels between the three samples of MCF-10A (used as reference for the epithelial normal cell line) and nine additional samples from different stages and types of breast cancer tissues (Pavon-Eternod et al. 2009).

For each biological sample, two probe arrays with a mixture of Cy3 and Cy5 dyes were used to produce the hybridization data. To reduce the staining bias, the cell line that was used as a reference (a non-cancer-derived breast epithelial MCF-10A) was dried with each of the two dyes, separately. The variation between the measured hybridization intensities was used to estimate the bias. While the experiment was not designed for estimating absolute measurements of tRNA genes, we considered the average of the hybridization intensities measured from the dye swapped arrays as a quantity of the tRNA hybridization. The complete data set (Pavon-Eternod et al. 2009) covers the 50 different nuclear tRNA probes. However, among these probes, only 30 were designed to match a single tRNA type. The rest were designed to match several tRNA isoacceptors. The 30 unique probes were used throughout our study. This set covers 25 of the isoacceptor groups.

Additional filtration was applied to ensure a perfect match for the hybridization reaction. To this end, the sequence of each probe was aligned using the tRNA genomic BLAST tool. The aligned sequences were filtered to include only the tRNA genes that had a perfect matched alignment of ≥20 sequential nucleotides. Only 22 legitimate probes passed this filter. We refined the assignment of a tRNA to its probe by defining higher constraints on the probe–tRNA hybridization.

Gene expression and codon usage analysis

Data from gene expression arrays of a normal cell line (MCF-10A) and a cancerous cell line (ZR-075) were used in this study. Genes that were reported as expressed were retrieved from Ensembl (Flicek et al. 2010). The Ensembl annotation of each probe was
based on the HG-U133A Affymetrix annotation file. Genes with no matched sequence or without an ATG initiator codon were discarded. Out of 22,215 potential human genes, 10,132 genes were further analyzed. The mRNA codon usage was computed by accumulating over all of the genes the product of the codon occurrences in each gene multiplied by the actual expression intensity of the gene. The intensity was retrieved from the relevant HG-U133A microarray. A naive view on gene expression is based on the Ensembl gene list (22,215 potential human genes). This analysis was applied for instances in which detailed transcriptomic data were missing.

**tRNA gene frequency (RGF) and relative synonymous codon usage (RSCU)**

For each tRNA isoacceptor that encodes a specific amino acid, the relative gene frequency was computed in the following way: In each amino acid isoacceptor group, let \( CN_i \) be the copy number of the \( i \)-th isoacceptor. Let \( M_{cui} \) be the average of the copy number of all isoacceptor within the \( j \)-th group. The \( i \)-th isoacceptor is part of the \( j \)-th group (Fig. 1A). The RGF of isoacceptor \( i \), \( RGF_i \), is defined as:

\[
RGF_i = \frac{CN_i}{M_{cui}}
\]

We partitioned the six-based codons of Leu, Ser, and Arg to their subgroups according to the genetic code table. The partition to six-based codons to four- and two-based codons. A total of 24 isoacceptor groups were considered; among them there are 21 isoacceptor groups with two or more codons.

Similarly, the relative synonymous copy number was computed: Let \( CU_i \) be the codon usage of codon \( i \). Let \( M_{cu} \) be the average of the codon usage within a group of the \( j \)-th amino acid synonymous codons. The \( i \)-th isoacceptor is part of the \( j \)-th group.

The RSCU of each codon, \( RSCU_i \), is defined as:

\[
RSCU_i = \frac{CU_i}{M_{cu}}
\]

**Kullback–Leibler divergence (dKL)**

The difference of the probability distribution between two data sets was computed using the Kullback–Leibler divergence definition (Prat et al. 2009). Let \( P \) and \( Q \) be the probability distribution of each data set. We applied this measure for normalized distribution of the copy numbers and the codon usage.

The dKL is defined as:

\[
dKL = \text{ave}(\sum P_i \log(P_i/Q_i) + \sum Q_i \log(Q_i/P_i))
\]

**Computing the tRNA adaptation index (tAI)**

tAI was computed according to dos Reis et al. (2004). This measure gauges the availability of tRNAs for each codon along an mRNA. Because codon–anticodon coupling is not unique due to wobble interactions, practically, several anticodons can recognize the same codon, with somewhat different efficiency.

Let \( n_i \) be the number of tRNA isoacceptors recognizing codon \( i \). Let \( tCGN_{ij} \) be the copy number of the \( j \)-th tRNA that recognizes the \( i \)-th codon, and let \( S_{ij} \) be the selective constraint on the efficiency of the codon–anticodon coupling. We define the absolute adaptiveness \( W_i \) for each codon \( i \) as:

\[
W_i = \sum_{j=1}^{ni} (1 - S_{ij}) tCGN_{ij}
\]

From \( W_i \) we obtain \( w_i \), which is the relative adaptiveness value of codon \( i \), by normalizing the \( W_i \) values (dividing them by the maximal of all the 61 \( W_i \)).

**Computing tAI for missing data**

Reliable experimental data are limited to the 30 tRNA genes for which 30 unique probes are used to estimate the tRNA levels. The rest of the tRNA genes are estimated based on the genomic copy number. In this case, for a tRNA that has no matching probe, the value of the tRNA abundance was normalized with the total genomic tRNA copy number in the following manner: Let \( t_p \) be the hybridization intensity value of probe \( i \). Let \( M_p \) be the mean of all of the probe hybridization intensities. Let \( M_{cu} \) be the mean of all genomic tRNA copy numbers. The tRNA abundance estimator \( tE_i \) for tRNAi is defined as:

\[
tE_i = \begin{cases} 
(t_p/M_p)/M_{cu} & \text{if there is data of tRNA, abundance :} \\
\text{Genomic copy number} & \text{Otherwise :}
\end{cases}
\]

**SUPPLEMENTAL MATERIAL**

Supplemental material is available for this article.

**ACKNOWLEDGMENTS**

We thank Mariana Pavon-Eternod for her assistance and personal help throughout this study. We thank Nathan Linial for his advice and useful suggestions. The work is partially supported by grants (to M.L.) Prospects (EU FrVII), the ISF 592/07, and the BSF 2007219 (to M.L.). S.M. is a fellow of the SCCB, the Sudarsky Center for Computational Biology.

Received October 6, 2011; accepted January 2, 2012.

**REFERENCES**


Duret L. 2000. tRNA gene number and codon usage in the *E. coli* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet* **16**: 287–289.


4.1.1. Appendix

Conservation of the relative tRNA compositions in healthy and cancerous tissues

Shelly Mahlab¹,&, Tamir Tuller²,*,&, Michal Linial³,*,&

1 School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel

2 Iby and Aladar Fleischman Faculty of Engineering, Department of Biomedical Engineering, Tel Aviv University, Israel.

3 Dept of Biological Chemistry, Institute of Life Sciences, Sudarsky Center for Computational Biology, The Hebrew University of Jerusalem, Israel

* To whom correspondence should be addressed.

* Equal contribution.
Table S1. The genomic tRNA copy number according to the 62 codons and anticodons

<table>
<thead>
<tr>
<th>Amino acids</th>
<th>Anti-codon</th>
<th>Codon</th>
<th>Copy Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>AGC</td>
<td>GCT</td>
<td>29</td>
</tr>
<tr>
<td>Ala</td>
<td>TGC</td>
<td>GCA</td>
<td>9</td>
</tr>
<tr>
<td>Ala</td>
<td>CGC</td>
<td>GCG</td>
<td>5</td>
</tr>
<tr>
<td>Ala</td>
<td>GCC</td>
<td>GCC</td>
<td>0</td>
</tr>
<tr>
<td>Arg</td>
<td>ACG</td>
<td>GTG</td>
<td>7</td>
</tr>
<tr>
<td>Arg</td>
<td>TCT</td>
<td>AGA</td>
<td>6</td>
</tr>
<tr>
<td>Arg</td>
<td>TCG</td>
<td>CGA</td>
<td>6</td>
</tr>
<tr>
<td>Arg</td>
<td>CCT</td>
<td>ACG</td>
<td>5</td>
</tr>
<tr>
<td>Arg</td>
<td>CGG</td>
<td>CGG</td>
<td>4</td>
</tr>
<tr>
<td>Arg</td>
<td>GCG</td>
<td>CGC</td>
<td>0</td>
</tr>
<tr>
<td>Asn</td>
<td>GTT</td>
<td>AAC</td>
<td>32</td>
</tr>
<tr>
<td>Asn</td>
<td>ATT</td>
<td>AAT</td>
<td>2</td>
</tr>
<tr>
<td>Asp</td>
<td>GTC</td>
<td>GAC</td>
<td>19</td>
</tr>
<tr>
<td>Asp</td>
<td>ATC</td>
<td>GAT</td>
<td>0</td>
</tr>
<tr>
<td>Cys</td>
<td>GCA</td>
<td>TGC</td>
<td>30</td>
</tr>
<tr>
<td>Cys</td>
<td>ACA</td>
<td>TGT</td>
<td>0</td>
</tr>
<tr>
<td>Gln</td>
<td>CTG</td>
<td>CAG</td>
<td>20</td>
</tr>
<tr>
<td>Gln</td>
<td>TGT</td>
<td>CAA</td>
<td>11</td>
</tr>
<tr>
<td>Glu</td>
<td>TTC</td>
<td>GAA</td>
<td>13</td>
</tr>
<tr>
<td>Glu</td>
<td>CTC</td>
<td>GAG</td>
<td>13</td>
</tr>
<tr>
<td>Gly</td>
<td>GCC</td>
<td>GGC</td>
<td>15</td>
</tr>
<tr>
<td>Gly</td>
<td>TCC</td>
<td>GGA</td>
<td>9</td>
</tr>
<tr>
<td>Gly</td>
<td>CCC</td>
<td>GGG</td>
<td>7</td>
</tr>
<tr>
<td>Gly</td>
<td>ACC</td>
<td>GTG</td>
<td>0</td>
</tr>
<tr>
<td>His</td>
<td>GTG</td>
<td>CAC</td>
<td>11</td>
</tr>
<tr>
<td>His</td>
<td>ATG</td>
<td>CAT</td>
<td>0</td>
</tr>
<tr>
<td>Ile</td>
<td>AAT</td>
<td>ATT</td>
<td>14</td>
</tr>
<tr>
<td>Ile</td>
<td>TAT</td>
<td>ATA</td>
<td>5</td>
</tr>
<tr>
<td>Ile</td>
<td>GAT</td>
<td>ATC</td>
<td>3</td>
</tr>
<tr>
<td>Leu</td>
<td>AAG</td>
<td>CTT</td>
<td>12</td>
</tr>
<tr>
<td>Leu</td>
<td>CAG</td>
<td>CTG</td>
<td>10</td>
</tr>
<tr>
<td>Leu</td>
<td>TAA</td>
<td>TTA</td>
<td>7</td>
</tr>
<tr>
<td>Leu</td>
<td>CAA</td>
<td>TTG</td>
<td>7</td>
</tr>
<tr>
<td>Leu</td>
<td>TAG</td>
<td>CTA</td>
<td>3</td>
</tr>
<tr>
<td>Leu</td>
<td>GAG</td>
<td>CTC</td>
<td>0</td>
</tr>
<tr>
<td>Lys</td>
<td>CTT</td>
<td>AAG</td>
<td>17</td>
</tr>
<tr>
<td>Lys</td>
<td>TTT</td>
<td>AAA</td>
<td>16</td>
</tr>
<tr>
<td>Met</td>
<td>CAT</td>
<td>ATG</td>
<td>20</td>
</tr>
<tr>
<td>Phe</td>
<td>GAA</td>
<td>TTC</td>
<td>12</td>
</tr>
<tr>
<td>Phe</td>
<td>AAA</td>
<td>TTT</td>
<td>0</td>
</tr>
<tr>
<td>Pro</td>
<td>AGG</td>
<td>CCT</td>
<td>10</td>
</tr>
<tr>
<td>Pro</td>
<td>TGG</td>
<td>CCA</td>
<td>7</td>
</tr>
<tr>
<td>Pro</td>
<td>CGG</td>
<td>CGG</td>
<td>4</td>
</tr>
<tr>
<td>Pro</td>
<td>GGG</td>
<td>CCC</td>
<td>0</td>
</tr>
<tr>
<td>Ser</td>
<td>AGA</td>
<td>TCT</td>
<td>11</td>
</tr>
<tr>
<td>Ser</td>
<td>GCT</td>
<td>AGC</td>
<td>8</td>
</tr>
<tr>
<td>Ser</td>
<td>TGA</td>
<td>TCA</td>
<td>5</td>
</tr>
<tr>
<td>Ser</td>
<td>CGA</td>
<td>TCG</td>
<td>4</td>
</tr>
<tr>
<td>Ser</td>
<td>ACT</td>
<td>AGT</td>
<td>0</td>
</tr>
<tr>
<td>Ser</td>
<td>GGA</td>
<td>TCC</td>
<td>0</td>
</tr>
<tr>
<td>stop</td>
<td>TCA</td>
<td>TGA</td>
<td>3</td>
</tr>
<tr>
<td>stop</td>
<td>TTA</td>
<td>TAA</td>
<td>2</td>
</tr>
<tr>
<td>stop</td>
<td>CTA</td>
<td>TAG</td>
<td>1</td>
</tr>
<tr>
<td>Thr</td>
<td>AGT</td>
<td>ACT</td>
<td>10</td>
</tr>
<tr>
<td>Thr</td>
<td>TGT</td>
<td>ACA</td>
<td>6</td>
</tr>
<tr>
<td>Thr</td>
<td>CGT</td>
<td>ACG</td>
<td>6</td>
</tr>
<tr>
<td>Thr</td>
<td>GGT</td>
<td>ACC</td>
<td>0</td>
</tr>
<tr>
<td>Trp</td>
<td>CCA</td>
<td>TGG</td>
<td>9</td>
</tr>
<tr>
<td>Tyr</td>
<td>GTA</td>
<td>TAC</td>
<td>14</td>
</tr>
<tr>
<td>Tyr</td>
<td>ATA</td>
<td>TAT</td>
<td>1</td>
</tr>
<tr>
<td>Val</td>
<td>CAC</td>
<td>GTG</td>
<td>16</td>
</tr>
<tr>
<td>Val</td>
<td>AAC</td>
<td>GTT</td>
<td>11</td>
</tr>
<tr>
<td>Val</td>
<td>TAC</td>
<td>GTA</td>
<td>5</td>
</tr>
<tr>
<td>Val</td>
<td>GAC</td>
<td>GTC</td>
<td>0</td>
</tr>
</tbody>
</table>
Table S2. tRNA gene frequency (RGF) and relative synonymous codon usage (RSCU).

<table>
<thead>
<tr>
<th>aa</th>
<th>anti codon</th>
<th>codon</th>
<th>RGF (cn)</th>
<th>RSCU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>AGC</td>
<td>GCT</td>
<td>2.045474</td>
<td>1.056763</td>
</tr>
<tr>
<td></td>
<td>TGC</td>
<td>GCA</td>
<td>0.837209</td>
<td>0.915095</td>
</tr>
<tr>
<td></td>
<td>GCG</td>
<td>GGG</td>
<td>0.465116</td>
<td>0.428597</td>
</tr>
<tr>
<td></td>
<td>GGC</td>
<td>GCG</td>
<td>0.599545</td>
<td></td>
</tr>
<tr>
<td>Arg (I)</td>
<td>ACG</td>
<td>CGT</td>
<td>1.647059</td>
<td>0.55916</td>
</tr>
<tr>
<td></td>
<td>TCG</td>
<td>CGA</td>
<td>1.411765</td>
<td>0.767182</td>
</tr>
<tr>
<td></td>
<td>CCG</td>
<td>CGG</td>
<td>0.941176</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCG</td>
<td>CCG</td>
<td>0.614518</td>
<td>1.263594</td>
</tr>
<tr>
<td>Arg (II)</td>
<td>TCT</td>
<td>AGA</td>
<td>1.090909</td>
<td>1.014585</td>
</tr>
<tr>
<td></td>
<td>CCT</td>
<td>AGG</td>
<td>0.909091</td>
<td>0.985344</td>
</tr>
<tr>
<td>Asn</td>
<td>GTT</td>
<td>AAC</td>
<td>1.682353</td>
<td>1.07459</td>
</tr>
<tr>
<td></td>
<td>ATT</td>
<td>AAT</td>
<td>0.117647</td>
<td>0.945251</td>
</tr>
<tr>
<td>Asp</td>
<td>GTC</td>
<td>GAC</td>
<td>2.000000</td>
<td>0.998288</td>
</tr>
<tr>
<td></td>
<td>ATC</td>
<td>GAT</td>
<td>0.943748</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCA</td>
<td>TGC</td>
<td>2.086692</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ACA</td>
<td>TGT</td>
<td>0.913308</td>
<td></td>
</tr>
<tr>
<td>Gin</td>
<td>CTG</td>
<td>CAG</td>
<td>1.290323</td>
<td>1.49149</td>
</tr>
<tr>
<td></td>
<td>TGG</td>
<td>CAA</td>
<td>0.709677</td>
<td>0.510851</td>
</tr>
<tr>
<td>Glu</td>
<td>TTC</td>
<td>GAA</td>
<td>1</td>
<td>0.848341</td>
</tr>
<tr>
<td></td>
<td>TCT</td>
<td>GAG</td>
<td>1</td>
<td>1.151659</td>
</tr>
<tr>
<td></td>
<td>GCC</td>
<td>GGC</td>
<td>1.905484</td>
<td>1.456593</td>
</tr>
<tr>
<td></td>
<td>TCC</td>
<td>GGA</td>
<td>1.16129</td>
<td>0.991999</td>
</tr>
<tr>
<td></td>
<td>CCC</td>
<td>GGG</td>
<td>0.903226</td>
<td>0.975528</td>
</tr>
<tr>
<td></td>
<td>ACC</td>
<td>GGT</td>
<td>0</td>
<td>0.657879</td>
</tr>
<tr>
<td>His</td>
<td>GTC</td>
<td>CAT</td>
<td>2</td>
<td>0.872451</td>
</tr>
<tr>
<td></td>
<td>ATG</td>
<td>CAT</td>
<td>0</td>
<td>0.837445</td>
</tr>
<tr>
<td></td>
<td>AAT</td>
<td>ATT</td>
<td>1.090909</td>
<td>1.095682</td>
</tr>
<tr>
<td></td>
<td>TAT</td>
<td>ATA</td>
<td>0.681818</td>
<td>0.488055</td>
</tr>
<tr>
<td></td>
<td>GAT</td>
<td>ATC</td>
<td>0.409091</td>
<td>1.416264</td>
</tr>
<tr>
<td>Leu (I)</td>
<td>AAG</td>
<td>CTT</td>
<td>1.92</td>
<td>0.667011</td>
</tr>
<tr>
<td></td>
<td>CAG</td>
<td>CTA</td>
<td>1.6</td>
<td>0.815552</td>
</tr>
<tr>
<td></td>
<td>TAG</td>
<td>CTA</td>
<td>0.48</td>
<td>0.355207</td>
</tr>
<tr>
<td></td>
<td>GAG</td>
<td>CTC</td>
<td>0</td>
<td>0.959119</td>
</tr>
<tr>
<td>Leu (II)</td>
<td>TAA</td>
<td>TTA</td>
<td>1</td>
<td>0.346073</td>
</tr>
<tr>
<td></td>
<td>CAA</td>
<td>TTG</td>
<td>1</td>
<td>0.58945</td>
</tr>
<tr>
<td>Lys</td>
<td>CTT</td>
<td>AAG</td>
<td>1.090909</td>
<td>1.126672</td>
</tr>
<tr>
<td></td>
<td>TTT</td>
<td>AAA</td>
<td>0.999697</td>
<td>0.857162</td>
</tr>
<tr>
<td>Met</td>
<td>CAT</td>
<td>ATG</td>
<td>2</td>
<td>1.070779</td>
</tr>
<tr>
<td></td>
<td>GAA</td>
<td>TTC</td>
<td>0</td>
<td>0.929221</td>
</tr>
<tr>
<td></td>
<td>AAA</td>
<td>TTT</td>
<td>0</td>
<td>0.929221</td>
</tr>
<tr>
<td>Pro</td>
<td>AGG</td>
<td>CCT</td>
<td>1.904762</td>
<td>1.145952</td>
</tr>
<tr>
<td></td>
<td>TGG</td>
<td>CCA</td>
<td>1.333333</td>
<td>1.106244</td>
</tr>
<tr>
<td></td>
<td>CGG</td>
<td>CCG</td>
<td>0.761905</td>
<td>0.462532</td>
</tr>
<tr>
<td></td>
<td>GGG</td>
<td>CCC</td>
<td>0</td>
<td>1.282597</td>
</tr>
<tr>
<td>Ser (I)</td>
<td>AGA</td>
<td>TCT</td>
<td>2.2</td>
<td>1.224916</td>
</tr>
<tr>
<td></td>
<td>TGA</td>
<td>TCA</td>
<td>1</td>
<td>0.984756</td>
</tr>
<tr>
<td></td>
<td>CGA</td>
<td>TCG</td>
<td>0.8</td>
<td>0.374714</td>
</tr>
<tr>
<td></td>
<td>GGA</td>
<td>TCC</td>
<td>0</td>
<td>0.312576</td>
</tr>
<tr>
<td>Ser (II)</td>
<td>ACT</td>
<td>AGC</td>
<td>1.221195</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGT</td>
<td>ACT</td>
<td>0</td>
<td>0.778005</td>
</tr>
<tr>
<td>Sup</td>
<td>TCA</td>
<td>TGA</td>
<td>1.5</td>
<td>1.497319</td>
</tr>
<tr>
<td></td>
<td>TTA</td>
<td>TAA</td>
<td>1</td>
<td>0.840384</td>
</tr>
<tr>
<td></td>
<td>CTA</td>
<td>TAG</td>
<td>0.5</td>
<td>0.662297</td>
</tr>
<tr>
<td>Thr</td>
<td>AGT</td>
<td>ACT</td>
<td>1.090909</td>
<td>0.995433</td>
</tr>
<tr>
<td></td>
<td>TGT</td>
<td>ACA</td>
<td>1.13252</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CTG</td>
<td>ACG</td>
<td>1.090909</td>
<td>0.465509</td>
</tr>
<tr>
<td></td>
<td>GGT</td>
<td>ACC</td>
<td>0</td>
<td>1.406538</td>
</tr>
<tr>
<td>Trp</td>
<td>CCA</td>
<td>TGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAA</td>
<td>TAA</td>
<td>1</td>
<td>1.123834</td>
</tr>
<tr>
<td></td>
<td>ATA</td>
<td>TAT</td>
<td>0.133333</td>
<td>0.886566</td>
</tr>
<tr>
<td>Val</td>
<td>CAC</td>
<td>GTG</td>
<td>1.375375</td>
<td>0.736809</td>
</tr>
<tr>
<td></td>
<td>CAT</td>
<td>GTT</td>
<td>1.375375</td>
<td>0.736809</td>
</tr>
<tr>
<td></td>
<td>TAC</td>
<td>GTA</td>
<td>0.625</td>
<td>0.474919</td>
</tr>
<tr>
<td></td>
<td>GAC</td>
<td>GTC</td>
<td>0</td>
<td>0.929736</td>
</tr>
</tbody>
</table>

*Note:* The first ranked RGF can wobble it, but it has its first ranked RGF cant wobble it.
Table S3. Source data for the calculated dKL, correlations and p-values for the tRNA compositions among all 12 tested breast tissue samples and CN.

Measurements include the correlations and the p-values for the hybridization intensities based on 30 uniquely assigned probes for 12 breast tissues. The CN column and row represent the analysis of the hybridization intensities with the genomics tRNA copy number (CN).

**Table S3**

*dKL matrix*

<table>
<thead>
<tr>
<th></th>
<th>normal breast tissue</th>
<th>breast cancer tissue</th>
<th>CN</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-01</td>
<td>0.075127</td>
<td>0.112349</td>
<td>0.285808</td>
</tr>
<tr>
<td>A-06</td>
<td>0.637946</td>
<td>0.586359</td>
<td>0.981072</td>
</tr>
<tr>
<td>S-23</td>
<td>0.923659</td>
<td>0.743659</td>
<td>0.665231</td>
</tr>
</tbody>
</table>

**Correlation matrix**

<table>
<thead>
<tr>
<th></th>
<th>normal breast tissue</th>
<th>breast cancer tissue</th>
<th>CN</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-01</td>
<td>0.916675</td>
<td>0.890022</td>
<td>0.732639</td>
</tr>
<tr>
<td>A-06</td>
<td>0.829</td>
<td>0.798649</td>
<td>0.864167</td>
</tr>
<tr>
<td>S-23</td>
<td>0.762664</td>
<td>0.743825</td>
<td>0.813902</td>
</tr>
</tbody>
</table>

**P-value matrix**

<table>
<thead>
<tr>
<th></th>
<th>normal breast tissue</th>
<th>breast cancer tissue</th>
<th>CN</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-01</td>
<td>1.14e-12</td>
<td>1.36e-10</td>
<td>4.16e-06</td>
</tr>
<tr>
<td>A-06</td>
<td>0.606e-09</td>
<td>9.36e-08</td>
<td>1.24e-07</td>
</tr>
<tr>
<td>S-23</td>
<td>0.96e-07</td>
<td>6.49e-06</td>
<td>9.36e-09</td>
</tr>
</tbody>
</table>

130
Figure S1. Expression levels of tRNA types for many of the deep sequencing experiments

The reads that match tRNA genes from 59 experiments is shown. Data was retrieved from deepBase. (A) A relative log abundance above and below the expectation value according to the 51 available isoacceptor tRNA types (note that tRNA that had no reads are not shown). (B) The correlation of the reads and the genomics tRNA copy numbers are plotted. The correlation coefficient (r) value and the statistical significant (p-value) are indicated (C, F). Note that the correlations and the p-values were calculated from the original dot plots. The data are shown in a log scale for data compression graphical reasons.
Figure S2. Expression levels of tRNA types for K562 cells by deep sequencing technology

The reads that match tRNA genes from K562 samples are counted and the total reads values are indicated. (A-C) The ENCODE dataset from small (20-200 nt) non-polyA RNA. (D-F) The dataset was extracted from the RNA Polymerase III (Pol-III) immuno-precipitation experiment. The reads were normalized and presented as a sorted relative reads value (A, D), a relative log abundance above and below the expectation value according to the 51 available tRNA types (note that tRNA that had no reads are not shown) (B, E). The correlation of the absolute number of reads and the genomics tRNA copy numbers are plotted. The correlation coefficient (r) value and the statistical significant (p-value) are indicated (C, F). Note that the correlations and the p-values were calculated from the original dot plots. The data are shown in a log scale for data compression graphical reasons.
Figure S3. Correlations of the reads from K562 cells and GM12878 by deep sequencing technology

The reads that match tRNA genes from K562 (A) and GM12878 (B) cells are counted and the total reads values are indicated. The correlations are based on the ENCODE and the RNA Polymerase III (Pol-III) immuno-precipitation experiment. The correlation coefficient (r) value and the statistical significant (p-value) are indicated. Note that the correlations and the p-values were calculated from the original dot plots. The data are shown in a log scale for data compression graphical reasons.
Figure S4. Expression levels of amino acids for tRNA fragments from Hela Cells

The reads that match tRNA fragments are based on {Cole, 2009 #234}. (A) A relative log abundance above and below the expectation value according the amino acids resolution. (B) The correlation of the reads and the genomics tRNA copy numbers are plotted. The correlation coefficient(r) value and the statistical significant (p-value) are indicated. Note that the correlations and the p-values were calculated from the original dot plots. The data are shown in a log scale for data compression graphical reasons.
4.2. Speed Controls in Translating Secretory Proteins in Eukaryotes - an Evolutionary Perspective

This chapter includes the following publication:

Abstract

Protein translation is the most expensive operation in dividing cells from bacteria to humans. Therefore, managing the speed and allocation of resources is subject to tight control. From bacteria to humans, clusters of relatively rare tRNA codons at the N'-terminal of mRNAs have been implicated in attenuating the process of ribosome allocation, and consequently the translation rate in a broad range of organisms. The current interpretation of “slow” tRNA codons does not distinguish between protein translations mediated by free- or endoplasmic reticulum (ER)-bound ribosomes. We demonstrate that proteins translated by free- or ER-bound ribosomes exhibit different overall properties in terms of their translation efficiency and speed in yeast, fly, plant, worm, bovine and human. We note that only secreted or membranous proteins with a Signal peptide (SP) are specified by segments of “slow” tRNA at the N’-terminal, followed by abundant codons that are considered “fast.” Such profiles apply to 3100 proteins of the human proteome that are composed of secreted and signal peptide (SP)-assisted membranous proteins. Remarkably, the bulk of the proteins (12,000), or membranous proteins lacking SP (3400), do not have such a pattern. Alternation of “fast” and “slow” codons was found also in proteins that translocate to mitochondria through transit peptides (TP). The differential clusters of tRNA adapted codons is not restricted to the N’-terminal of transcripts. Specifically, Glycosylphosphatidylinositol (GPI)-anchored proteins are unified by clusters of low adapted tRNAs codons at the C’-termini. Furthermore, selection of amino acids types and specific codons was shown as the driving force which establishes the translation demands for the secretory proteome. We postulate that “hard-coded” signals within the secretory proteome assist the steps of protein maturation and folding. Specifically, “speed control” signals for delaying the translation of a nascent protein fulfill the co- and post-translational stages such as membrane translocation, proteins processing and folding.

Introduction

In dividing cells, the process of translation elongation consumes most of the cell energy and resources [1–3]. The rate of translation must be tightly controlled for coping with the cell demands and its limited resources. Specifically, translation efficiency is determined by the amount of proteins that are produced from the coding mRNA. In a more mechanistic view, translation efficiency is reflected by the preferable allocation of ribosomes on the mRNA [4]. Sequence-based features such as mRNA folding energy, positioning of individual amino acids (AAs) and codons govern the translation efficiency [5–7]. Failure in coordinating the ribosomal flow leads to ribosomal drop-off [3], translation errors [8], frameshift [9] and protein misfolding [10]. Direct measurements of ribosome density from in vivo studies confirmed that translational rates differ between transcripts [11]. Moreover, the rate may vary by several folds on the same mRNA [2,12,13].

Several factors govern protein translation rate and accuracy (see discussion in [3,14,15]). A dominant parameter in dictating translation rate is the nature of the codons at the initial segment of the transcripts [16]. Other features include the competition on ribosome binding [17], mRNA folding energy [5], accessibility of specific tRNAs [18] and CG content [5]. A dominating parameter of translation efficiency from E. coli to human is the codon usage [19,20]. The coding usage of a broad range of organisms positively correlates with cellular proteins’ expression levels and thus, indirectly, with translation efficiency [21,22].

In all eukaryotes, the decoding of mRNAs to proteins obeys the same rules [23]. The genomic tRNA copy number (CN) strongly correlates with the needs for intracellular tRNA levels [24]. This property is best captured by the tRNA adaptation index (tAI) [19] that balances between the decoding rules and the tRNA CN [25]. Indeed, in humans, tAI appropriates the actual abundance of tAI in healthy and diseased cells [26].

In eukaryotes, a distinction should be made between proteins that are translated by the soluble, cytosolic ribosome (CYTO-Rb) and the membrane-bound ribosomes (MEM-Rb). The latter cover the proteins destined to the secretory systems (endoplasmic reticulum (ER), Golgi, endosomes, lysosomes, plasma membrane and the extracellular space) [27]. A common feature of the secretory proteins is the presence of signal peptide (SP) at the N’-terminal [28]. Alternatively, membranous proteins that lack SP (e.g., many G-protein coupled receptors) use their first TMD as a membrane signal. Translation of the secretory proteins at the ER
Author Summary

Measurements of translation by ribosomal profiling and additional large-scale methods support the notion that the elongation speed and ribosomal occupancy are tightly regulated. We revisited the proteomes of a number of organisms, from yeast to human, and focused on the appearance of codons' clusters that impact the speed of translation elongation. Thus, transcripts are analyzed according to their encoded "traffic signs." Specifically, translation by free- or endoplasmic reticulum (ER)-bound ribosomes differs substantially with respect to the codon clusters' distribution at the beginning of the coding region. Discretization of all transcripts to consecutive segments exposed the uniqueness of secreted and membranous proteins that have a signal peptide (SP). Similarly, a non-random codon distribution characterized proteins with "targeting peptides" for mitochondria and for GPI-anchor, while the bulk of the proteome carry no significant pattern of their codons. We conclude that translation via an ER co-translocation process imposes unique constraints on translation efficiency that match with the fate of the proteins as secreted, membranous, mitochondrial-targeted or GPI-anchored. Tuning the translation of a nascent protein is essential for coping with the constraints imposed by membrane-bound translation for a successful ER translocation and protein processing for maturation and folding.

membranes is a multiphase process that is based on coordinated steps of translation, translocation and folding [13,29,30].

In this study, we hypothesized that proteins of CYTO-Rb and MEM-Rb translation differ in their translation elongation management. A local tRNA adaptation pattern at the N'-terminal which starts with segments of lowly adapted tRNAs, followed by segments of highly adapted tRNAs, is characteristic of secreted and membranous SP-proteins but not identified in the bulk of the proteins or in other regions of the transcripts. Such patterns are shared by a large number of eukaryotic proteomes and found also in proteins that are designated to the mitochondria. The impact of "traffic signs" on the management of translating ER-bound ribosomes is discussed in view of recent experimental evidence on translation rates.

Results

Translation elongation efficiency is approximated by tRNA adaptation index

An estimation of the effect of the tRNA abundance on the efficiency of the translation is captured by the tRNA adaptation index (tAI) [See Materials and Methods]. The pairing of tRNA with the mRNAs is not unique in the case of the Wobble pairing (Figure 1A). Each organism differs by the number and the relative appearance of tRNA isoacceptors for decoding the 20 amino acids (AAs, 61 codons). Synonymous codons are associated with a broad range of tAI values (Figure 1B). Some AAs (e.g., Arginine) are encoded by 6 codons but the range of their tAI values is still very narrow. On the other hand, a broad range of tAI values is associated with AAs that have only two codons each (e.g., Asparagine and Cysteine) (Figure 1B).

The tRNAs copy number (CN) is subjected to evolutionary forces and thus differs substantially throughout the evolutionary tree. For example, there are 287 tRNA genes in the budding yeast S. cerevisiae but as many as 3790 tRNA genes in Bos Taurus. The tAI value that is assigned to each codon varies substantially among different organisms. While the correlations among human, D. melanogaster, C. elegans are moderate, the correlations with B. taurus or A. thaliana (flowering plant) are negligible (Figure 1C). The tAI codon values for each organism is listed in Table S1.

Translation efficiency marks are encoded in the human secretory proteome

The translation of proteins in eukaryotes is executed in two settings: Proteins that are translated by free ribosomes (coined cytoplasmic ribosome, CYTO-Rb) and ER bound ribosomes (coined membranous ribosome, MEM-Rb). We partitioned the entire proteomes into four non-overlapping groups (Table S2):

(i) Signal Peptide (SP) proteins that are not located at the membrane (SP not TMD). These are mostly secreted proteins (e.g., hormone peptides, growth factors).

(ii) SP proteins with TMD. These are proteins that contain at least one TMD but are transllocated to the ER via an SP recognition mode. Additional step leads to a protein maturation following the removal of the SP (e.g., HLA class I histocompatibility antigens, Cadherins).

(iii) Integral membrane proteins that lack SP (TMD not SP). The initial TMD is used for insertion of the protein to the translocon pore (i.e., translocon). The topology of these proteins is determined by the presence of a stop signal along the sequence. The first TMD serves as an anchor signal.

(iv) Proteins that lack SP or TMD and are translated by free ribosomes (CYTO-Rb, simply refer to as “Cytosolic”).

Recall that the final destination of these proteins may not be restricted to the cytosol (e.g, nuclear proteins).

Groups (i-iii) compose the secretory proteome (Figure 2A). The human proteome consists of 18,434 proteins. Among them 26% include at least one TMD and an additional 9.5% are secreted proteins that contain SP. A similar partition is reported for fly, worm and bovine (Figure 2B) and other model organisms. The tAI of each coding sequence is computed (see Materials and Methods), and the average “global tAI” for the analyzed proteins’ group was defined (see Materials and Methods). Each of the three protein groups that together compose the secretory proteome displays a distinct global tAI (Figure 2C). For example, the p-value of the human secreted proteins (marked as “SP-not TMD” group) relative to membranous proteins without SP (TMD not SP) is 2.58e-11. The calculated p-values of the secreted proteins with respect to membranous proteins with SP (TMD and SP) and the cytosolic group are 1.0e-14 and 9.0e-12, respectively.

Comparing the average global tAI values for the secretory and cytosolic protein groups in different organisms is shown in Table 1. The main observation (Figure 2C) demonstrates that secreted proteins that have SP tend to have higher global tAI relative to the proteins of the membranous groups (TMD, with or without SP). While the absolute values of the global tAI are different for each organism (based on codon tAI, Table S1), the trend of low tAI for the membranous proteins relative to the secreted proteins is surprisingly robust (Figure 2C). We extended the analysis to include also yeast and plant representatives. The average values of the calculated global tAI values for (i) cytosolic proteins, (ii) SP-no TMD (iii) SP and TMD and (iv) TMD not SP are listed in Table S3.

We show the statistical significance among each pair of the protein groups for 6 organisms (Table 1). The statistical difference between the two exclusive sets of membranous proteins (with/without SP) is minimal (with p-value>1.0e-4, Table 1). For
Figure 1. **tRNA isoacceptors and adaptation index.** (A) Illustration of the decoding by tRNA. The alanine (Ala) charged tRNAs that recognize GCU and GCC belong to the same isoacceptor. Decoding is performed according to the wobble rules [73]. Alanine (Ala) is decoded by three groups of isoacceptor tRNAs. The genomic tRNA copy number (CN) from *H. sapiens* is marked. Specifically, the number of genes for Ala is 43 (the sum of the CN of all isoacceptor groups). Codons are always read by the 5' to 3' directionality from DNA or mRNA. (B) The range of codon tAI that can be assigned to each AA in *H. sapiens* is shown. Codon tAI is determined by the CN of tRNAs for that codon and according to the coupling of tRNA at the wobble position. The tAI for each codon is marked by a colored dot. Tryptophan (Trp) and Methionine (Met) are encoded by a single codon. For the other AAs...
example, the p-values of the global tAI values for the yeast-secreted proteins relative to other groups range from 1.67e-12 to 6.48e-23 (Table 1). A striking observation is that secreted proteins and the soluble fraction (i.e., CYTO-Rb translation) specify high average global tAI values with regard to the membranous proteins. A similar trend was observed in all six tested organisms (included yeast and flowering plant, Table S3).

**Global tAI correlates with mRNA expression levels and protein abundance**

Many determinants govern the protein abundance in eukaryotic cells [11]. The contribution of sequence-dependent determinants to the rates of translation and degradation has been estimated [31]. A positive correlation between the gene tAI and its expression was determined from the signature of gene expression microarrays [32]. We tested whether the average higher global tAI that was associated with the secreted (SP non-TMD) and the cytosolic proteins (Table S3) relative to membranous proteins reflects a difference in the expression levels. We took advantage of the experiments with high coverage of the yeast proteome and compared the protein abundance and the global tAI. We used a resource from mass spectrometry (MS) peptide counts [33] (total of 4012 proteins, Figure 3A) and the quantitative data from GFP-tagged proteins [34] (total of 2279 proteins, Figure 3B). We found substantial agreement between the results from these complementary technologies (compare 3A and 3B). The strongest correlation was noted between the global tAI values and the cytosolic proteins. However, the significance of the correlation between the global tAI and the proteins of the secreted proteome is rather weak (SP

A range of tAI values are shown according to the number of codons (2, 3, 4 and 6 codons). Note that Arg, Ser and Leu that are decoded by 6 codons each, do not necessarily have a wide range of tAI values. The minimal and maximal tAI values for each of the AAs are colored green and red, respectively. (C) Clustering of multicellular model organisms by the correlation calculated according to a vector of the tAI values (61 codons). The Spearman correlation coefficient between each pair of species is color-coded. The tAI codon values for each organism is listed in Table S1.

doi:10.1371/journal.pcbi.1003294.g001

**Figure 2. The secretory proteome.** (A) Partition of the secretory proteome with respect to membrane topologies is shown. The secreted proteins (red background) contain a Signal peptide (SP, red string) that is cleaved in the ER lumen. The site for cleavage by the SP protease is colored yellow. The membranous fraction is divided according to the presence (purple background) or absence (green background) of SP. All the three groups are translated by MEM-Rb. (B) Pie diagrams show the partition of the secretory proteome: (i) Proteins that have TMD but lack of SP sequence (TMD non-SP), (ii) Proteins that have SP but each protein has one or more TMD (SP and TMD) and (iii) Secreted proteins with SP in their N-terminus (marked in red, SP non-TMD). The rest of the proteins are soluble proteins that are translated by CYTO-Rb. The majority of the secretory proteome in all the 4 model organisms - human (**H. sapiens**), fly (**D. melanogaster**), worm (**C. elegans**) and bovine (**B. taurus**) are membranous proteins without SP (green). For these proteins, ER translocation is mediated via internal TMDs. For the detailed number of proteins in each organism see Table S2. (C) Average global tAI values for each groups of the secretory proteome as in (B). The histograms show analysis of the entire secretory proteomes from **H. sapiens**, **D. melanogaster**, **C. elegans** and **B. taurus**. Similar trends apply for Yeast (**S. cerevisiae**) and plant (**A. thaliana**). The statistical significance is based on the p-value calculated from the Kolmogorov–Smirnov (KS) test. The statistical significance are marked by asterisks. With p-values E-5 to E-10 (*) and <E-10 (**), (for detailed statistical analysis see Table 1).

doi:10.1371/journal.pcbi.1003294.g002

PLOS Computational Biology | www.ploscompbiol.org 4 January 2014 | Volume 10 | Issue 1 | e1003294
not TMD). We suggest that the relatively high global tAI is associated with an overall expression level for the majority of the proteins that are translated by free ribosomes (i.e., accounts for 78% and 81% of the analyzed proteins, Figure 3A and 3B, respectively). However, a high expression level is not supported for the secreted protein group. Additional parameters such as protein length, AA usage and CG content were also tested. The length of the proteins from the group “SP and TMD” was significantly longer than the rest of the proteins (P value = 1e-4). But the secreted proteins group (SP not TMD) and the “TMD not SP” group that differs in their tAI (Figure 2C) have no difference in protein length (p value = 0.133). All other correlations show a borderline statistical significance. We concluded that the tAI is strongly associated with protein abundance only for the cytosolic proteins. The same trend was found for the human proteome (data analyzed from [35]).

A robust signal at the N’-terminal specifies the secreted proteome

The secreted proteins showed significantly higher global tAI values (Figure 2C, Table S3). We tested the possibility that the tested protein groups may carry segmental information in addition to their global tAI values. To analyze the segmental properties of the proteomes, we discretized the transcripts to segments of 30 codons. The same notations were applied for the C’-terminus, starting from the last codon of the protein (Figure 4A). The results are presented as “Relative tAI,” which is defined as the current segments’ tAI divided by the calculated value of the global tAI of the coding sequence. This measure allows comparing the trends among organisms. Using the Relative tAI values (and not the absolute tAI values) cancels out the inherent difference in expression levels that are associated with the tested proteins groups (Figures 2–3).

Among the analyzed model organisms, the annotations for the human proteome are accurate and complete. According to the four groups partition (Figure 2B and the cytosolic fraction), the SP-containing proteins are characterized by an occurrence of lowly adapted tRNAs (coined LATS) at the N’-terminal (~45 codons) followed by highly adapted tRNAs (HATS) (Figure 4B). Notably, proteins that contain SP with or without TMD display a similar profile. All protein groups converged at segment N3 (codons position 60–90, Figure 4B). It is important to note that the “Relative tAI” profile of the entire proteome (combined all 4 groups, marked “All”, Figure 3B) shows no outstanding position-based pattern. Additional segments (e.g., N4) provided no additional information and will not be discussed further.

Figure 4C shows the cumulative distribution of tAI values for each of the analyzed protein groups for N1 and C1 segments from a human proteome. The statistical difference between the N1 and C1 segments is significant (Table 2). Actually, both the N1 and the C1 segments differ significantly from a random selection of a 30-codon segment (Kolmogorov-Smirnov (KS) test, Figure 3C, Table 2). The calculated p-values versus the random sets range between 1.0e-15 to 1.0e-22 for N1, and 1.0e-12 to 1.0e-27 for C1. More importantly, the statistical tests show significant p-values (7.6e-6 to 2.1e-57) for the characteristics of the N1 segment among the four protein groups, while the p-values for the C1 segments are statistically insignificant (Table 2).

The tAI segmental analysis was extended to other model organisms including B. taurus (Figure 4D), D. melanogaster (Figure 4E), C. elegans (Figure 4F) and S. cerevisiae (Figure 4G). Assessing the significance of the differences in the “Relative tAI” values for the different segments of the four protein groups is achieved by comparing the maximal range of the computed average relative tAI among the four groups. For example, the “Average Relative tAI” of the N1 in H. sapiens spans as much as 0.053 while the C1 deviates by only 0.007. We demonstrated these range differences of N1, N3 and C1 segments for all the tested organisms (Figure 4H). A similar pattern is generalized and the range of “Average Relative tAI” of N1 is significantly higher than that of N3 or C1. In this view, the range in values of segment N3 is considered a statistical noise.

As many of the secreted proteins (e.g., hormones, growth factors) are short proteins, we tested the effect of protein length on the observed segmental tAI profile. We confirmed that the impact of the protein length of the segmental local tAI is negligible. Specifically, we partitioned the SP-proteins to very short (90–240 AAs) and very long (>1,000 AAs) protein groups. We found that the trend of the tAI profiles is insensitive to the length. The “very short” and “very long” proteins originated from the same distribution (t-test, p-value = 0.72).

We tested the differential tAI segmental profiles of membranous proteins (composed of the groups of “SP and TMD” and “TMD not SP”) according to the separation to single (marked as types I–IV) and multi-pass proteins (Figure 5A). This type of partition tests whether the membrane topology governs the characteristics of the tAI segmental profile (shown in Figures 4B–4G). It is evident that the existing of SP dominates the profile irrespectively to the number of TMDs or the protein topology within the membrane (Figure 5B). The analysis is limited to yeast and humans due to the poor annotations on membranous protein topologies for the other model organisms.

Table 1. Statistical KS tests for the global tAI values that were calculated for 6 model organisms’ proteomes.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Groups</th>
<th>TMD non-SP</th>
<th>SP and TMD</th>
<th>Cytosolic</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. sapiens</td>
<td>SP non-TMD</td>
<td>2.58e-11</td>
<td>1.08e-14</td>
<td>9.01e-12</td>
</tr>
<tr>
<td></td>
<td>TMD non-SP</td>
<td>0.0374</td>
<td>4.84e-4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SP and TMD</td>
<td>2.18e-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. taurus</td>
<td>SP non-TMD</td>
<td>3.96e-34</td>
<td>8.11e-9</td>
<td>9.18e-3</td>
</tr>
<tr>
<td></td>
<td>TMD non-SP</td>
<td>9.6e-14</td>
<td>7.42e-85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SP and TMD</td>
<td>2.22e-15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>SP non-TMD</td>
<td>8.33e-10</td>
<td>5.36e-6</td>
<td>3.3e-4</td>
</tr>
<tr>
<td></td>
<td>TMD non-SP</td>
<td>0.488</td>
<td>7.19e-29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SP and TMD</td>
<td>8.82e-11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. elegans</td>
<td>SP non-TMD</td>
<td>7.79e-22</td>
<td>5.01e-11</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>TMD non-SP</td>
<td>2.31e-4</td>
<td>2.84e-57</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SP and TMD</td>
<td>1.22e-13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>SP non-TMD</td>
<td>6.48e-23</td>
<td>1.67e-12</td>
<td>3.3e-16</td>
</tr>
<tr>
<td></td>
<td>TMD non-SP</td>
<td>0.476</td>
<td>1.14e-11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SP and TMD</td>
<td>6.6e-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. thaliana</td>
<td>SP non-TMD</td>
<td>1.12e-17</td>
<td>1.24e-30</td>
<td>1.34e-6</td>
</tr>
<tr>
<td></td>
<td>TMD non-SP</td>
<td>3.28e-13</td>
<td>2.67e-10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SP and TMD</td>
<td>2.77e-28</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Partition of the proteomes to 4 exclusive groups is according to UniProtKB annotations for TMD and SP. Statistical significance <1.0e-5 is shown in bold. doi:10.1371/journal.pcbi.1003294.t001
membranous proteins that lack SP, the first TMD acts as the anchor signal. We further tested whether a codon dependent signal is encoded in the TMD. To this end, we aligned all sequences from the “TMD not SP” group by their first TMD (Figure S1). We found that the segmental tAI values of the first TMD differs from the observation of the SP-proteins. Actually, the

**Figure 3. Correlation between yeast protein abundance and global tAI.** (A) Mass spectrometry (MS) data were from the yeast quantitative proteome [33]. Protein abundance is measured from the match of the MS peptide-spectrum. Each spectrum is associated with a peptide that is re-assigned to its parent protein. The analysis covered 4012 proteins divided as follows: SP non-TMD: 87; TMD non-SP: 582; SP and TMD: 60; Cytosolic: 3283. (B) Quantitative proteomics [34] was measured by estimating the fluorescence from the tagged-GFP. The analysis covered 2279 proteins divided as follows: SP non-TMD: 67; TMD non-SP: 383; SP and TMD: 47; Cytosolic: 1782. The protein abundance and the global tAI are plotted and the correlation coefficient (r) and the p-values are indicated.

doi:10.1371/journal.pcbi.1003294.g003
within the same segment.

Table 2. Statistical differences (KS test) between segments of tAI values for partition of the human proteome and randomized sequences.

<table>
<thead>
<tr>
<th></th>
<th>SP non-TMD</th>
<th>TMD non-SP</th>
<th>SP and TMD</th>
<th>Cytosolic</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP non-TMDa</td>
<td>1</td>
<td>7.57e-06</td>
<td>1.25e-07</td>
<td>2.77e-28</td>
<td>3.83e-20</td>
</tr>
<tr>
<td>TMD non-SP</td>
<td>0.000722</td>
<td>1</td>
<td>6.83e-23</td>
<td>3.49e-13</td>
<td>3.98e-15</td>
</tr>
<tr>
<td>SP and TMD</td>
<td>0.001162</td>
<td>0.149803</td>
<td>2.08e-57</td>
<td>1.60e-22</td>
<td></td>
</tr>
<tr>
<td>Cytosolic</td>
<td>0.030616</td>
<td>0.004063</td>
<td>0.007944</td>
<td>1</td>
<td>1.79e-22</td>
</tr>
<tr>
<td>Random</td>
<td>1.33e-22</td>
<td>8.41e-23</td>
<td>1.11e-12</td>
<td>4.47e-27</td>
<td>1</td>
</tr>
</tbody>
</table>

Upper and lower triangles are based on 30-codon segments identified as N1 and C1, respectively. Statistical significance <1.0e10 is shown in bold. doi:10.1371/journal.pcbi.1003294.t002

“anchored TMD” shares no local tAI characteristics. We concluded that it is not the hydrophobicity per se that dictates the local tAI properties but instead, the SP sequences are characterized by clusters of lower adapted codons followed by clusters of highly adapted segments.

Generalizing speed controls toward organelle destination and subcellular localization

The robust phenomena of differential codon usage according to their tAI property along the transcript is not restricted to the N’-terminal segment. The Glycosylphosphatidyl inositol (GPI) anchored proteolipids reach the ER through an SP dependent process. For these proteins, an additional modification occurs following a proteolytic cleavage at a C’-terminal peptide of the nascent peptide [36]. We tested whether a signal for GPI lipid anchoring is encoded by segmental tAI measurements.

We separated the proteins that are predicted as GPI-anchor proteins [37]. Figure 6A shows a histogram for the cleavage site with respect to the last codon (marked as codon 0). In the majority of the cases, the cleavage sites are positioned within the C1 segment (codon marked as -25). The average segmental tAI for the 128 human GPI-proteins is shown (Figure 6B). Remarkably, the AAs composition of the GPI-anchor proteins is poorly conserved. Still, the GPI-anchor proteins are characterized by the significance of LATS at their final segment (C1, ~30 codons, Figure 6B). Thus, GPI-anchor proteins are marked by evolutionary signals at both, the N’- and C’-termini.

As opposed to the previously mentioned cases of GPI-anchored and SP-proteins that are modified at the ER on the nascent chain, translocation of mitochondrial proteins occurs as a post-translational stage. Hundreds of proteins reach the different compartments of the mitochondria (and chloroplasts in plants) by sophisticated mechanisms [38,39]. Many of these mitochondrial targeted proteins have a cleavable Transit Peptide (TP) in their N’-terminals. There are 499 proteins annotated to have TP in humans. Figure 6C shows the cleavage sites with respect to the initiator Methionine. For the majority of the proteins, the cleavage sites are positioned within the N1 or the N1-intermediate segments. The similarity of the local segmental tAI to the profile of the SP-proteins is evident (Figure 6D). TP adopts a more extreme value (“Relative tAI” of 0.95 in H. sapiens) for an extended segment relative to the SP-proteins (Figure 6D).

An overlap in the segmental profiles for the SP and TP protein is striking. Figure 6E demonstrates that when the AA compositions of the SP and the TP are compared, the overlap in the AAs usage is minimal. These results postulate as to the generality of the phenomenon. Notably, the marked difference in codon usage of the TP and SP segments argues for an unrestricted selection that supports a pattern of LATS followed by HATS. Such a design may be used as a general trend for management of protein targeting to sub-cellular compartments and organelles.

The profile at the N’-terminal segments is determined by preferred selection of codons

A key sequence feature of the SP is the central helical region that is dominated by Leu and Ala with some occurrence of Val, Phe and Ile. We show that the SP proteins have a preferable use of some amino acids (e.g., Leu and Trp), but a limited use of Asn, Asp, Ser, Thr and Arg.

There are two possible explanations for the observed profile at the N1-segment of the proteins with SP sequences: (i) The AAs that determine the SP are enriched with “slower” codons (i.e., lower tAI codon values); (ii) The codons at the initial segment that compose the SP reflect an evolutionary selection process. Both explanations may fulfill the global demands of MEB-Rb translation mode. In order to distinguish between these possibilities, we counted the codon usage in the SP of each of the relevant proteins, and the codon usage in segments of non-SP proteins. For some codons, the deviation between the usage in SP and non-SP is substantial (Figure 7A). For example, the use of Cys is preferable in SP-proteins, while Lys is rarely used in the segment that covers the SP sequence. Additionally, we tested the existence of an evolutionary signal that can account for the preferential selecting of codons in the N’-terminal segments of the SP-proteome. This is performed for any AA, regardless of its actual tendency to be used. Specifically, we questioned whether a selected codon in the SP sequence is randomly chosen from a background of the complete proteome codon usage data.

We show the preferred usability of a specific codon in view of its tAI value (Figure 7A, empty frames). For example, the AA valine (Val) is encoded by four codons. Among these codons, the codons that are mostly used for the SP-proteins are the ones with low tAI values (codons GTG) while the ones with maximal tAI value (codon GTG) are rarely used (Figure 7A). In order to assess the statistical power of such observations, we compared the actual local tAI for the SP segment (as in Figure 7A) with that of
simulated sequences that are composed of identical amino acids but are encoded by codons that were randomly selected from their synonymous codons, according to the tAI distribution in the entire genome (Figure 7B). While the tAI distributions are quite similar (dKL = 0.001), the mean value of the actual SP local tAI value was lower with respect to the randomized sequences (0.3143 and 0.3209 for the original SP and the synonymous codons tAI 1000 randomized tests, respectively). Importantly, the distributions differ significantly from the replaced sequences according to the codon usage distribution (p-value = 1.3e-07).

We concluded that in addition to the preselected AAs for the SP sequences (Figure 7A), an evolutionary signal is attributed to the selection of preferred codons in the SP sequences (Figure 7B).

Prototypic profiles of translational efficiency - the human proteome

The N-terminal segmental profile of SP proteins dominated over 3,100 protein sequences in humans (Figures 4B–4G). To ensure an unbiased analysis of the human proteome, we clustered by means of an unsupervised mode all ~18,400 human proteomes according to their segmental tAI profile (illustrated in Figure 4A). We focused on clusters that are dominated by LATS at the N1 segment (Figure 8, clusters 1–4). Enrichment tests according to the clusters’ annotations were performed. The most significantly enriched cluster’s annotation consists of secreted, signal, glycoprotein and disulfide-bridge (p-value of enrichment is 5.4e-18). An additional set of enriched annotations includes the plasma membrane and membranous proteins. These annotations are fully consistent with MEM–Rb translation (for a detailed analysis, see Table S4). Therefore, the clusters of most significant LATS values followed by HATS are associated with secreted proteins, membranous proteins, extracellular matrix and receptors, all of which belong to SP-containing proteins.

Based on a global, unbiased clustering, proteins that are signified by a characteristic pattern are identified. For example, a profile with several consecutive HATS (Figure 8 cluster 6,170...
Figure 6. Analysis of the local segmental tAI profiles for GPI-anchored and Transit peptide (TP)-proteins. (A) Histogram of the cleavage site relative to the end of the coding transcript for GPI-anchored proteins. Length is measured relative to the stop codon. (B) Relative tAI profile at C'-terminal segments for 128 human GPI-anchored proteins at the C'-terminal region. (C) Histogram of the cleavage site relative to the initiator Methionine for the TP-proteins. (D) Relative tAI profile of 499 human TP proteins at the N'-terminal region. (E) Relative codon usage in SP- and TP-
proteins. Y-axis scale is the relative codon usage in SP sequences divided by the relative codon usage in the TP sequences. Codons that belong to the same AA are colored as a group. doi:10.1371/journal.pcbi.1003294.g006

proteins) matches ribosomal proteins. Such a profile is expected for proteins that are expressed at high amounts and a translation speed that reaches maximal efficiency (i.e., the number of proteins that are produced per transcript). Ribosomal proteins are known by their high expression, efficient translation and the preferable use of abundant codons. A detailed analysis of proteins clusters

Figure 7. Signal sequence codon usage analysis of the human proteome. (A). Codon usage fold change in SP versus non-SP proteins. The relative codon usage in signal sequences is divided by the relative codon usage in sequences of same length distribution, originated from non-SP proteins. Y-axis is shown as the fold change subtracted by one marking the codons that are more commonly used and those that are underrepresented in the signal sequences. The values of codon tAI are indicated by the empty frame to indicate the absolute tAI value for each codon (as in Figure 1B). (B). tAI distribution of the original signal sequences (blue) and of the signal sequences in which each codon was randomly replaced by a synonymous one according to their codon usage distribution (red). The significance of the mean values of the two distributions is shown.

doi:10.1371/journal.pcbi.1003294.g007
The concept that arises from our study supports the notion of evolutionary dependent marks for a “speed control” management. We have shown that such property is encoded in the initial segment of the SP-proteins (secreted and membranous), TP-proteins (mitochondria targeted), as well as for the terminal segment of GPI-anchored proteins but not the anchor TMD sequences. Thus, the observed segmental tAI profile also acts at the level of “final destination” of proteins. The TP-proteins and the addition of the GPI-moiety [40] are post-translational processes. In the case of TP-proteins, the observed segmental tAI profile (Figure 5B) may act as a “time delayer” to ensure safe folding. Importantly, the observed signal for “speed control” management is missing for the bulk of the proteins that are translated by free ribosomes. It was proposed that the lowly adapted tRNAs at the initial segment of proteins govern the ribosomal allocation properties as expressed by ribosome density and translation speed [41]. In this report, we propose that the evolutionary encoded signal is mainly associated with membrane bound translation. We postulate that it is a general design for complying with the mechanistic and kinetic demands of a restricted subset of the proteome.

Investigating the trend of the local segmental tAI (e.g., Figures 4–5) for protein families allows us to challenge the importance of their profile in view of their function. We focused on 25 human proteins that carry Matrix Metalloproteinses (MMPs) functions [42]. This diverse group consists of membranous (6 proteins) and secreted proteins (18 proteins, Table S5). MMPs contribute to the modulation cancer and metastasis. The different MMPs regulate apoptosis, inflammation, migration, adhesion and vascularization [43]. We noted that the average local tAI profile (Figure S2) of the MMP family resembles the overall N’-terminal segmental trend of the SP-proteome (i.e., initial segment of LATs following by HATS). Interestingly, it is mostly the subset of the membranous MMPs (with/without TMD or with GPI anchor) rather than the secreted MPPs that dominates this pattern. The pattern of the local tAI and the variability in this profile among paralogs and functionally related proteins is under current investigation.

The partition of the complete proteome to four disjointed groups is based on their apparent proteins’ localization. Evidently, other partitions are feasible. We tested the impact of our predetermined partition on the robustness of the observed pattern assigned for the SP-proteome: (i) We confirmed that further partition of the SP-membranous proteins to proteins with a single- or multi-TMDs (Figure 5) had no effect on the observed pattern of the entire group. (ii) The results of an unsupervised clustering procedure showed that a large fraction of the human proteome matches a small number of dominant patterns (Figure 8, Clusters 1–6). Focusing on the clusters that show a pattern similar to that of the SP-proteome revealed a significant enrichment of key terms that include ER lumen, vesicle trafficking, extracellular proteins, receptors, hormones, plasma membrane and such (Table S4). Interestingly, we identified several SP-proteins that belong to small families (e.g., defensins) that exhibit a unique tAI segmental pattern which is different from the dominant secretory clusters (clusters 1–4, Figure 8). Defensins are host-defense secreted peptides of the innate immune system. Defensins resulted from recent duplications and some were shown as specific to the primate lineage [44]. We are currently studying the translational efficiency of such outliers.

A causal relation of the tAI segmental pattern and the apparent translation efficiency is somewhat indirect (discussed in [16]). The estimation of the abundance of tRNAs in vivo (computationally and experimentally) showed the strong correlation to their genomic copy number [26] under a broad set of conditions. However, subtle effects of tRNA concentration at the ribosome A-site, the activity and extent of the RNA modifying enzymes [45] and the

Figure 8. Clustering of all human proteome according to their segmental tAI values for the N’-terminal. A total of 18,434 proteins are included in the analysis and clustered by the calculated tAI for 5 consecutive overlapping segments at the N’-terminal region of the proteins. Unsupervised clustering resulted in several dominating clusters that are numbered 1–6. Red and green colors mark the low and high segmental tAI values, respectively (according to the scale). For details on annotation enrichment for each cluster, see Table S4.

doi:10.1371/journal.pcbi.1003294.g008
actual fraction of the loaded/unloaded tRNAs adds to the dynamic modeling of ribosome allocation and queuing [46].

A quantitative view of the need for allocating the resources for translation was proposed based on experimental [2] and evolution considerations [3,14]. While most of the analysis is based on E. coli and S. cerevisiae [3], the impact of the different determinants on in vivo translation efficiency in humans and other multicellular organisms remained an open issue [47]. The observed pattern of conserved optimal and non-optimal codons in clusters was proposed as an evolutionary evolved rhythm for the ribosomal speed in accordance with the secondary structure of the translated polypeptides [48].

Additional hardcoded signals are encoded by the CG content, the Shine-Delgarno (SD) and the Kozak sequences around the terminal segments, there was no outstanding signal in any of these measurements, the “hydrophobicity” per se cannot account for our findings: (i) The hydrophobic AAs are not particularly associated with low tAI values (Figure 1B, Table S1). (ii) The C’-terminal helical segment of the GPI-precursor lies in between the secreted SP and TMD segments in terms of hydrophobicity [53]. (iii) Despite a poor correlation of tAI values among organisms (excluding yeast and humans), the statistical confidence at the codon resolution (Figures 4–6) failed to show the pattern of C’-terminal LATS, despite the prominent presence of a TMD in the C’-terminal segment. (iv) The TMD from the group of “TMD not SP” showed that the hydrophobicity cannot account for low adapted codons (Figure S1).

In accordance with our view, the evolution rate for SP sequences was calculated to be 10 fold higher when compared to the mature proteins. Specifically, it was suggested that SP sequences have undergone positive selection [54]. We argue that the variability in the SP sequences is a reflection of the translation “hard-coded” speed control signals that covers these segments. Additional sequence determinants for translation efficiency include the GC content, transcript and coding length, over-representation of correlated codons [55], and the tendency for mRNA secondary structures. We showed that the GC and the coding length do not constitute the basis for our reported observations.

From an evolutionary perspective, it was proposed that an optimal strategy in enhancing translational efficiency is observed under tRNA shortage [18]. However, in addition to purely sequence-based determinants, a number of context-dependent attributes (often hard to separate) govern the translational speed [56]. This includes the presentation of secondary structures, the accessibility of ribosomes and masking of the transcript by RNA binding proteins [57]. Isolating these determinants is context-dependent and naturally also cell specific (e.g., some cells may contain RNA binding proteins that interfere with the ribosome flow). Whether the tAI segmental profile directly governs the speed parameters for multi-cellular eukaryotes is yet to be tested.

Sophisticated imaging technologies determined the parameters of the translation elongation rate at a codon resolution [58]. In addition, in vivo experimental measures by ribosomal profiling [2,13] provided detailed data on the steady state of the ribosome positioning during translation. Our current analyses provide an additional layer to the qualitative outlook of the process of elongation [59].

Mechanistic constraints for ER bound translation

Several models were developed to capture the translation kinetics of the secretory proteome [60–62]. Based on this view, the signal that was exposed in this report could also serve to enhance the capacity of the mRNA to engage in a productive ER targeting process. An efficient reuse of the mRNA on MEM-Rb, once the mRNA is “occupied” by an already docked ribosome, is an attractive proposal [13,63].

Our analysis focused on the MEM-Rb translation. We revisited the mechanistic demands of the secretory proteome [30]. In addition to the need of managing the ribosomal flow for any transcript, special constrains are imposed for the MEM-Rb translation. In mammals, the co-translocation of SP-containing proteins is mediated mostly by the signal recognition particle (SRP) [64]. Once the SRP recognizes the emerging SP from the ribosome [65], a conformational change leads to slowing of translation. Apparently, this attenuation in translation rate is necessary for the nascent chain to diffuse to the ER membrane [47]. The interaction of the SRP with its receptor (SR) and its release serve as an internal “timer” for resuming translation [66], and for production of functional proteins [67].

Recently, the SRP-independent insertion route was systematically assessed in yeast [68] and mammals [69]. The dependency of the hydrophobicity index of the N’-terminal segments of the proteins and the tendency to bind the SRP revealed that a substantial fraction of the yeast secretome is actually SRP-independent and this fraction mainly applies to SP-proteins and
to the subset of the GPI proteome [68]. Thus, the notion of a "timer" for translation and translocation may not be limited to SRPs but to the need for a rich network of proteins and chaperones that coordinate their actions to ensure appropriate translocation and targeting.

A role for the codons' distribution along the transcripts as a “time delay” should be considered. With this notion, the generality for transcripts for SP-, TP- and GPI-anchor proteins is striking. We suggest that attenuation of events such as the SP proteolytic cleavage (not necessarily in the end of the LATS), the speed of folding, the cleavage of GPI to promote the locking of the protein at the membrane surface, and recycling of the mRNA to ensure additional rounds of translation are all encoded in the codon organization profile. A similar signature across a range of organisms from yeast to humans indicates a robust, evolutionary refined phenomenon.

Materials and Methods

Proteins’ coding sequences and experimental data

The list of proteins for each group of each organism was taken from UniProtKB based on a “reviewed” set. For SP proteins we used the UniProtKB (Based on SignalP4.0 [52]). Only proteins marked with “signal” and “cleaved site” were considered. The SP-anchored proteins were excluded from the SP-proteins group. In addition, the proteins marked as “fragment” were excluded. A similar protocol was applied for GPI-anchored and TP (transit-peptide) and predicted Tail-anchored (TA) Type IV. The canonical variants from UniProtKB were mapped to their matched RefSeq nucleotide sequences. A gene that had no matched sequence, or had a sequence that lacked the ATG initiator codon, was discarded. The corresponding coding sequences were extracted from the RefSeq database. Only proteins that start with an initiator Methionine and end with a stop codon are compiled.

Signal peptide sequences were retrieved from the proteins coding sequences according to their position that were marked by UniProtKB. The codon usage for these sequences was counted and defined as SP codon usage. The codon usage of sequence from proteins that are not annotated as SP proteins was counted as non-SP codon usage. Those sequences began at the first position of the coding sequence and terminated at a position that was randomly selected from the signal sequence length distribution. Sequences that were randomly replaced were created by replacing each codon in the sequence with a codon from its synonymous codons by a random choice according to the codon usage of each AA. Randomized tests were performed 1000 times.

A high coverage (>70%, 4,500 proteins) mass spectrometry (MS) yeast experiment [33] was used for protein abundance measurements. Protein levels span more than four orders of magnitude. Independent yeast protein quantitation was extracted from the GFP library measurements [34]. Briefly, each protein from the GFP-tagged yeast library was counted by flow cytometry measurement (~2,500 proteins). For human protein abundance, the MS data resource for the high-coverage of 11 human cell-lines [35] was used.

tAI measurements

An estimation of the effect of the tRNA abundance on the efficiency of the translation rate of codons is captured by the tRNA adaptation index (tAI) [19]. Global tAI measurement gauges the availability of tRNAs for each codon along the mRNA. Data of genomic tRNA copy numbers were taken from the Genomic tRNA Database (http://grtndb.ucsc.edu/) using human genome hg19 (NCBI Build 37.1, Feb 2009) [70]. For each tRNA isoacceptor, the number of gene copies (excluding Pseudogenes and Selanocysteine tRNAs) was counted. The codon tAI and global tAI for the model organisms was calculated as above from Genomic tRNA Database (Table S1).

A codon-anticodon coupling is not unique - a factorization for each of the wobble pair was used [19]. Formally, let ni be the number of tRNA isoacceptors recognizing codon i. Let tCGNij be the copy number of the jth tRNA that recognizes the ith codon, and let Sij be the selective constraint on the efficiency of the codon-anticodon coupling. We have used the Sij scaling for the Wobble nucleoside-nucleoside pairing as described in [41]. We define the absolute adaptiveness, Wi, for each codon i as:

From Wi we obtain wi, which is the relative adaptiveness value of codon i, by normalizing the Wi’s values (dividing them by the maximal of all the 61 Wi).

The final tAI of a gene (referred as Global tAI) is the geometric mean of its codons (excluding the stop codon). A geometric mean was calculated in an identical way for calculating the segmental tAI (e.g., 30-codons, SP-segment, TMD segment). Local tAI is calculated by dividing each coding sequence into several overlapping windows, each containing 30 codons. Relative tAI value is defined as the ratio of the segmental, local tAI (i.e., 30-codons segment) to the calculated global tAI of the protein (for the entire protein length). A relative tAI value <1.0 signifies the preference of rarely adapted tRNA codons (“slow” codons) in the analyzed segment relative to the codon composition of the entire coding sequence. Global tAI and C1 segment tAI were computed by excluding the stop codon from their sequences. For sequences that are shorter than 180 amino acids, only local segmental tAI were calculated. This was applied to avoid overlap between N’ and C’ terminal windows.

Proteins’ clustering

Protein clustering was performed for a matrix of 18,434 rows (each represents a mRNA-mapped coding sequence), and five columns (each represents a window of 30 codons from the N’-terminus segments marked N1 to N3. The functional annotation enrichment of the resulted clusters was according to Fisher Exact Test enrichment scheme with hypergeometric distribution and multiple hypothesis corrections [71].

Statistical analysis and simulations

Different data distributions were compared using the standard Matlab statistical tools such as Kolmogorov–Smirnov (KS) and t-tests. The KS test compared any two samples while quantifying the empirical cumulative distribution functions of the two. The p-value is calculated under the null hypothesis that the samples are drawn from the same distribution. Thus, the lower p values indicate more significant differences between the two examined samples. The difference in the probability distribution between the two datasets was computed using Kullback-Leibler divergence (dKL) (see detailed in [72]). For testing the similarity of the segmental tAI profile to randomly created genes, we created random gene sets with the same codon preference and same length distribution. We selected a set of 1000 genes. The simulation was performed by 1000 repetitions of the protocol.
Supporting Information

Figure S1  Reanalysis of the local tAI of the human “TMD not SP” proteins, according to their first TMD that serves as anchor signal. (PDF)

Figure S2  Local tAI of 25 human Matrix metalloproteinases (PDF)

Table S1  The tAI codons values for 6 model organisms. Each of the 61 codons are indicated by the calculated tAI. For each tRNA isoacceptor, the number of gene copies (excluding Pseudogenes and tRNA for Selanocysteine) was counted. (DOCX)

Table S2  Partition of the complete proteomes to 4 groups. (DOCX)

Table S3  Global tAI values for complete proteomes partitioned to 4 groups for 6 eukaryotic organisms. (DOCX)

Table S4  Annotation enrichment summary for clusters 1–4, Figure 7. (DOCX)

Table S5  Identifier of 25 proteins of the human Matrix metalloproteinases, input list of Figure S2. (PDF)

Acknowledgments

We thank Nathan Linial for his advice and suggestions throughout the project. We thank Tamir Tuller and Manor Askenazi for useful discussions.

Author Contributions

Conceived and designed the experiments: SM ML. Performed the experiments: SM ML. Analyzed the data: SM ML. Contributed reagents/materials/analysis tools: SM ML. Wrote the paper: SM ML.

References


4.2.1 Appendix

Figure S1: Reanalysis of the local tAI of the human “TMD not SP” proteins, according to their first TMD that serves as anchor signal.
Figure S2: Local tAI of 25 human Matrix metalloproteinases

Table S1. Codons tAI for H. sapiens, S. cerevisiae, C. elegans, D. melanogaster, B. taurus and A. thaliana.

<table>
<thead>
<tr>
<th>codon</th>
<th>H.sapiens</th>
<th>B.taurus</th>
<th>D. melanogaster</th>
<th>C.elegans</th>
<th>S.cerevisiae</th>
<th>A. thaliana</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>0.547238</td>
<td>0.061532</td>
<td>0.286807</td>
<td>0.416667</td>
<td>0.431034</td>
<td>0.173333</td>
</tr>
<tr>
<td>AAG</td>
<td>0.779954</td>
<td>0.050456</td>
<td>0.713193</td>
<td>0.994444</td>
<td>1</td>
<td>0.282133</td>
</tr>
<tr>
<td>AAT</td>
<td>0.458698</td>
<td>0.013853</td>
<td>0.251816</td>
<td>0.243889</td>
<td>0.297352</td>
<td>0.093653</td>
</tr>
<tr>
<td>ACA</td>
<td>0.17284</td>
<td>0.00789</td>
<td>0.286855</td>
<td>0.250047</td>
<td>0.30795</td>
<td>0.10668</td>
</tr>
<tr>
<td>ACC</td>
<td>0.207373</td>
<td>0.008741</td>
<td>0.341468</td>
<td>0.34</td>
<td>0.487685</td>
<td>0.096</td>
</tr>
<tr>
<td>ACG</td>
<td>0.228111</td>
<td>0.007258</td>
<td>0.235182</td>
<td>0.302222</td>
<td>0.160099</td>
<td>0.1008</td>
</tr>
<tr>
<td>ACT</td>
<td>0.288018</td>
<td>0.01139</td>
<td>0.478011</td>
<td>0.472222</td>
<td>0.67734</td>
<td>0.133333</td>
</tr>
<tr>
<td>AGA</td>
<td>0.172814</td>
<td>0.032344</td>
<td>0.143403</td>
<td>0.222222</td>
<td>0.738916</td>
<td>0.12</td>
</tr>
<tr>
<td>AGC</td>
<td>0.251152</td>
<td>0.015683</td>
<td>0.286807</td>
<td>0.25</td>
<td>0.123153</td>
<td>0.173333</td>
</tr>
<tr>
<td>AGG</td>
<td>0.199309</td>
<td>0.08687</td>
<td>0.189293</td>
<td>0.21</td>
<td>0.29803</td>
<td>0.145067</td>
</tr>
<tr>
<td>AGT</td>
<td>0.129954</td>
<td>0.009043</td>
<td>0.125908</td>
<td>0.10975</td>
<td>0.054064</td>
<td>0.076093</td>
</tr>
<tr>
<td>ATA</td>
<td>0.144052</td>
<td>0.005523</td>
<td>0.09565</td>
<td>0.194506</td>
<td>0.123233</td>
<td>0.06692</td>
</tr>
<tr>
<td>ATC</td>
<td>0.541475</td>
<td>0.009656</td>
<td>0.344168</td>
<td>0.44</td>
<td>0.637931</td>
<td>0.1824</td>
</tr>
<tr>
<td>ATG</td>
<td>0.576037</td>
<td>0.026033</td>
<td>0.573614</td>
<td>0.527778</td>
<td>0.738916</td>
<td>0.32</td>
</tr>
<tr>
<td>ATT</td>
<td>0.53318</td>
<td>0.013411</td>
<td>0.478011</td>
<td>0.611111</td>
<td>0.827525</td>
<td>0.253333</td>
</tr>
<tr>
<td>CAA</td>
<td>0.201613</td>
<td>0.0142</td>
<td>0.191205</td>
<td>0.583342</td>
<td>0.554187</td>
<td>0.106667</td>
</tr>
<tr>
<td>CAC</td>
<td>0.288018</td>
<td>0.019059</td>
<td>0.239006</td>
<td>0.587778</td>
<td>0.492611</td>
<td>0.133333</td>
</tr>
<tr>
<td>CAG</td>
<td>0.611751</td>
<td>0.032943</td>
<td>0.443595</td>
<td>0.381111</td>
<td>0.238916</td>
<td>0.154133</td>
</tr>
<tr>
<td>CAT</td>
<td>0.12644</td>
<td>0.009985</td>
<td>0.104924</td>
<td>0.315028</td>
<td>0.216256</td>
<td>0.058533</td>
</tr>
<tr>
<td>CCA</td>
<td>0.230446</td>
<td>0.007101</td>
<td>0.239039</td>
<td>0.972239</td>
<td>0.615776</td>
<td>0.600023</td>
</tr>
<tr>
<td>CCC</td>
<td>0.228111</td>
<td>0.006816</td>
<td>0.240918</td>
<td>0.12</td>
<td>0.08867</td>
<td>0.1632</td>
</tr>
<tr>
<td>CCG</td>
<td>0.18894</td>
<td>0.006216</td>
<td>0.315488</td>
<td>0.422222</td>
<td>0.197044</td>
<td>0.258667</td>
</tr>
<tr>
<td>CCT</td>
<td>0.31682</td>
<td>0.009466</td>
<td>0.334608</td>
<td>0.166667</td>
<td>0.123153</td>
<td>0.226667</td>
</tr>
<tr>
<td>CGA</td>
<td>0.172831</td>
<td>0.009468</td>
<td>0.478059</td>
<td>0.305608</td>
<td>4.31E-05</td>
<td>0.080012</td>
</tr>
<tr>
<td>CGC</td>
<td>0.145161</td>
<td>0.015052</td>
<td>0.344168</td>
<td>0.38</td>
<td>0.310345</td>
<td>0.0864</td>
</tr>
<tr>
<td>CGG</td>
<td>0.199309</td>
<td>0.013285</td>
<td>0.152964</td>
<td>0.153333</td>
<td>0.061576</td>
<td>0.092267</td>
</tr>
<tr>
<td>CGT</td>
<td>0.201613</td>
<td>0.014161</td>
<td>0.478011</td>
<td>0.527778</td>
<td>0.431034</td>
<td>0.12</td>
</tr>
<tr>
<td>CTA</td>
<td>0.086437</td>
<td>0.003945</td>
<td>0.095626</td>
<td>0.083383</td>
<td>0.184729</td>
<td>0.133349</td>
</tr>
<tr>
<td>CTC</td>
<td>0.228111</td>
<td>0.00568</td>
<td>0.172084</td>
<td>0.36</td>
<td>0.061576</td>
<td>0.128533</td>
</tr>
<tr>
<td>CTG</td>
<td>0.286866</td>
<td>0.008362</td>
<td>0.413002</td>
<td>0.165556</td>
<td>0.059113</td>
<td>0.082667</td>
</tr>
<tr>
<td>CTT</td>
<td>0.31682</td>
<td>0.007889</td>
<td>0.239006</td>
<td>0.5</td>
<td>0.027032</td>
<td>0.165853</td>
</tr>
<tr>
<td>GAA</td>
<td>0.374424</td>
<td>0.32738</td>
<td>0.286807</td>
<td>0.472222</td>
<td>0.923645</td>
<td>0.16</td>
</tr>
<tr>
<td>GAC</td>
<td>0.460829</td>
<td>0.039696</td>
<td>0.669216</td>
<td>0.75</td>
<td>0.923645</td>
<td>0.346667</td>
</tr>
<tr>
<td>GAG</td>
<td>0.379032</td>
<td>0.220725</td>
<td>1</td>
<td>0.817778</td>
<td>0.418719</td>
<td>0.224533</td>
</tr>
<tr>
<td>Codon</td>
<td>Frequency1</td>
<td>Frequency2</td>
<td>Frequency3</td>
<td>Frequency4</td>
<td>Frequency5</td>
<td>Frequency6</td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>GAT</td>
<td>0.202304</td>
<td>0.020664</td>
<td>0.293786</td>
<td>0.32925</td>
<td>0.40548</td>
<td>0.152187</td>
</tr>
<tr>
<td>GCA</td>
<td>0.288102</td>
<td>0.022091</td>
<td>0.09566</td>
<td>0.250061</td>
<td>0.30795</td>
<td>0.133355</td>
</tr>
<tr>
<td>GCC</td>
<td>0.601382</td>
<td>0.018617</td>
<td>0.413002</td>
<td>0.44</td>
<td>0.487685</td>
<td>0.1536</td>
</tr>
<tr>
<td>GCG</td>
<td>0.236175</td>
<td>0.01969</td>
<td>0.173996</td>
<td>0.191111</td>
<td>0.098522</td>
<td>0.136</td>
</tr>
<tr>
<td>GCT</td>
<td>0.835253</td>
<td>0.024359</td>
<td>0.573614</td>
<td>0.611111</td>
<td>0.67734</td>
<td>0.213333</td>
</tr>
<tr>
<td>GGA</td>
<td>0.259217</td>
<td>0.277683</td>
<td>0.286807</td>
<td>1</td>
<td>0.246305</td>
<td>0.16</td>
</tr>
<tr>
<td>GGC</td>
<td>0.432028</td>
<td>0.053832</td>
<td>0.669216</td>
<td>0.416667</td>
<td>0.985222</td>
<td>0.306667</td>
</tr>
<tr>
<td>GGG</td>
<td>0.342166</td>
<td>0.017357</td>
<td>0.095631</td>
<td>0.138942</td>
<td>0.184815</td>
<td>0.093353</td>
</tr>
<tr>
<td>GGT</td>
<td>0.18966</td>
<td>0.032804</td>
<td>0.293786</td>
<td>0.182917</td>
<td>0.432512</td>
<td>0.134627</td>
</tr>
<tr>
<td>GTA</td>
<td>0.144041</td>
<td>0.017357</td>
<td>0.095631</td>
<td>0.138942</td>
<td>0.184815</td>
<td>0.093353</td>
</tr>
<tr>
<td>GTC</td>
<td>0.228111</td>
<td>0.013064</td>
<td>0.206501</td>
<td>0.38</td>
<td>0.62069</td>
<td>0.144</td>
</tr>
<tr>
<td>GTG</td>
<td>0.535714</td>
<td>0.039475</td>
<td>0.365201</td>
<td>0.211111</td>
<td>0.182266</td>
<td>0.136533</td>
</tr>
<tr>
<td>GCT</td>
<td>0.31682</td>
<td>0.018144</td>
<td>0.286807</td>
<td>0.527778</td>
<td>0.862069</td>
<td>0.2</td>
</tr>
<tr>
<td>TAA</td>
<td>0.028805</td>
<td>0.003156</td>
<td>0.262268</td>
<td>0.30313</td>
<td>0.061576</td>
<td>0.157197</td>
</tr>
<tr>
<td>TAC</td>
<td>0.423963</td>
<td>0.034079</td>
<td>0.43021</td>
<td>0.527778</td>
<td>0.492611</td>
<td>1</td>
</tr>
<tr>
<td>TAG</td>
<td>0.009217</td>
<td>0.009687</td>
<td>0.262268</td>
<td>0.027778</td>
<td>0.019704</td>
<td>0.157197</td>
</tr>
<tr>
<td>TAT</td>
<td>0.205818</td>
<td>0.020356</td>
<td>0.188862</td>
<td>0.231694</td>
<td>0.216256</td>
<td>0.439</td>
</tr>
<tr>
<td>TCA</td>
<td>0.115236</td>
<td>0.003945</td>
<td>0.095641</td>
<td>0.250042</td>
<td>0.246373</td>
<td>0.120049</td>
</tr>
<tr>
<td>TCC</td>
<td>0.207373</td>
<td>0.031618</td>
<td>0.275335</td>
<td>0.3</td>
<td>0.487685</td>
<td>0.368533</td>
</tr>
<tr>
<td>TCG</td>
<td>0.152074</td>
<td>0.006784</td>
<td>0.221797</td>
<td>0.357778</td>
<td>0.140394</td>
<td>0.091733</td>
</tr>
<tr>
<td>TCT</td>
<td>0.288018</td>
<td>0.021434</td>
<td>0.382409</td>
<td>0.416667</td>
<td>0.67734</td>
<td>0.499187</td>
</tr>
<tr>
<td>TGA</td>
<td>0.086406</td>
<td>0.019734</td>
<td>0.047801</td>
<td>0.027778</td>
<td>0.061576</td>
<td>0.157197</td>
</tr>
<tr>
<td>TGC</td>
<td>0.892857</td>
<td>0.356473</td>
<td>0.334608</td>
<td>0.361111</td>
<td>0.246305</td>
<td>0.2</td>
</tr>
<tr>
<td>TGG</td>
<td>0.229263</td>
<td>0.130952</td>
<td>0.397706</td>
<td>0.342222</td>
<td>0.389163</td>
<td>0.186667</td>
</tr>
<tr>
<td>TGT</td>
<td>0.391964</td>
<td>0.239578</td>
<td>0.146893</td>
<td>0.158528</td>
<td>0.108128</td>
<td>0.0878</td>
</tr>
<tr>
<td>TTA</td>
<td>0.144009</td>
<td>0.006311</td>
<td>0.191205</td>
<td>0.138889</td>
<td>0.492611</td>
<td>0.08</td>
</tr>
<tr>
<td>TTC</td>
<td>0.374424</td>
<td>0.024928</td>
<td>0.382409</td>
<td>0.416667</td>
<td>0.67734</td>
<td>0.213333</td>
</tr>
<tr>
<td>TTG</td>
<td>0.247696</td>
<td>0.013853</td>
<td>0.25239</td>
<td>0.238889</td>
<td>0.773399</td>
<td>0.158933</td>
</tr>
<tr>
<td>TTT</td>
<td>0.164372</td>
<td>0.013641</td>
<td>0.167878</td>
<td>0.182917</td>
<td>0.297352</td>
<td>0.093653</td>
</tr>
</tbody>
</table>
Table S2. Number of proteins according to 4 distinct groups partition in six different organisms.

<table>
<thead>
<tr>
<th>Organism</th>
<th>SP non-TMD</th>
<th>TMD non-SP</th>
<th>non-SP non-TMD</th>
<th>SP and TMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. sapiens</td>
<td>1758</td>
<td>3395</td>
<td>11887</td>
<td>1394</td>
</tr>
<tr>
<td>B. taurus</td>
<td>636</td>
<td>926</td>
<td>3655</td>
<td>265</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>302</td>
<td>475</td>
<td>2198</td>
<td>119</td>
</tr>
<tr>
<td>C. elegans</td>
<td>296</td>
<td>542</td>
<td>2243</td>
<td>106</td>
</tr>
<tr>
<td>A. thaliana</td>
<td>1258</td>
<td>1868</td>
<td>7084</td>
<td>423</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>215</td>
<td>1074</td>
<td>4658</td>
<td>93</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>H. sapiens</th>
<th>B. taurus</th>
<th>D. melanogaster</th>
<th>C. elegans</th>
<th>A. thaliana</th>
<th>S. cerevisiae</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP no-TMD</td>
<td>0.3246</td>
<td>0.0268</td>
<td>0.3352</td>
<td>0.4101</td>
<td>0.1705</td>
<td>0.4206</td>
</tr>
<tr>
<td>TMD no-SP</td>
<td>0.3208</td>
<td>0.0244</td>
<td>0.3192</td>
<td>0.3646</td>
<td>0.1686</td>
<td>0.3696</td>
</tr>
<tr>
<td>SP and TMD</td>
<td>0.3201</td>
<td>0.0253</td>
<td>0.3202</td>
<td>0.3703</td>
<td>0.1666</td>
<td>0.3663</td>
</tr>
<tr>
<td>Cytosolic</td>
<td>0.3224</td>
<td>0.0270</td>
<td>0.3415</td>
<td>0.4004</td>
<td>0.1701</td>
<td>0.3883</td>
</tr>
</tbody>
</table>

Table S3. Global tAl values for the proteins according to 4 distinct groups in 6 different organisms.
Table S4. Functional enrichments resulted for the hierarchical clustering of the human proteome. The clustering is based on the vector of 5 consecutive tAI segments values (covers 90 codons from the N'-terminal region) and 18,434 protein sequences.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Number of proteins</th>
<th>Dominating InterPro and GO terms</th>
<th>C-EASE score&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Most significant p-value</th>
<th>Most significant FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,3</td>
<td>2160</td>
<td>Glycoprotein, signal peptide, disulfide bond, Secreted, extracellular region</td>
<td>9.995</td>
<td>5.38e-18</td>
<td>8.14e-15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cell membrane, plasma membrane, transmembrane region, intrinsic to membrane</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2160</td>
<td>Oxidoreductase, electron carrier activity, oxidation reduction</td>
<td>7.731</td>
<td>1.15e-05</td>
<td>13.10</td>
</tr>
<tr>
<td>1</td>
<td>1527</td>
<td>Extracellular matrix, extracellular region</td>
<td>3.605</td>
<td>5.66e-04</td>
<td>0.93</td>
</tr>
<tr>
<td>2</td>
<td>1162</td>
<td>Receptor, signal peptide, g-protein coupled receptor, Olfactory receptor, G protein-coupled olfactory receptor, class II, sensory transduction</td>
<td>4.078</td>
<td>6.81e-5</td>
<td>0.10</td>
</tr>
<tr>
<td>2</td>
<td>1629</td>
<td>Transit peptide, mitochondrion</td>
<td>2.904</td>
<td>1.11e-6</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1162</td>
<td>Gap junction, cysteine-rich domain, Connexins, wide pore channel activity, cell-cell junction</td>
<td>2.454</td>
<td>1.50e-5</td>
<td>0.028</td>
</tr>
<tr>
<td>4</td>
<td>1154</td>
<td>Connexins, wide pore channel activity, cell-cell junction</td>
<td>5.034</td>
<td>3.76e-7</td>
<td>6.29e-4</td>
</tr>
</tbody>
</table>

<sup>a</sup>C-EASE score is roughly the average enrichment for all keywords in the cluster transformed to -log10 units.
Chapter 5

Discussion

In this thesis I presented my contribution to the field of gene expression regulation, which dictates the behavior of any multi-cell organisms and underlying the ability of cells to adapt and response to stress and other changing condition. Since gene expression profile has a direct impact on cell phenotype, and therefore defines cell state and identity, its regulation involves multiple layers. In my studies I focused on post-transcriptional regulation using computational modeling that benefit from the accumulated experimental knowledge. Specifically, I researched the impact of miRNA regulation in two cellular dimensions: First, the analyzing their overlooked presence associated with the supraspliceosome complex (SF) within the nucleus, and second, modeling their interaction network and the regulatory impact on the gene expression profile of the cell.

At an additional layer of post transcriptional regulation, I focused on translation. I further investigated the regulation of the translation machinery in view of the encoded efficiency of the translation elongation process.

5.1. Computational model of miRNA regulation

The field of miRNA regulation has been extensively studied in the last two decades (He and Hannon, 2004; Chen and Rajewsky, 2007; Bartel, 2018). Nowadays, about 60% of the human coding genes are considered to be regulated by miRNAs in a cellular context (He and Hannon, 2004; Jonas and Izaurralde, 2015). To understand the impact of this regulation, experimental techniques as well as computational models were developed. In fact, most of our current knowledge on the specificity of miRNA-mRNA regulatory network is based on computational prediction tools (Peterson et al., 2014). Most of these predictors are made according to the results from experimental data, free energy calculation, and a collection of sequence-based features. At first, the validation of miRNA-mRNA interaction was done mainly by a specific miRNA manipulation following the
analysis of changes of the expression levels of different genes (Hausser and Zavolan, 2014). As the technology advances, new large-scale methodologies were developed. One of such important methodology in miRNA investigation is the CLIP-Seq (cross-linking immunoprecipitation) method that combines UV cross-link of RNA and proteins with immunoprecipitation of Argonaute (AGO) protein, and further analyze by deep sequence analysis the attach RNA molecules (Licatalosi et al., 2008; Thomson, Bracken and Goodall, 2011). Another important methodology that had been developed recently is CLASH and CLEAR-(covalent ligation of endogenous Argonaute-bound RNAs)-CLIP, that generates miRNA–mRNA chimeras (Helwak et al., 2013) by adding a ligation step, thus report on pairs of miRNA and their binding segments from mRNAs. This methodology exposed non-canonical binding sites and a large number of coding region interactions. Those protocols reveal many novel miRNAs (Kozomara and Griffiths-Jones, 2014). Unfortunately, they suffer from low coverage and poor consistency (Chi, Hannon and Darnell, 2012). Therefore there are no clear evidences of the miRNA-mRNA interaction network (Betel et al., 2010; Helwak et al., 2013; Seok et al., 2016).

Since the regulatory network of miRNA–mRNA interaction has a many-to-many mode, any change in one miRNA may result in direct effects on its target genes, but also non-directional effects on the non-target genes. Modeling the impact of such interaction network on the final gene expression profile is non-trivial. As described above, gene expression has many levels of regulation that involve different factors and regulators. Different aspects of this interaction had been modeled (Schulz et al., 2013; Nitzan et al., 2014), however the quantitative aspect at a cell level is mostly ignored. In my work, I developed a quantitative stochastic model that considers the expression profile of both miRNAs and mRNAs, and the refined knowledge of their probability of interactions and the potential effect once an interaction is materialized. This computational model enables us to simulate the interaction network, while exploring a large set of parameters.

5.2. Computational view of translation efficiency

There are many aspects that regulates the translation process (Pisarev, Hellen and Pestova, 2007; Sonenberg and Hinnebusch, 2009). In recent studies the notion of slow translated codons at the beginning of the coding sequence, followed by effective translated
codons was found, and was the basis for the translation ‘ramp’ model (Gingold and Pilpel, 2011). In my study of translation regulation, I focused on the sequence-based feature of gene codon usage in view of its impact on translation elongation rate. Specifically, I examined the coding region elongation rate according to the tRNA adaptation index (tAI). This measurement considers the availability of the corresponding tRNA of each codon, and was found to be a good approximation of the elongation efficiency (Reis, Savva and Wernisch, 2004; Tuller et al., 2010).

5.3. Gene sets display different regulation sensitivity

As the many levels of gene expression regulation overlap in time and space, we found that each gene may have different sensitivity at each regulation level. Using the miRNA-mRNA interaction model described in chapter 3, I performed thousands of simulations, each represent a case of overexpression of a specific miRNA, in three different cells (HeLa, HEK293 and MCF-7) and covering 250 different miRNAs. Each of the manipulated miRNA was over-expressed by 7 increasing degrees of overexpression, from the default level (marked as x1) to x1000 folds higher expression. The analysis of such exhaustive simulation revealed a general behavior of sensitivity of the different genes to miRNA manipulation. I found that there is a relatively large set of genes (about 20% of the reported genes) that are exceptionally stable. Those genes had final retention of at least 85% for almost all miRNA manipulations. Although this set of genes was identified independently for each of the three tested cell types, I found a statistically significant overlapping set of genes. This suggest that the identity of those genes reflects a strong evolutionary signal of stability to any possible miRNA changes. The annotations of the miRNA stable genes are enriched by different aspects of translation including the structure of the ribosomes. On the other hand, I found an interesting small set of sensitive genes, those are genes that their final retention is at most 50%, under almost all miRNA manipulations that were performed. However, this miRNA–sensitive set of genes was unique for each of the tested cell type. Another interesting dimension resulted in an unexpected finding is the observation that different genes have different dynamic along the simulation of miRNA acting in the cell, as some genes tend to degrade fast and other mainly degrade towards the end of the
simulation. This finding corroborates the claim that within the cells, a competition occurs that is sensitive to dynamics of changing the amounts and ratio of the molecular entities.

In view of translation elongation, unsupervised clustering of the segmental tAI of the coding sequence revealed that the signal of “speed control” suggested by the ‘ramp’ model cannot be globally generalized and it is valid for the set of genes that have signal peptide (SP) at the beginning of their coding sequence. The set of SP-gene in human includes about 3100 proteins. This pattern of translation efficiency of the SP-proteome seems to be evolutionary conserved, as it was also found in *S. cerevisiae, C. elegans, D. melanogaster* and *B. taurus*. Indeed, the evolution rate for SP sequences was calculated to be 10-fold higher when compared to the mature proteins. Specifically, it was suggested that SP sequences have undergone positive selection (Li *et al.*, 2009).

5.4. Regulation in localizations

In living cells, many processes are localized and thus dictate a local concentration and molecular availability. In my work, I analyzed the effects of the miRNA–mRNA interaction network within the cytoplasm, where the most common and canonical interactions occur and where the mature mRNAs are found. However, sporadic findings claim that miRNAs function in the nucleus, presumably for non-classical function (Roberts, 2014). In chapter 2, I describe my findings indicating the abundance of miRNA within the HeLa SF, the large splicing machinery that is located at the nucleus. We analyzed small RNA that were associated with the SF, and found the reads were aligned to 207 different miRNAs (along with a tail of low expressing miRNAs). We found differences in the miRNA expression profile between the SF and the cytoplasmic one. As the miRNA repertoire in the cytoplasm is much wider than the one found in the SF, there are many miRNAs that were not found there. Interestingly, I found pre-miRNA sequences that are exclusive to the SF such hsa-mir-663a, hsa-mir-6087, hsa-mir-7704. A natural assumption would be that the miRNAs that the reads were aligned to within the SF would be miRNAs that are encoded within a host gene, and indeed the majority of the miRNA that were found are associated with host genes, as is the case in miRNA acting in the cytosol. However, I found that many of the sequences derived from independent miRNA, i.e., miRNA that are not encoded within a hosting intron. As the final stages of canonical miRNA biogenesis
occurs in the cytoplasm, we would expect to find either complete pre-miRNA sequences
are sequences that would randomly align to any sequence along the pre-miRNA. I found
that 64% of the pre-miRNA fragments were mature miRNA, that implies their reshuffling
from the nucleus back to the cytoplasm. The finding of many reads that are aligned to
various regions of the pre-miRNA suggest a local nuclear processing through yet unknown
mechanism. The finding of the different pre-miRNA derived sequences in the SF is a non-
canonical functionality within the nucleus. We found that some the sequences that were
found in the SF are complementary to other gene sequences. One example is hsa-mir-7704.
This miRNA overlaps with the 5’ end of the HOXD1 gene, and also displays a perfect
complementary to the 5’ UTR of the HAGLR gene. Therefore, it can play a role in
coordinating their expression patterns. Our results show an inverse correlation between the
expression of HAGLR and miR-7704 in two different cell lines, as expected from a direct
competition on transcription. It is likely that whereas HAGLR acts as a negative regulator
of HOXD1, miR-7704 helps enhancing HOXD1 transcription by neutralizing the
expression of its antisense repressor.

In view of translation regulation, as described in chapter 4 we found that the localization
of the translation has a great impact on its dynamic. We found that proteins with signal
peptide, that are translated on the ribosomes that are bounded to the ER membrane, proteins
that have transit peptide and are targeted to the mitochondria, and proteins with terminated
with Glycosylphosphatidyl inositol (GPI)-anchored proteins that are also translated by the
ER ribosomes have a pattern of slowly translated codons at the sequence that is responsible
to their localization. In all these instances, an enzymatic step is needed for the protein
maturation and thus a delay in translation is encoded by the selection of low codons as
reflected by the segmental tAI score. Their translation elongation efficiency pattern meets
the notion of the ‘ramp’ model, suggesting that a slow translated segment may act as a time
delayer to ensure safe folding. Importantly, we did not observe the ‘ramp’ pattern in
proteins that are translated by free ribosomes. Therefore, we propose that the evolutionary
encoded signal is mainly associated with membrane bound translation. We postulate that
it is a general design for complying with the mechanistic and kinetic demands of a restricted
subset of the proteome.
5.5. Quantitative view of regulation

In our analysis of gene expression regulation, we showed the quantitative aspects of the different sets of regulation on gene expression and translation. At the level of translation regulation, we analyzed several cellular settings including healthy, transformed cell lines, and cancerous tissues. As the absolute tRNA levels are known to increase in transformed cells by about 20-fold (Pavon-Eternod et al., 2009), we showed that the tRNAs that change their overall expression roughly maintain their relative concentrations upon a wide range of conditions. Moreover, the change in codon usage among cell lines of different identity is negligible, as we found strong significant resembles of tRNA expression levels of 12 different breast tissues (three healthy and nine cancerous). Therefore, we postulated that the genomic tRNA copy number is a reliable and valid approximation for their expression levels. As mentioned above, the impact of a specific miRNA manipulation is usually studied at the level of its own target gene, ignoring the multiple influences at a cellular level. However, modeling the miRNA – mRNA regulation according to its stochastic and probabilistic approach showed that this level of regulation has an effect on the total mRNA proportions and quantities. I found that simulating the miRNA interaction network by providing the gene expression profile of one cell type and the miRNA expression profile of another cell type result in an available mRNA profile that mimics the cell type from which the miRNA profile was considered. Another interesting finding is that there is an importance to the extent of overexpression manipulation. We found that although most of the genes are indifferent to the level of overexpression, about 30% behave differently in different overexpression level of a specific miRNA. Moreover, we found that running the miRNA-mRNA simulator using miRNA expression from one cell and mRNA expression profile from another cell result a final expression profile that is more similar to the latter one. This is a direct illustration for the dominant role of miRNA composition for cell identity. In this view it provides additional rationale to the observation in which stem cells and cancer cells are best defined by their miRNA profiles. As a direct outcome from this finding, the ability to manipulate the miRNA in cells (and less so their mRNAs) can be used in future therapy, benefiting from the ability to switch between cell states and their identities.


A review of the role of miRNA and mRNA in gene expression regulation.

miRNA molecules are short, single-stranded RNA molecules that play a crucial role in gene expression regulation. They function by binding to complementary sequences in the 3' untranslated regions (UTRs) of messenger RNA (mRNA) molecules, preventing their translation into proteins. The binding of miRNA to mRNA results in degradation or suppression of translation, thereby silencing the expression of specific genes.

miRNA molecules are expressed in a wide variety of tissues and cells, and their expression levels can be affected by various factors, including microenvironmental conditions. The levels of miRNA expression are tightly regulated to ensure proper gene expression and cell function.

miRNA molecules are involved in various biological processes, including development, differentiation, and disease. They have been implicated in the regulation of cancer, inflammation, and autoimmune diseases.

In conclusion, the role of miRNA in gene expression regulation is essential for maintaining cellular homeostasis and proper gene expression. Further studies are needed to fully understand the complex regulatory mechanisms involved in miRNA-mediated gene expression regulation.
Translation

Regulation

Regulation of gene expression is one of the most important processes in the cell. It is composed of several different levels of control, and is studied in depth at its various stages, starting with the control of the synthesis and transport of mRNA from the nucleus to the cytoplasm, and additional control processes that occur in the cytoplasm and include control of the level of the mRNA itself, as well as control of the translation process. One of the mRNA control processes in the cytoplasm includes control of microRNA (miRNA). miRNAs are small molecules of RNA not encoding, which are about 22-25 nucleotides in length. In humans, there are about 1900 genes encoding miRNA, and about 2500 mature miRNA. Today, we estimate that about 60% of the mRNAs are regulated by miRNA. Usually, the consequences of miRNA control are a decrease in the level of mRNA and translation inhibition, and this is done by base pairing in the region of the miRNA site (MBS), which is usually found in the 3'UTR of the mRNA. It is important to note that the interactions of the miRNA with their targets actually form a complex network in which one miRNA can identify and bind to several regions of several dozen and even hundreds of different mRNAs. In the different regions of its control, the dynamics and general characteristics of such a complex network are partially understood. In most studies, the effect of miRNA is evaluated by manipulating it and monitoring the consequences on the target gene of that miRNA. Most of what we know today about miRNA is based on the analysis of the results of experiments in which the expression of a specific miRNA is increased or decreased. We found that generally, the level of control that is caused by the binding of a miRNA to the mRNA is moderate and is usually very small. The competition that arises from the network configuration between different sites of the same miRNA was described in the context of the ceRNA hypothesis. In my work, I developed a general framework that models the interactions between miRNA and mRNA using a probabilistic approach. The final output of the simulation is the level of preservation of each gene after modeling the interactions between the miRNA and mRNA, which is based on the interaction affinity, the initial amounts of miRNA and mRNA in the cell, and the availability of each connection site. Surprisingly, different simulations in which the expression of more than 250 different miRNAs were increased revealed a list of genes whose level of expression decreased significantly in almost all experiments of increased expression of each miRNA. In addition, we found a large group of genes that were very stable, and were completely insensitive to any manipulation done on the different miRNAs. Furthermore, we found that different cell types have different sets of sensitive genes, while there is a great overlap between the sets of stable genes. A new discovery.

शुभार�भि क्रियाकार तत्कालीन नियन्त्रण एक महत्त्वपूर्ण समारोहोत्तम है जो जीवन के नियमों में समस्याओं का निर्माण, संभाल और नियंत्रण करने के लिए उपयोगी है। यह महत्वपूर्ण है क्योंकि यह जीवन के क्षेत्रों में समस्याओं का संभाल और नियंत्रण करने के लिए उपयोगी है। यह महत्वपूर्ण है क्योंकि जीवन के क्षेत्रों में समस्याओं का संभाल और नियंत्रण करने के लिए उपयोगी है। यह महत्वपूर्ण है क्योंकि जीवन के क्षेत्रों में समस्याओं का संभाल और नियंत्रण करने के लिए उपयोगी है। यह महत्वपूर्ण है क्योंकि जीवन के क्षेत्रों में समस्याओं का संभाल और नियंत्रण करने के लिए उपयोगी है। यह महत्वपूर्ण है क्योंकि जीवन के क्षेत्रों में समस्याओं का संभाल और नियंत्रण करने के लिए उपयोगी है। यह महत्वपूर्ण है क्योंकि जीवन के क्षेत्रों में समस्याओं का संभाल और नियंत्रण करने के लिए उपयोगी है।
עבודה זו נעשתה בהדרכת של:

פרופסור מיכל ליניאל
חקר רמות הבקרה שלلاحプラス השיעוט:
מיזול הדינמי של אנטראקטיזות
מולקולריות בタンינים חים

תיברレーション לשם קבלת תואר דוקטור לפילוסופיה

מת
שלי מחלב-איביב

הוגש לסינט האוניברסיטה העברית בירושלים

יונ 2018