עבודת גמר (תזה) לתואר
דוקטור לפילוסופיה

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

Thesis for the degree
Doctor of Philosophy

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

מאת
הילה גינגולד

By
Hila Gingold

יחסי גומלין דינמיים בין שימוש בקודונים והמאגר התאי של מולקולות רנ"א מוביל משפיעים על יעילות התרגום ועשויים למלא תפקיד בעיצוב גורלו של התא

Dynamic interplay between codon usage and the tRNA pool affects translation efficiency and may play a role in shaping the cell fate

מנחה:
פרופ' יצחק פלפל

Advisor:
Prof. Yitzhak Pilpel

ניסן התשע"ג

March 2013

Table of Contents

1. **Abstract**

In a comprehensive model for translational efficiency the process should be thought of in terms of demand vs. supply, with supply being the tRNAs availability, and the demand captured by the actual representation of the various codons in the transcriptome. Prevailing models for translation elongation efficiency of genes often assume that the process occurs at constant efficiency and fidelity for each gene throughout organism life. Towards a next-generation model of translation elongation we study the factors that govern dynamics supply of the tRNAs along with potential complementary dynamics of the demand from the codon usage of genes. The reasoning behind this notion is that if the gene's codons are highly represented in the transcriptome at a given condition, then its translational efficiency might be compromised. This thesis consists of three chapters that study various aspects of the dynamism in demand and supply during translation.

In the first chapter we reveal a global tendency of distinct species to increase the representation of low-efficiency codons in the translated transcriptome upon stressful conditions, implying for poor translation of stress-related genes, presumably due to lack of sufficient evolutionary optimization pressure on their codon usage.

Considering translation efficiency as a dynamic attribute, we examine in the second chapter the potential changes in supply and demand in translation elongation upon cancer. Utilizing data from customized microarrays we detected recurring changes in the tRNA pool of human cancerous cells. Intriguingly, we found that the cancerous tRNA pool is predicted to selectively boost the translation efficiency of genes associated with proliferation processes. Moreover, we show that such differential effect of the cancerous tRNA pool on translation efficiency is governed by a so-far unrecognized dichotomy in the codon usage of proliferation- and differentiation-related genes. Specifically, cancer appears to predominantly enhance the expression of tRNAs whose corresponding codons are enriched among the proliferative genes, and repress the expression of the tRNAs whose corresponding codons are enriched among the differentiation-related genes. In fact, we show that the cancerous tRNA pool boosts the translation of the same genes whose mRNAs is elevated upon cancer, suggesting that changes in translation efficiency may mediate "switching" between proliferation and differentiation modes in normal physiology and in cancer.

In the third chapter we challenge the traditional conception of translation efficiency by suggesting that local pools of “recycled” tRNAs in the vicinity of the codon at the ribosome A-site may boost translation efficiency. Consistent with our hypothesis, we found that in subsequent occurrences of the same amino acids, highly expressed genes tend to use same codons repetitively in cases that the same amino acid is encoded. In the fourth chapter I discovered that opposing to the prevailing notion that associates translation accuracy with preference of high-efficiency codons, conserved positions of certain amino acids of yeast genes tend to be encoded with codons that are relatively immune to translation errors, even if these codons are inferior in terms of translation efficiency. Together my thesis provides foundations for a new model for translation efficiency and fidelity, a model in which dynamism of all components is captured towards a more faithful description of the conversion of the transcriptome into the proteome.

2. Introduction

Gene expression is one of the most central molecular processes in living cells. Organisms invest a considerable amount of their resources, including energy, raw material and information bandwidth, to carry out the process while optimizing efficiency, responsiveness and accuracy. During evolution, organisms evolved sophisticated means to achieve all of these goals and to balance between them when needed. Efficiency of gene expression consists of the throughput of the process on one hand and of its costs on the other (Dekel and Alon 2005). The costs of the process are numerous and they consist of investment of building blocks, energy and allocation of cellular resources, such as the ribosomes and tRNAs (Stoebel et al. 2008). Accuracy can be described as the probability that the translated protein will be error-free and match the sequence prescribed by the encoding gene sequence, in addition to the likelihood that it will fold properly within the cell (Drummond and Wilke 2008; Zhou et al. 2009). The advent of modern genomics and systems biology has revolutionized our understanding of the diversity of molecular and systems-level mechanisms that control and optimize translation efficiency and accuracy (Arava et al. 2003; Dittmar et al. 2004; Lackner et al. 2007; Hendrickson et al. 2009; Ingolia et al. 2009).

The translation process is highly regulated by a variety of structural elements and sequence motifs (Kozak 1986; Jackson et al. 2010), and it is responsive to biological and environmental conditions (Loh and Song 2010; Spriggs et al. 2010). While classical studies delineate regulation at the level of initiation as the key factor in translation control (Jackson 2010 (Sonenberg and Hinnebusch 2009; Jackson et al. 2010), recent evidences reveal that regulation at the level of elongation plays a major role in shaping translation efficiency (Cannarozzi et al. 2010; Tuller et al. 2010) and translation accuracy (Drummond and Wilke 2008), and even affects the folding of proteins (Sorensen et al. 1989; Komar et al. 1999; Kimchi-Sarfaty et al. 2007; Drummond and Wilke 2008; Zhou et al. 2009; Tuller et al. 2010). This body of recent works and emerging concepts are reviewed in our recent paper (Gingold and Pilpel 2011).

The diverse regulatory mechanisms that act at the level of translation elongation are associated with the redundant nature of the universal genetic code. The apparent redundancy of the genetic code allows the choice between alternative codons for all but two of the amino acids. Although such alternative codons are traditionally termed

'synonymous', it is now well accepted that the implications of the choice between them on the translation process is far from being equivalent. The differential effect of synonymous codons on translation is typically attributed to two main factors - differences in the secondary structure of the transcripts that is determined by the nucleotide sequence (Andersson and Kurland 1990; Kudla et al. 2009), and differences in the amounts of their corresponding tRNAs in cells.
On the one hand, the tightness of the mRNAs secondary structure might control both the ribosome binding and the rate of its flow across them. On the other hand, the speed at which a codon is translated is expected to increase with the availability of its cognate amino acid-loaded tRNAs. Hence, the astronomical number of alternative nucleotide sequences that could still code for the same protein leaves many degrees of freedom that evolution could use for achieving control without affecting the protein sequence.

Non-random usage of synonymous codons was observed decades ago, and was interpreted as reflecting selective pressures for translational selection (Ikemura 1985; Shields et al. 1988; Stenico et al. 1994; Moriyama and Powell 1997). Formal measures of translation efficiency of genes have been developed, where the common models either measure the codon bias of genes - i.e., the non-random assignment of codons to amino acids, or additionally consider the availability of its corresponding tRNAs (Ikemura and Ozeki 1983; Sharp and Li 1987; dos Reis et al. 2004).
As genomic data for coding sequences and measured levels of gene expression accumulate, the early evidences are now more established. A consistent trend of increased usage of codons that correspond to the most abundant tRNAs, especially in highly expressed genes, was detected in bacteria (Lithwick and Margalit 2003). In yeast species it was found that entire gene modules, pathways and complexes might show coordinated selection for translation efficiency in some species, but not in others, depending on lifestyle needs. For instance, while genes belonging to fermentative pathways are codon-optimized in anaerobic species, respiratory genes show selection of optimal codons in aerobic yeasts (Man and Pilpel 2007), and in other related cases (Jiang et al. 2008). Selection for translation efficiency was shown also in some multicellulars such as *C. elegans*, *D. melanogaster* and *Arabidopsis thaliana* (Duret and Mouchiroud 1999; Duret 2000; Heger and Ponting 2007; Drummond and Wilke 2008). Yet, attempts to demonstrate selection for translation efficiency in human, and to further correlate it with expression levels, yield

contradictory results, with some works suggesting that efficiency of translation is under selection, while others suggest to the contrary—reviewed in (Chamary et al. 2006). Translational selection is also emerging in the context of virus-host interactions. Several studies showed codon bias in genes of bacteriophages towards their bacterial host codon bias (Sharp et al. 1984; Carbone 2008; Lucks et al. 2008; Bahir et al. 2009), suggesting selection for efficient translation of the viral genes. A comprehensive analysis showed that the specific sets of viral-encoded tRNA genes were selected by the virus during evolution, presumably as they may boost translation efficiency of virus's own genes (Bailly-Bechet et al. 2007).

The extent of adaptation between the cellular tRNA pool and the codon usage of genes is typically thought of in the context of an evolutionary time scale, and not in the context of physiological time scale processes. In particular, the tRNA pool is typically considered to be constant throughout the life of the cell and across cell types, tissues and organs of a multi-cellular organism. Yet, this notion of fixed tRNA pool was recently challenged. Measurements of the human tRNA pool in different tissues indicate for variation in the availability of the various tRNA types between different cell types (Dittmar et al. 2006). Moreover, condition-dependent expression of tRNAs was reported for yeast - the cellular tRNA pool seems to change in the transition from fermentation to respiration (Tuller et al. 2010). Likewise, the tRNA pool might change during development - the replacement of seven suboptimal codons by optimal ones in the ADH gene of Drosophila led to in vivo increase of its activity in third-instar larva, but in the adult flies it resulted in reduced activity of this gene (Hense et al. 2010). This result might reflect differences in tRNA pools between larvae and adult flies, though the authors of that consider additional possibilities.

Traditionally, the extent of adaptation between the codon usage of a given gene to the cellular tRNA pool was assumed to mainly affect the gene's expression. Intriguingly, recent studies suggest that codon bias of individual genes may regulate global processes in the cell and even determine its fate. A dramatic discovery was recently published describing an interferon-induced human gene that attenuates viral infection by sequestrating tRNAs that are favored by the viral genes' codon usage. Thus, by altering tRNA availability, the host cell selectively inhibits translation of viral genes (Li et al. 2012). Another dramatic demonstration illustrates the difference in the oncogenic potential of two homologs of the Ras oncogene that differ in the extent of adaptation of their codons to the tRNA pool (Lampson et al. 2013).

Interestingly, not only codon bias of individual genes but also differential codon usage of particular functional gene sets can be of great physiological consequence. One prime example is the observation that human cell cycle genes are enriched with low-efficiency codons, suggesting a potential mode of control of cell cycle through changes in translation efficiency (Frenkel-Morgenstern et al. 2012). Similarly, in cyanobacterium and *Neurospora*, low-efficiency codons were shown to be associated with control of circadian rhythmicity (Xu et al. 2013; Zhou et al. 2013).

3. Determinants of translation efficiency and accuracy – published review

(A review by Gingold & Pilpel was published in *Mol Syst Biol*)

Abstract

Proper functioning of biological cells requires that the process of protein expression be carried out with high efficiency and fidelity. Given an amino-acid sequence of a protein, multiple degrees of freedom still remain that may allow evolution to tune efficiency and fidelity for each gene under various conditions and cell types. Particularly, the redundancy of the genetic code allows the choice between alternative codons for the same amino acid, which, although 'synonymous,' may exert dramatic effects on the process of translation. Here we review modern developments in genomics and systems biology that have revolutionized our understanding of the multiple means by which translation is regulated. We suggest new means to model the process of translation in a richer framework that will incorporate information about gene sequences, the tRNA pool of the organism and the thermodynamic stability of the mRNA transcripts. A practical demonstration of a better understanding of the process would be a more accurate prediction of the proteome, given the transcriptome at a diversity of biological conditions.

Molecular Systems Biology 7; Article number 481; doi:10.1038/msb.2011.14
Citation: ***Molecular Systems Biology*** **7:481**

www.molecularsystemsbiology.com

molecular systems biology

REVIEW

Determinants of translation efficiency and accuracy

Hila Gingold and Yitzhak Pilpel*

Department of Molecular Genetics Weizmann Institute of science, Rehovot, Israel
* Corresponding author. Department of Molecular Genetics, Weizmann Institute of science, Herzel, Rehovot 76100, Israel. Tel.: +97 28 934 6058; Fax: +97 28 934 4108; E-mail: pilpel@weizmann.ac.il

Received 29.10.10; accepted 15.2.11

Proper functioning of biological cells requires that the process of protein expression be carried out with high efficiency and fidelity. Given an amino-acid sequence of a protein, multiple degrees of freedom still remain that may allow evolution to tune efficiency and fidelity for each gene under various conditions and cell types. Particularly, the redundancy of the genetic code allows the choice between alternative codons for the same amino acid, which, although 'synonymous,' may exert dramatic effects on the process of translation. Here we review modern developments in genomics and systems biology that have revolutionized our understanding of the multiple means by which translation is regulated. We suggest new means to model the process of translation in a richer framework that will incorporate information about gene sequences, the tRNA pool of the organism and the thermodynamic stability of the mRNA transcripts. A practical demonstration of a better understanding of the process would be a more accurate prediction of the proteome, given the transcriptome at a diversity of biological conditions.
Molecular Systems Biology **7**: 481; published online 12 April 2011; doi:10.1038/msb.2011.14
Subject Categories: RNA; proteins
Keywords: codon usage; translation accuracy; translation efficiency; tRNA

Introduction

Expression of genes is one of the most central molecular processes in living cells. Organisms invest a considerable amount of their resources, including energy, raw material and information bandwidth, to carry out the process, while optimizing efficiency, responsiveness and accuracy. During evolution, organisms evolved sophisticated means to achieve all of these goals and to balance between them when needed. Efficiency of gene expression consists of the throughput of the process on one hand and of its costs on the other (Dekel and Alon, 2005). The costs of the process are numerous and they consist of investment of building blocks and energy and allocation of cellular resources, such as the ribosomes and tRNAs (Stoebel *et al*, 2008). Accuracy can be described as the probability that the translated protein will be error free and match the sequence prescribed by the encoding gene sequence, in addition to the likelihood that it will fold properly within the cell (Drummond and Wilke, 2008; Zhou *et al*, 2009). The advent of modern genomics and systems biology has revolutionized our understanding of the diversity of molecular and systems-level mechanisms that control and optimize translation efficiency and accuracy (Arava *et al*, 2003; Dittmar *et al*, 2004; Lackner *et al*, 2007; Hendrickson *et al*, 2009; Ingolia *et al*, 2009).

The apparent redundancy of the genetic code, in which most of the amino acids can be translated by more than one codon, offers evolution the opportunity to tune the efficiency and accuracy of protein production to various levels while maintaining the same amino-acid sequence. The various codons that correspond to the same amino acid are often considered 'synonymous,' yet their corresponding tRNAs might differ in their amounts in cells and thus also in the speed in which they will be recognized by the ribosome (Varenne *et al*, 1984; Sorensen *et al*, 1989). Also, the alternative nucleotide sequences of the various codon choices for a protein might give rise to transcripts with different secondary structure and stability, which may affect translation (Kudla *et al*, 2009) and even folding (Komar *et al*, 1999; Kimchi-Sarfaty *et al*, 2007). The number of alternative nucleotide sequences that could still code for the same protein is astronomical, leaving many degrees of freedom that evolution could use for achieving control without affecting the protein sequence. While the non-random usage of synonymous codons is often correctly assumed to reflect the action of neutral drift, in an increasing number of cases it now turns out to reflect the result of natural selection, perhaps mainly for tuning efficiency and accuracy of translation (Drummond and Wilke, 2008; Cannarozzi *et al*, 2010; Tuller *et al*, 2010a). The translation process is highly regulated by diverse structural elements and sequence motifs during each of the initiation, elongation and termination steps. Recent studies have enlightened our understanding of translational regulation, for both natural and stress conditions (Loh and Song, 2010; Spriggs *et al*, 2010). In this review, we will focus on the dissimilar, sometimes even opposite effect of different synonymous codons on both translation efficiency and accuracy.

Quantification of translation efficiency

During evolution, cells evolved means to tune the efficiency of translation of different genes to different desired levels. Some gene products are needed in higher amounts than others, while the expression of others, such as regulatory proteins tends to be low. Perhaps more challenging are genes that need to be translated at various levels in different conditions (Takagi *et al*, 2005; Lu *et al*, 2006; Ingolia *et al*, 2009). A more formal

Table I Traditional measures of translation elongation efficiency

Index name	The model by which translation efficiency of a gene is estimated	Properties of translation elongation efficiency measure			
		Explicitly consider the tRNAs availability	Considers the effect of amino-acid composition	Discrimination between translation efficiency of individual codons	Complexity[a] of implementation for many species
The frequency of use of optimal codons, F_{op} (Ikemura, 1981)	The measure quantifies the fraction of optimal codons in a gene	Yes	No	Low[b]	High
Codon Bias Index, CBI (Bennetzen and Hall, 1982)	Measure of the fraction of codon choices, which is biased to *n* preferred codons (relative to random usage of synonymous codons)	Yes	No	Low[b]	High
The codon adaptation index, CAI (Sharp and Li, 1987)	The geometric mean of the ratios of the frequency of each codon in highly expressed genes to the frequency of its most abundant synonymous codon	No	Partially[c]	Partially[d]	Moderate
The 'effective number of codons', Nc (Wright, 1990)	Measures the extent to which the codon usage of a gene departs from equal usage of synonymous codons	No	No	None	Very low
The tRNA Adaptation Index, tAI (dos Reis *et al*, 2004)	The geometric mean of the availability of the tRNAs that serve each codon	Yes	Yes	High	Low

[a]The complexity of implementation is evaluated by the nature of the required input data. Trivially, all measures weight the number of occurrence of each of the 61 codons in the gene of interest. Additionally, the tAI measure requires the identification of all tRNA genes in the genome and their classification according to their anticodons, whereas the CAI measure requires a reference set of known highly expressed genes. The implementation of the F_{op} and CBI measures obligates a reference set of identified 'optimal' or 'preferred' codons, which are dominantly used in highly expressed genes, respectively.
[b]The measure classifies codons into only two categories.
[c]The score weights different patterns of distribution of synonymous codons. Yet, the values of two hypothetical genes that differ from each other by their amino-acid composition, but use only the most abundant codons, are identical.
[d]Codons that do not appear in the reference set were assigned with a fixed frequency.

treatment of the question 'what is the optimal level of expression of a given protein' suggests that the level should be such that the benefit due to expression of the gene should exceed the costs of its production at that level (Dekel and Alon, 2005). Evolving a genome-wide translation regulation regime thus amounts to determining the efficiency of translation of various genes at different conditions, cell types and tissues.

The various genes in the genome, depending on their sequence, might be more or less efficient in consuming the cellular resources of translation, including the ribosomes, the tRNAs, the aminoacyl tRNA synthetases, amino acids, translation factors and energy. A major challenge is to model and predict translation efficiency from the sequences of genes. A sign of success in the future would be the ability to predict protein abundances genome wide in various cell types and conditions.

Traditional computations of translation elongation efficiency (see Table I) may consider the mRNA coding sequence alone and may additionally include explicit inspection of the tRNA pool. Models of the first type, which measure the codon bias of genes—i.e., the non-random assignment of codons to amino acids—revealed decades ago that a striking correlation exists between codon usage and expression levels (Grantham *et al*, 1981; Bennetzen and Hall, 1982; Gouy and Gautier, 1982). In these models, genes that have a codon usage pattern reminiscent of selected 'elite' highly expressed genes are likely to be highly expressed too. The most common index of this sort is the codon adaptation index, CAI (Sharp and Li, 1987). The CAI defines the relative adaptiveness of an individual codon encoding a given amino acid as the ratio of the codon's frequency in highly expressed genes to the frequency of the most abundant codon for that amino acid. The CAI for a gene is then calculated as the geometric mean of the relative adaptiveness values of all the codons along that gene.

The second type of measures explicitly considers the tRNA pool, gauging the availability of tRNA at each codon along the gene. The correspondences between tRNA concentration and translation elongation speed are based on earlier observations, indicating that translation elongation rate is positively correlated with the tRNA concentrations of the translated codons (Varenne *et al*, 1984). In *E. coli*, codons corresponding to highly abundant tRNAs are translated as much as sixfold faster than their synonymous tRNA counterparts that occur at lower concentrations (Sorensen *et al*, 1989). Following early works (Ikemura, 1981; Ikemura and Ozeki, 1983), the tRNA Adaptation index, tAI (dos Reis *et al*, 2004) was developed. The tAI follows the mathematical model of the CAI, but it estimates the translation efficiency of a given gene by assessing the availability of the tRNAs that serve each codon rather than the codon usage itself. As tRNA levels are typically not readily measured, the amount of the different tRNAs in cells is often deduced from the copy number of the tRNA-coding genes in the genome. The usage of tRNA gene copy number as a proxy of tRNA abundance is supported by several observations (Dong *et al*, 1996; Percudani *et al*, 1997; Kanaya *et al*, 1999; Tuller *et al*, 2010a). When calculating the tAI, the tRNA availability of a given codon incorporates both the approximated tRNA levels of its fully-matched tRNA, as well as

contributions from tRNAs that contribute to translation through Crick's wobble rules (Crick, 1966). An obvious advantage of the tAI over the CAI is that it alleviates the need to identify *a priori* the 'elite' set of highly expressed genes as a reference. Instead, it only requires the identification of all tRNA genes in the genome and their classification according to their anti-codons. The tAI measure enables a convenient implementation for many species, and yet, its assumptions regarding the relative strength of imperfect codon–anticodon pairing should be further tuned (Ran and Higgs, 2010). Nonetheless, in studies in a collection of yeast species, both measures correlated highly with mRNA levels (Pearson's correlation 0.6–0.7) in a genome-wide survey (Man and Pilpel, 2007).

But should we expect tAI and CAI values of genes to correlate with the corresponding mRNA or protein abundances? To begin with, mRNA and protein abundances are often correlated between themselves (de Sousa Abreu *et al*, 2009; Vogel *et al*, 2010) so that any measure that correlates with one of them might show above-random levels of correlation with the other. Ideally, a measure of translation efficiency should correlate with the ratio of protein to mRNA level, and indeed the tAI has been shown to correlate with measures of this sort. In *S. cerevisiae*, the simple correlation between tAI and protein-to-mRNA ratio is very weak compared with the correspondence between tAI and mRNA levels, and yet it is still statistically significant (Pearson's correlation=0.123, *P*-value=1.47×10^{-9}). The correlation between protein abundance and tAI, given the genes' mRNA levels, however, is higher (Pearson's partial correlation=0.38, *P*-value=8.54×10^{-81}; Tuller *et al*, 2010b). Similarly, significant positive correlations were detected between tAI and protein levels for sets of yeast proteins having the same mRNA levels (Man and Pilpel, 2007). Furthermore, in *S. cerevisiae*, the contribution of codon choice to the variations in the mRNA–protein correlation remains of prime importance even where RNA decay and protein half-life are taken in consideration (Wu *et al*, 2008). Interestingly though, measures such as CAI and tAI have been shown (especially in unicellulars) to correlate with both mRNA and protein levels, yet probably due to completely different reasons (Figure 1). More intuitive is the correlation with protein levels—high CAI or tAI values for genes should increase translation efficiency and thus increase protein levels at a given mRNA level. Less intuitive is the correlation between mRNA levels and CAI or tAI. Non-optimal codon usage of genes can be detrimental to the cell as it will increase the sequestration of ribosomes during translation, while usage of preferred codons might optimize the allocation of ribosomes to certain genes (Andersson and Kurland, 1990; Kudla *et al*, 2009). The interesting point is that the weight of such effects depends on mRNA levels, so that wasteful sequestration of ribosomes on a low copy mRNA will have a minor effect on the cellular ribosomal pool. Thus, the evolutionary pressure to optimize the codons of genes should increase with their mRNA levels, thereby presumably creating the correlation between mRNA levels and measures such as CAI and tAI.

Advanced challenges in assessing translation efficiency and accuracy

The tAI and the CAI measures predict gene expression with reasonable accuracy, yet alleviating some of the assumptions

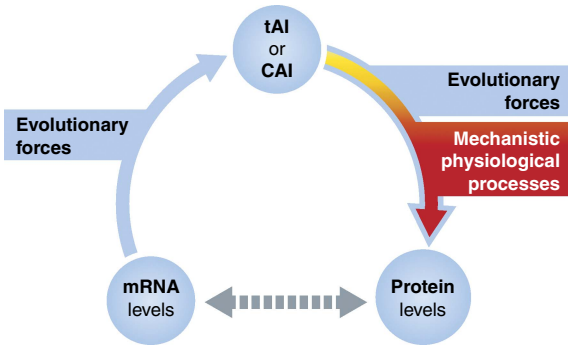


Figure 1 mRNA levels have an evolutionary effect on translation efficiency, which in turn affects protein levels on a physiological timescale. The positive correlation between mRNAs level to measures of translation efficiency, such as CAI or tAI, might reflect an evolutionary pressure to optimize the codon usage of highly expressed mRNAs so as not to sequester too many ribosomes—the faster the elongation rate is, the shorter the time in which a ribosome is bound to any particular mRNA. The extent of evolutionary pressure to optimize a gene should thus positively correlate with its mRNA level. On the other hand, the positive correlation between translation efficiency measures and protein abundance probably acts on a much faster timescale, of mechanistic physiological processes, and it is also governed by evolutionary forces. The codon usage of proteins that are needed at high expression levels is adjusted to achieve high-translation efficiency at a given mRNA level. The significant correlation between the tAI and protein-to-mRNA ratio suggests the causal effect on protein levels.

on which they are based might lead to more accurate models of translation efficiency (see Figure 2).

First, we need to estimate the concentration of amino acid-loaded tRNAs. The life cycle of a tRNA molecule is complicated, it requires transcription, further processing including base modification and charging with amino acid. Recent measurements (Zaborske *et al*, 2009) are beginning to supply estimates on availability of 'ready-to-translate' tRNAs and in general such abundance levels might deviate from the copy number of the tRNA genes, and even from just the concentration of the tRNA molecules in the cell. For example, amino-acid starvation differentially affects the charging levels of isoaccepting tRNA species, leading to wide variation in the sensitivity of the translation rate of individual codons to amino-acid deficiency (Sorensen, 2001; Elf *et al*, 2003).

Second, not only the global codon usage of a gene, but also the order of the high- and low-efficiency codons along the gene may affect translation efficiency. According to measures such as CAI and tAI, the order of high- and low-efficiency codons along the transcript is ignored. Recent analysis of multiple genomes revealed a trend in which the first approximately 30–50 codons in genes preferentially correspond to more rare tRNAs (Tuller *et al*, 2010a). Such genic sections form 'low-efficiency ramps', which might deliberately attenuate the ribosome during early elongation. The authors showed that such a profile is particularly pronounced in highly expressed genes and, at least in yeast, it is inversely correlated with ribosomal density (experimentally measured by Ingolia *et al* (2009)). This correspondence with the experimentally measured ribosomal density data is an indication that the translation efficiency profile is probably a speed profile, aiming to control the rate of flow of the ribosomes by localizing an early traffic bottleneck (Figure 2A). It was proposed that such deliberate early attenuation enables a jam-free flow of ribosomes once they passed that region, thus reducing the probability of

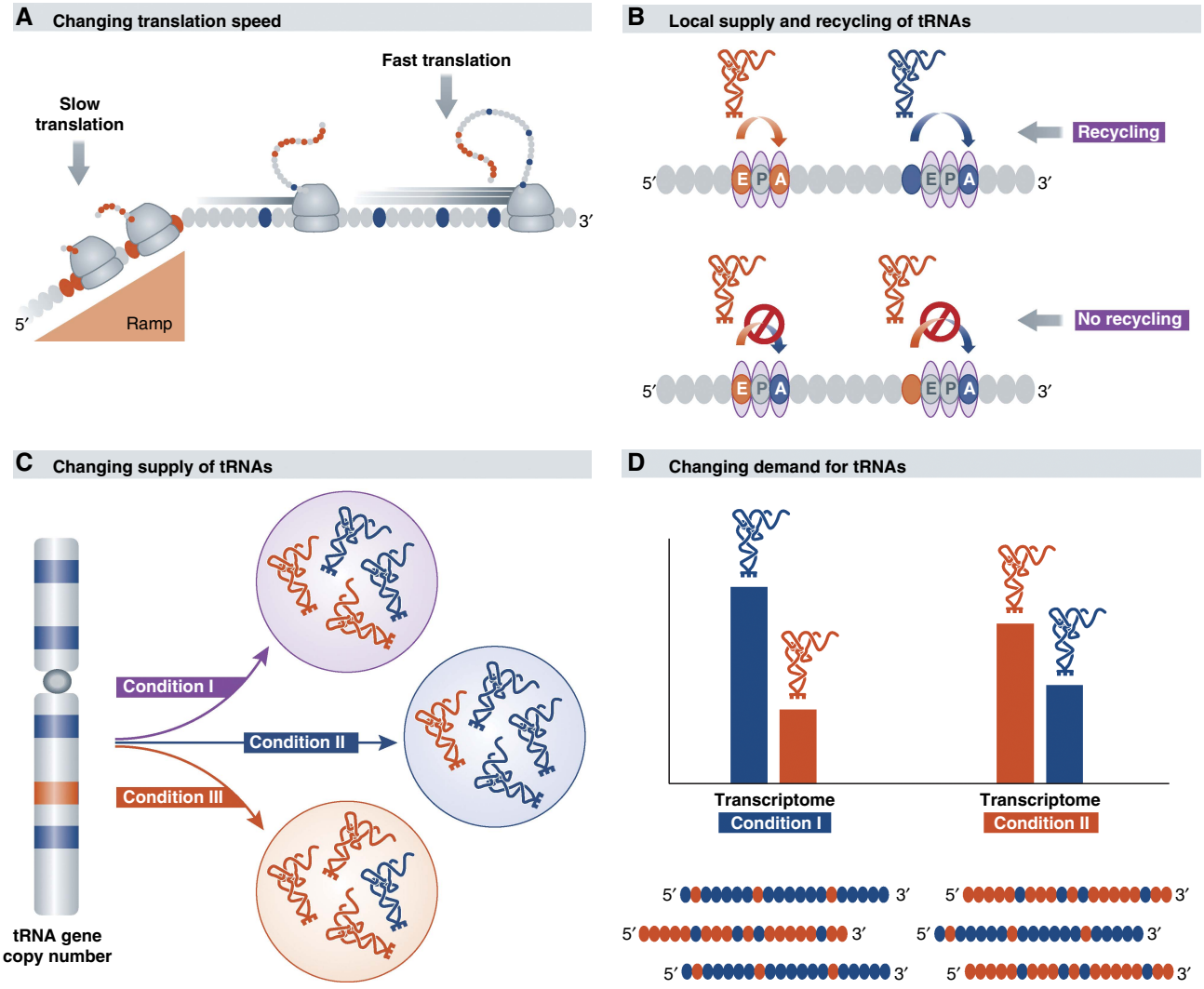


Figure 2 Advanced challenges in assessing translation efficiency. New evidences challenge the common simplified assumptions in assessing translation efficiency. Shown in all sub-figures are two codon types, which may differ in their translation elongation efficiency, a 'blue' and a 'orange', served respectively by a 'blue' and a 'orange' types of tRNA. Some of the amino acids on the polypeptides are also colored blue or orange, reflecting the different efficiency of the codons that code for them. The following lines of further research into the mechanisms of translation are suggested: (**A**) The order of high- and low-efficiency codons (the later are colored in orange) is meaningful and can be utilized by evolution to design an optimal schedule for ribosomal flow on transcripts. In particular, the slow 'ramp' observed in the 5′ end, especially of highly expressed genes, may avoid jamming of ribosomes once they passed it. (**B**) A local concentration of a tRNA molecule that was just released from the ribosome is high in the vicinity of the subsequent codons. Thus, although some tRNAs might be at low concentration over the entire cell volume, they might be present at relatively higher level in proximity of the codons they just finished translating. According to this possibility, the efficiency of translation of a codon depends also on whether that codon was used a few codons upstream on the same mRNA molecule. An indication for the mechanism might be that similar codons tend to cluster together on mRNA sequences. (**C**) Regulation of expression of the tRNAs could lead to dynamic changes in their availability in time or space dimensions, e.g., under various conditions, differential developmental stages, or at different tissues. (**D**) The efficiency of translation is a function of the ratio between the supply and the demand for each tRNA. The demand for different tRNAs, namely—the actual representation of the 61 codons at the transcriptome, might vary between different cell types, different environmental conditions and different time points along organism's life. Here, during the transition from condition I to II, the transcriptome changes from mainly consisting of genes that are rich in the blue codon to genes that more heavily biased towards the orange one; as a result the demand for the corresponding orange tRNA increases in the second condition.

ribosome fall-off. Such a design could increase the productivity of expression while minimizing the costs of the process. This reasoning is consistent with indication of increasing selection against frameshifting errors towards the 3′ end of coding sequences (Huang *et al*, 2009).

Third, local pools of elevated availability of required tRNAs might promote translation elongation efficiency. An implicit assumption of traditional models such as tAI is that all codons utilize the same global tRNA pool. Surprisingly, a recent observation (Cannarozzi *et al*, 2010) implied that the availability of the same tRNAs might be different on different positions along the same mRNA (Figure 2B). This study showed that in subsequent occurrences of the same amino acids, genes tend to deliberately use codons that are translated by the same cognate tRNA. Similar to the ramp design, this trend was shown to be predominantly obeyed by rapidly induced genes, hinting that this is another means to boost translation efficiency. The authors hypothesized that codons at the ribosome A-site can utilize recycled tRNAs from the codons that were just translated. To further establish their hypothesis,

they synthesized variants of the *green fluorescent protein* (*GFP*) gene in which the internal arrangement of synonymous codons either maximized or minimized the potential reuse of tRNAs from near-by position, and observed the expected increase or decrease in expression.

From a kinetic point of view this hypothesis is not trivial. First, it requires that the diffusion of the recycled tRNA will be slow enough compared to the rate of translation elongation. This situation may even necessitate or predict the existence of 'local translation factories' nearby the ribosome, which will supply the re-charging services to the recycled tRNA. Studies indicating the capacity of aminoacyl–tRNA synthetases to interact with the ribosome (Kaminska *et al*, 2009) and reporting on colocalization of protein translation components (Barbarese *et al*, 1995) may serve as supported evidence.

Fourth, the tRNA pool might change dynamically rather than being constant (Figure 2C). According to the simplest models, the tRNA pool is assumed to remain constant throughout the life of a cell and in different cell types of the body. Yet measurements of the tRNA pool in different tissues and cell types showed interesting differences, suggesting that the same gene might be translated differently in each such environment (Dittmar *et al*, 2006). Similarly, in the transition from fermentation to respiration in yeast, the tRNA pool also seems to change (Tuller *et al*, 2010a). Likewise, the tRNA pool might change during development. The replacement of seven suboptimal codons by optimal ones in the *ADH* gene of *Drosophila* led to *in vivo* increase of its activity in third-instar larva, but in the adult flies it resulted in reduced activity of this gene (Hense *et al*, 2010). This result might reflect differences in tRNA pools between larvae and adult flies, though the authors consider additional possibilities.

Finally, the demand for the various tRNAs, presented by the transcriptome, might change dynamically too (Figure 2D). Presumably, the efficiency of translation is a function of the ratio between the supply and the demand for each tRNA. If a given tRNA is highly expressed, but the codons that correspond to that tRNA are highly represented in the transcriptome present at a given condition, then translation efficiency from that tRNA might be compromised in that condition. Interestingly, different codons do indeed fluctuate in their representation in the transcriptome at various conditions (H Gingold, Z Bloom, O Dahan and Y Pilpel, in preparation) emphasizing the need for parallel assessment of the representation of the codons in the transcriptome and the tRNA pool in a richer model of translation efficiency.

Challenging the above assumptions of the simple models may thus result in a more comprehensive model of translation efficiency. Such a richer model might not only improve protein level predictions, it might also explain tissue and condition variation in protein levels, the effects of mutations on translation efficiency, stochastic fluctuation in protein level and rapidity of expression response to signals and changes.

Evolutionary selection for codon—tRNA adaptation

What are the indications that genes were selected during evolution to optimize their translation efficiency? On the face of it one may ask 'why not select for better translation efficiency even if it were to contribute only minutely to fitness?' The answer comes from population genetics that teaches us that traits are fixated in populations not only according to their fitness gain but also due to random drift caused by neutral mutations. In that respect, neutral mutations act like thermal noise in thermodynamic systems; they may prevent fixation of traits with positive, yet small fitness value. The effective population size (Hartl and Taubes, 1998) of a species determines how small the fitness value of a mutation can be while still allowing its fixation. Qualitatively, the rule is simple—the larger the species' effective population size, the higher the probability of fixation. The question of whether the genes in a genome are indeed subject to selective pressure to enhance translation efficiency is thus *a priori* open until rigorous criteria are met, and one would expect that while microbial species, with typically large population sizes, might manifest it, small effective population size species, such as human, might not (Bulmer, 1991; dos Reis and Wernisch, 2009).

As genomic data for coding sequences and measured levels of gene expression started accumulating, the indications of selective pressures for translational selection suggested by early evidences (Ikemura, 1985; Shields *et al*, 1988; Stenico *et al*, 1994; Moriyama and Powell, 1997) are becoming well established. A consistent trend of increased usage of codons that correspond to the most abundant tRNAs, especially in highly expressed genes, was detected in bacteria (Lithwick and Margalit, 2003). In yeast species it was found that entire gene modules, pathways and complexes might show coordinated selection for translation efficiency in some species, but not in others, depending on lifestyle needs. For instance, while genes belonging to fermentative pathways are codon-optimized in anaerobic species, respiratory genes show selection of optimal codons in aerobic yeasts (Man and Pilpel, 2007), and in related cases (Jiang *et al*, 2008). Selection for translation efficiency was shown also in some multicellulars such as *C. elegans*, *D. melanogaster* and *Arabidopsis thaliana* (Duret and Mouchiroud, 1999; Duret, 2000; Heger and Ponting, 2007; Drummond and Wilke, 2008). Yet, as expected from the above population theoretic arguments, attempts to demonstrate selection for translation efficiency in human, and to further correlate it with expression levels, yield contradictory results—reviewed in Chamary *et al* (2006). Some studies found no evidence for translational selection in human (Kanaya *et al*, 2001; dos Reis *et al*, 2004), suggesting that synonymous codons in human are not selected to maximize translation efficiency (Lercher *et al*, 2003). Conversely, other studies do indicate weak, yet significant, translational selection in human, according to estimates of codon usage adaptation to the global tRNA pool (Comeron, 2004; Lavner and Kotlar, 2005). Future related studies may further the exploration of tissue-specific expression patterns of tRNA isoaccpetors (Dittmar *et al*, 2006), and would ultimately be incorporated into more comprehensive measures of translation elongation efficiency.

Translational selection is also emerging in the context of adaptation between viruses and their hosts. Several studies showed codon bias in genes of bacteriophages towards their bacterial host codon bias (Sharp *et al*, 1984; Carbone, 2008; Lucks *et al*, 2008; Bahir *et al*, 2009), suggesting selection for efficient translation of the viral genes. Interestingly, the

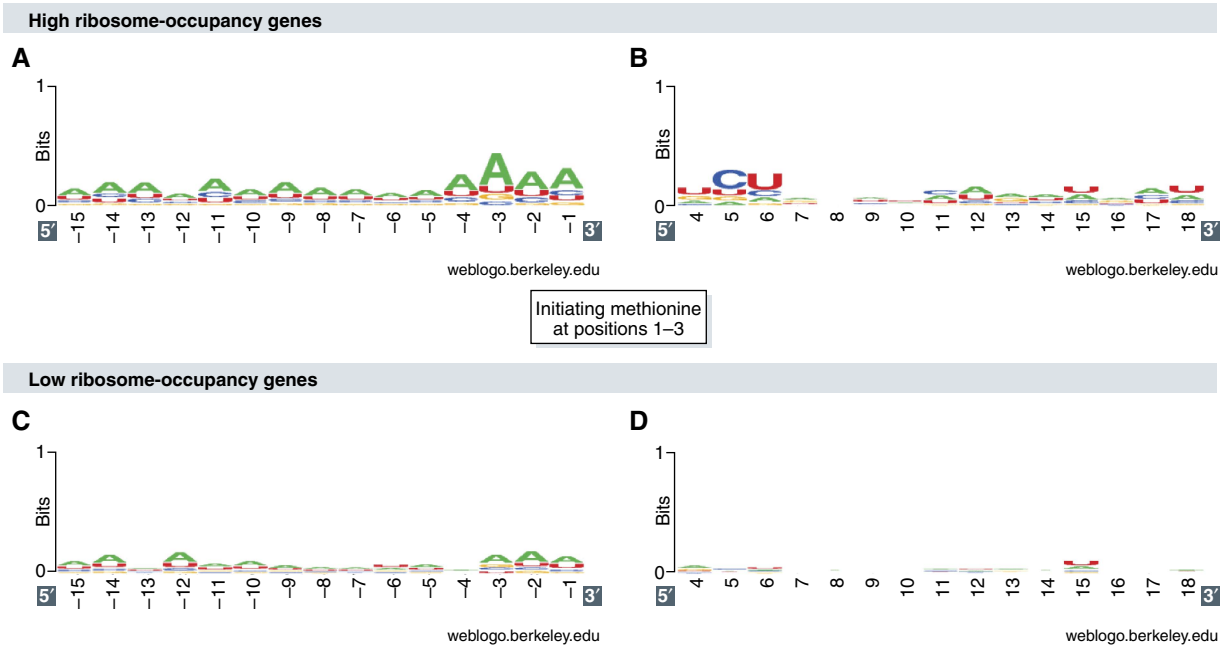


Figure 3 Sequence motifs in the vicinity of the initiation site and ribosome occupancy. The figure displays sequence motif logos of the sequence spanning between positions −15 and +18 relative to the initiating AUG for two yeast genes sets—high ribosome-occupancy genes and low ribosome-occupancy genes (Arava *et al*, 2003). The sequence logos show an interesting signature of enrichment in Adenine nucleotides upstream to the initiating AUG codon in genes with high ribosome occupancy (**A**), accompanied with particular nucleotide preference at positions +5 and +6 (**B**). The 5′ UTR sequence of low ribosome-occupancy genes is also enriched with Adenine nucleotides (**C**), yet to a much lower extent. Genes with low ribosome occupancy show no nucleotide preference downstream to the initiating AUG codons (**D**). For this display, high ribosome-occupancy and low ribosome-occupancy genes (204 and 206 genes, respectively) were defined as genes at the top and at the bottom of the ribosome-occupancy distribution (occupancy > 0.85, or occupancy < 0.6 correspondingly). The 5′ UTR sequences of the investigated genes were derived from the study by Nagalakshmi *et al* (2008); the coding regions were downloaded from SGD web site. Sequence logos were created using WebLogo (Crooks *et al*, 2004).

genomes of some viruses may contain a small selection of tRNA genes that might be added to the cellular tRNA pool and participate in translation upon infection. Why are such tRNA genes selected to be included in the typically very compact viral genome? A comprehensive analysis showed that the specific sets of viral-encoded tRNA genes were selected by the virus during evolution, presumably as they may boost translation efficiency of virus's own genes (Bailly-Bechet *et al*, 2007). An interesting possibility is that the viral tRNA genes might allow the virus to infect also hosts of a wide spectrum of codon usage, thus increasing the bandwidth of potential hosts, by alleviating the need to adapt precisely to the codon usage of each host separately.

Sequence-dependent determinants of translation-initiation rate

The overall speed of translation is determined by the rates of its three major steps—initiation, elongation and termination. The initiation step is regulated by a variety of structural elements and sequence motifs, some of which are uniquely associated with either prokaryotic or eukaryotic organisms (Kozak, 2005; Jackson *et al*, 2010). Such structural elements in eukaryotes are the 7-methylguanosine cap and the poly-(A) tail, which synergistically enhance translation-initiation efficiency (Gallie, 1991) via circularization of the mRNA, which in turn is mediated by interactions with eukaryotic-initiation factors (Tarun and Sachs, 1996; Kahvejian *et al*, 2005). In addition to a contribution of the 3′ end of the transcript to initiation, binding and assembly of the ribosome for a round of translation is governed by the sequence and the mRNA secondary structure in the vicinity of the start codon. In prokaryotes, ribosome binding occurs at the purine-rich Shine-Delgarno (SD) sequence (Shine and Dalgarno, 1974), located a few nucleotides upstream from the start codon, which is complementary to a sequence near the 3′ end of 16S rRNA (Steitz and Jakes, 1975; Jacob *et al*, 1987). In eukaryotes, translation initiation follows a scanning mechanism of the mRNA by the ribosome. The 40S ribosomal subunit enters at the 5′ end of the mRNA and migrates linearly until it encounters the first AUG codon (Kozak, 2002). The ribosome will initiate that first AUG codon if it is flanked by a short sequence motif, known as 'Kozak sequence' (Kozak, 1986).

An important question is whether different variations on the sequence motif in the vicinity of the translation start site are associated with, and perhaps even determining, difference in translation-initiation efficiency. It was previously shown that the 5′ untranslated sequence of yeast mRNAs is rich in A-residues, and that highly expressed genes commonly use the Serine UCU codon as second triplet in the open-reading frame (Hamilton *et al*, 1987). More recently, using data on genome-wide ribosome density (Ingolia *et al*, 2009), Robbins-Pianka *et al* (2010) reported on reduced predicted secondary structure in 5′ UTRs, especially in high ribosome-density genes in yeast. Genome-wide measurements of occupancy and density of ribosomes on mRNA enable us to systematically examine how sequence in the vicinity of the initiation site may affect initiation efficiency. Figure 3 shows a sequence motif logo of the sequence flanking the AUG start codon for two sets of

S. cerevisiae genes—low ribosome-occupancy genes and high ribosome-occupancy genes, based on Arava's analysis of ribosome occupancy (Arava *et al*, 2003). Clearly, high ribosome-occupancy genes show a motif with moderate information content, whereas the low ribosome-occupancy motif shows little or no consensus. Specifically, the analysis shows the preferred usage of the A nucleotide along the 15 positions upstream to the start codon, and in particularly at positions −4 to −1, in high ribosome-occupancy genes. This analysis suggests a hierarchy between genes in the fit of their 5′ UTR sequences to a canonical-initiation motif, which may determine the relative initiation efficiency of each gene in the genome. In addition, for high-occupancy genes, the sequence logo shows a pointed elevated usage of nucleotides C and U, in the 5th and 6th positions in the open-reading frame. Interestingly, the second codon position shows elevated tAI values on average (Tuller *et al*, 2010a) suggesting a selection for high-translation efficiency for efficient release and recycling of the initiator methionine tRNA. Indeed, this signal is more pronounced in genes with high ribosome occupancy compared with genes with low occupancy (H Gingold and Y Pilpel, unpublished data, 2011).

Association between mRNA folding and translation rate

The mRNA molecules in the cell often assume a secondary and a tertiary structure that might be tight for some genes, and loose for others. For translation to proceed, such structure must be threaded through the ribosome. Here is thus another opportunity to regulate and induce wide variation in translation efficiency of genes—the tightness of their mRNA structure might control both the ribosome binding and the rate of its flow across them. Early evidences indicate that the stability of base pairing at the ribosome-binding site or in its vicinity is a major determinant of translation-initiation efficiency in prokaryotes (Schauder and McCarthy, 1989). In eukaryotic organisms, tight secondary structures along the 5′ UTR were shown to reduce translation efficiency, especially if they are located in proximity to the translation start site, presumably by obstructing ribosome binding (Wang and Wessler, 2001).

The effect of mRNA structure on translation was traditionally deciphered by inspecting natural genes from various genomes (Jia and Li, 2005). Now, synthetic biology may to complement this picture by allowing researchers to manipulate one property of a gene, while keeping many others constant. Recently, Kudla *et al* (2009) provided a good example for this modern trend by synthesizing a library of 154 *GFP* genes that varied randomly at synonymous sites, while encoding the same amino-acid sequence. They expressed the GFP genes in *E. coli*, and detected 250-fold variation in expression levels. They found that tight structure at the 5′ end of the mRNA inhibits translation, whereas loose structures promote it. These results are consistent with the notion that the initiation step is of prime importance in determining gene expression levels. In prokaryotes, ribosome binding occurs at the SD sequence (Shine and Dalgarno, 1974) located upstream from the start codon. Interestingly, it was shown before that masking of the initiation site by tight secondary structure can be offset by a stronger-than-normal SD interaction (de Smit and van Duin, 1994; Olsthoorn *et al*, 1995). As Kudla *et al* (2009) only varied the coding region of GFP, this possibility was not tested in their recent study.

The association between the stability of secondary structures in the translation-initiation region and translation efficiency is further supported by large-scale computational analysis (Gu *et al*, 2010), indicating a genome-wide trend of reduced mRNA stability near the start codon for both prokaryotic and eukaryotic species. Here too the trend was found to be enhanced among highly expressed genes, suggesting an effect of translation efficiency.

Determining the overall rate of translation: one key factor or a 'combination lock'?

While it is widely accepted that mRNA folding and codon–anticodon adaptation are the key factors in the determination of initiation and elongation rates, respectively, the identity of the rate-limiting step of the overall translation efficiency remains controversial. Surprisingly, and in contradiction to many studies of natural genes, Kudla *et al* (2009) indicate that the variation in protein expression levels in the GFP library is not derived at all from codon bias differences (measured by the Codon Adaption Index). They proposed instead that the mRNA folding at the beginning of the transcript has the predominant role in shaping expression level of individual genes, whereas selection for codon bias aims to increase the global rate of protein synthesis by reducing the ribosomes sequestering on the mRNA. A related study inspected *E. coli* and *S. cerevisiae* and found similar trends of relatively loose secondary structure stability near 5′ ends of genes (Tuller *et al*, 2010b). The authors investigated the interplay between folding energy and codon bias in determining translation efficiency across all the genes of *E. coli* and *S. cerevisiae*. Unlike the results obtained by Kudla *et al* (2009) for synthetic genes, Tuller *et al* (2010b) observed a significant correlation between codon bias and protein abundance (normalized to mRNA level), but no direct correlation between folding energy and protein abundance. These authors did find, however, that the strength of association between codon bias and protein expression is modulated by folding energy. Part of the reason for this apparent discrepancy between the natural and synthetic genes was suggested to be the different distribution of folding energy values between the two gene sets (Tuller *et al*, 2010b).

Future studies will probably investigate the separate contribution of the diverse determinants of translation efficiency to the overall rate of translation. Such an analysis was carried out for the *Desulfovibrio vulgaris* bacteria, aiming to assess the contribution of sequence features associated with the initiation, elongation and termination steps to the variation in mRNA–protein correlation (Nie *et al*, 2006). Ideally, such studies will take into consideration *in vivo* estimation of mRNA decay and protein degradation as potential confounding factors. This reasoning is consistent with recent studies indicating for higher conservation of protein abundance than mRNA levels across different species, hence implying for major role of either translational or protein degradation

control in maintaining proteins in desired levels (Schrimpf *et al*, 2009; Laurent *et al*, 2010).

An important challenge is to appropriately consider features in the mRNA that affect translation. For example, in addition to its prime effect on ribosome binding and initiation, the secondary structure of mRNA governs the movement of the ribosome during elongation too, suggesting a broader effect of mRNA structure on translation (Wen *et al*, 2008). In that respect, modern investigations broaden the scope of the classical ribosome attenuation model that was originally described as a mechanism relevant to amino-acid biosynthetic genes only (Yanofsky, 1981).

It is interesting to note the difference between the expressions of natural genes in their natural genome compared to man-made heterologous expression systems, in which one often expresses a gene from one species in another species. In both cases, the need to optimize expression of a given protein often arises, but beyond that some of the actual considerations might be very different. A native gene in its natural genome can be highly expressed but only to the extent that the benefit from the gene will not exceed the costs associated with its production. Some of the costs are direct, e.g., consumption of raw material and energy, and some are indirect, e.g., sequestration of the gene expression apparatus. Thus, even the most highly expressed genes in a natural context must be 'considerate' of the rest of the genes in the genome. The situation could be different in artificial systems, especially in the biotechnology context in which a more 'selfish-gene' approach could be justified. Here high expression of a gene in a host may be justified even if overall fitness of the host cell is significantly compromised, as long as the system is economically cost-effective. Another prime difference is that heterologous systems often reach very high expression levels, much beyond even highly expressed genes in their natural genomes. The design considerations of the genes' sequence and their interaction with the cellular machinery in the two cases might thus be very different. We anticipate that future studies will expand upon existing attempts to design nucleotide sequences (given amino-acid sequence constraints) that optimize either fitness of the host or productivity of a given desired protein (Kudla *et al*, 2009; Welch *et al*, 2009; Navon and Pilpel, 2011).

Codon choice may affect translation fidelity

So far we have discussed the effect of codon choice and mRNA structure on the throughput of translation, but these parameters could also govern the fidelity and accuracy of the process. In the stochastic search for the right tRNA, the ribosome might incorrectly bind a tRNA with a one base-mismatch relative to the codon, often termed 'near-cognate tRNA' (tRNAs with more than one base-mismatch relative to the codon typically do not pass the initial screen; Rodnina and Wintermeyer, 2001). If a near-cognate tRNA binds to the A-site of the ribosome, the wrong amino acid might be incorporated, creating a 'missense translational error'. The frequency of such translation errors *in vivo* was estimated to be 10^{-5} in yeast cells (Stansfield *et al*, 1998), but more recent measurements in *B. subtilis* showed a surprisingly high rate of 10^{-2} (Meyerovich *et al*, 2010). Missense errors can also be caused by erroneously charged tRNAs, with an overall error rate of 1 per 10 000 (Ibba and Soll, 2000). Missense errors that might disrupt protein function impose metabolic costs of wasted synthesis; if the loss of function is accompanied with improper folding, the damage might be even more pronounced. The misfolded protein may interact with other cellular components, causing protein aggregation (Bucciantini *et al*, 2002), disruption of membrane integrity (Stefani and Dobson, 2003) and it may ultimately result in cell dysfunction and disease—reviewed in Gregersen, 2006.

Translation can thus be thought of in terms of a competition process between the cognate and near-cognate tRNAs for a given codon, where the higher the concentration of correct tRNAs, the lower the probability of binding the wrong ones. Indeed in *E. coli*, the frequency of missense errors is diminished by ninefold if the same amino acid is translated by a codon that corresponds to an abundant tRNA rather than a low-abundance one (Precup and Parker, 1987).

The association between selection on synonymous site and translation accuracy was quantitatively examined for the first time by Akashi (1994). Akashi (1994) showed higher frequencies of preferred codons in evolutionarily conserved amino-acid positions among *Drosophila* species. Comparing only 38 orthologous genes among fly species, Akashi (1994) found that the frequency of preferred codons is significantly higher at conserved amino-acid positions compared with non-conserved ones. Akashi (1994) thus suggested that selection favors optimal codons at sites where misincorporations are most likely to disrupt protein functions. This type of pioneering analysis was later applied in the full genome era to *E. coli* (Stoletzki and Eyre-Walker, 2007), yeast, worm, mouse and human (Drummond and Wilke, 2008), verifying the significant association between optimal codons and evolutionary conservation, supporting Akashi's early notion that in the very same positions where evolution conserved the amino acid against DNA replication mutations it also insisted on the preferred codons that would minimize the chance for translation errors. Drummond and Wilke (2008) carried out molecular-level evolutionary simulation of the effects of misfolding due to translation errors on fitness. They concluded that selection acts on translation accuracy, but only if misfolding imposes a direct fitness cost. Their study suggested that selection for translation accuracy, although intuitively associated with production of functional proteins, might mainly be derived by the need to globally prevent the toxic consequences of misfolding errors. Selection against misfolding errors were further shown to not only associate with the usage of preferred codons but also with preference of misfolding-minimizing amino acids (Yang *et al*, 2010).

Selection pressure against misfolding is directly supported by studies that focus on structurally sensitive sites, where mutations are highly disruptive. Buried amino-acid residues were shown to be preferentially encoded by more optimal codons compared with solvent-exposed residues (Zhou *et al*, 2009). This is consistent with evidences for higher sensitivity of protein core residues, compared with surface residues, to mutations that occur during DNA replication (Tokuriki *et al*, 2007). The hypothesis of selection against mistranslation-induced protein misfolding is further sustained by a very

different and yet complementary approach (Warnecke and Hurst, 2010). These authors demonstrated coordinated utilization of *cis*-acting (preferred codons) and *trans*-acting (molecular chaperons) elements as a strategy for misfolding prevention. They show that proteins, which attain their native structure spontaneously, or at least without the aid of the bacterial chaperonin GroEL, are enriched with preferred codons at structurally sensitive sites, compared with proteins that need the chaperonin for folding. The study thus suggests that the chaperonin alleviates the need to optimize codons as a means to prevent translation-mediated misfolding. Further, in the context of translation accuracy, selection pressures on synonymous sites also appear to act against frameshifting errors (Farabaugh and Bjork, 1999), and to reduce the cost of nonsense errors (Gilchrist *et al*, 2009).

But 'errors' are sometimes beneficial, and the ability to introduce them when needed may have even been selected for. A striking recent example showed that under certain stresses, a 'programmed translation error' may occur, which leads to increased misincorporation of methionine residues into the mammalian proteome (Netzer *et al*, 2009). Unlike the misincorporation errors discussed above, this phenomenon appears to feature elevation in misacylation of Met residues in non-Met tRNAs. This observation is striking because methionine has a radical oxygen-protective capacity and sure enough operates predominantly under oxidative stress.

The strategic role of the rare: advantageous usage of disadvantageous codons

In the previous sections we described the benefits associated with the usage of codons that correspond to abundant tRNAs—such codons may enhance the speed and accuracy of the translation elongation step. However, it is of interest to understand whether codons which belong to the opposite side of the scale, namely, codons that correspond to the least abundant tRNAs, are also preferred in selected cases, or whether their usage is simply the outcome of the absence of selection for abundant codons (Sharp and Li, 1986). High frequencies of rare codons in lowly expressed genes were observed in many genomes, including human (Lavner and Kotlar, 2005). Rare codons have the potential to slow down the translation elongation rate (Pedersen, 1984), due to the relatively long dwell time of the ribosome in its search for rare tRNAs. Several studies suggest that gene-wide codon bias in favor of slowly translated codons serves as a regulatory means to obtain low expression levels of protein when desired, for example, in the case of regulatory genes, or where excess of the protein appears to be detrimental or lethal to the cell (Konigsberg and Godson, 1983; Zhang *et al*, 1991). The level of protein secondary structure was also found to be associated with codon usage. Particularly, it was found that fast folding α-helical sequences are preferentially encoded by fast codons, whereas slower folding β-sheets strands, loops and disordered structures are enriched with rare (slow) codons (Thanaraj and Argos, 1996a).

More subtle are the cases in which only specific regions within a gene might be strategically selected to feature slow codons. For example, choice of slow codons was suggested to affect co-translational folding—reviewed in Tsai *et al*, 2008. A simple model suggests that the strategic usage of rare codons provides a pause during translation, during which an already translated segment of a protein may be folded in the absence of an otherwise potentially interfering segment that is not yet translated (Komar *et al*, 1999; Tsai *et al*, 2008). Supporting this notion is a study in which 16 consecutive rare codons in a gene were replaced by synonymous optimal ones in *E. coli*. Although the optimal codons enhanced the translation speed, they appear to have reduced folding as deduced by a 20% decrease in the encoded enzyme's specific activity (Komar *et al*, 1999). Such a manipulation in another gene of *E. coli* resulted in elevated *in vivo* misfolding and aggregation rates (Cortazzo *et al*, 2002). A small and yet significant similar effect was also obtained in yeast in a similar experiment (Crombie *et al*, 1992, 1994). Removal of translational attenuation sites in the bacterial *SufI* gene by an alternative approach, in which a global increase of the translation rate was obtained by adding a large excess of naturally rare tRNAs, also resulted in perturbed folding (Zhang *et al*, 2009). The hypothesis that rare codons are employed to temporally separate the synthesis of defined portions of the protein is consistent with the observation that boundaries between domains—proteins' independent folding modules—are enriched with clusters of rare codons (Thanaraj and Argos, 1996b).

In the last decade, the awareness of the fascinating biology of intrinsically unstructured proteins has grown significantly (Gsponer *et al*, 2008). The function of such proteins often depends on them being unstructured, and hence there have been extensive computational (Uversky *et al*, 2000) and experimental (Tsvetkov *et al*, 2008) efforts to identify such proteins genome wide. Common to such attempts is the search for signals in the protein amino-acid sequence that determine its lack of structure. A plausible hypothesis is that obtaining an unfolded structure also requires instructions from the nucleotide sequence, and in particular that coupled translation-folding determines unstructureness. Could it be that the strategic choice of certain codons, e.g., fast codons in domain boundaries, can actually serve to reverse the above-mentioned folding-promoting design, so that a protein will be unfolded? In general, is there a code of translation efficiency that is needed to create an unfolded protein? Can the effect of codon choice on folding pathways be simply referred to as either 'beneficial' or 'deleterious?' The answer is probably 'no.' A naturally occurring mutation in the human *MDR1* gene, involving a synonymous rare-to-frequent codon substitution, led to slight alternation in the native tertiary structure of the protein and subsequent change in its substrate specificity (Kimchi-Sarfaty *et al*, 2007). The wide potential impact of the co-translational folding timing is further manifested by a recent observation that codon usage might affect post-translation modification and folding, and as a consequence the stability of a protein due to a forced choice between ubiquitination and an alternative modification (Zhang *et al*, 2010). More generally, an interesting possibility is that proper post-translation modification of proteins, which sometimes takes place during the 'pioneering round of translation' while the nascent chain emerges from the ribosome, may require a certain optimal tempo of translation. We may thus anticipate that some modifications, including myristylation that occur co-translationally (Wilcox *et al*, 1987) or others such as glycosylation, may require a certain rate of

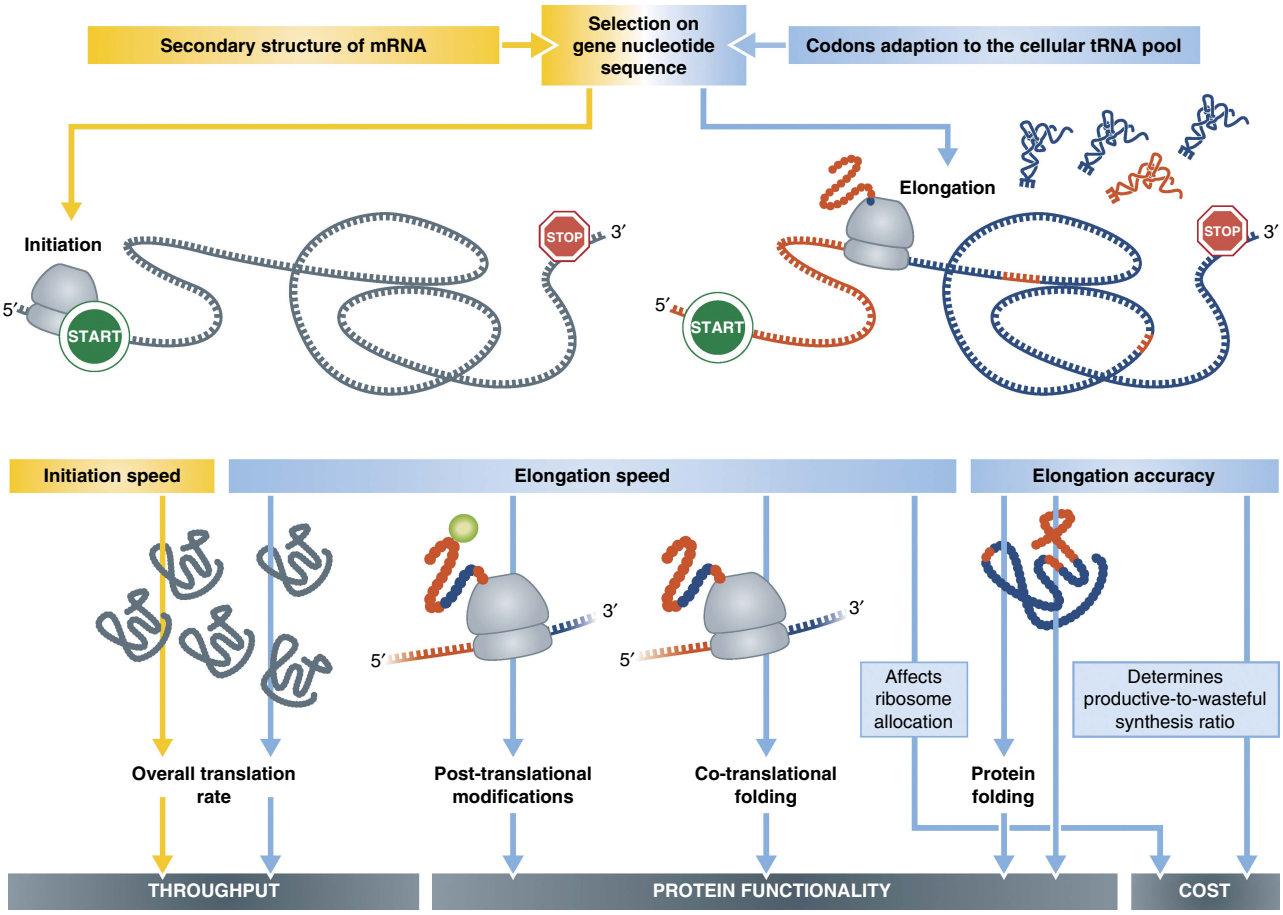


Figure 4 The predominant effect of selection on synonymous site on gene expression. Synonymous codons correspond to the same amino acid and yet might differ from each other by their adaptation to the cellular tRNA pool and also by their contribution to the secondary structure and the stability of the transcript. The effect of each attribute on translation properties and the further consequences on gene expression is marked with pale blue (for codon–anticodon adaptation) or yellow (for mRNA folding).

translation in their vicinity. Thus, the nucleotide sequence that codes for the protein, and not only its amino-acid sequence, may determine the modifications. In that respect it is interesting to note that highly predictive amino-acid motifs for some modifications remains elusive, and it might thus be that inclusion of nucleotide sequence information may facilitate the distinction between functional and non-functional post-translation modification sites.

Summary

In this review, we discuss in detail the implication of selection on synonymous site to translation properties. An overall view of the effect of codon choice on gene expression is shown in Figure 4. In summary, our understanding of the process of translation has been revolutionized in the genome and systems biology era. Two important characteristics of the process, its efficiency and its fidelity, are now understood much better than just a few years ago. Still, the challenges ahead will be to integrate all of the knowledge and insight that has accumulated from these various studies, and create a consistent model of the translation process that will predict the proteome under various conditions and cell types. Such a model will greatly enhance our understanding of genomes and cellular circuits, will help to elucidate the basis of cell-to-cell variation and will shed light on the molecular basis of diseases.

Current points of debate have to do with the relative role of codon choice and mRNA structure in affecting translation, the relative contribution of control at the level of translation initiation versus elongation, the relative extent of selection for efficiency versus accuracy and the role of random drift versus selection in shaping genes sequence. Even further, translation itself constitutes only one of several steps in the gene expression process, and gene expression as a whole poses only part of the constraints that genes' sequences must obey. The same nucleotide should also support other features such as nucleosome positioning, appropriate splicing (Warnecke *et al*, 2009) and higher order structural elements of the DNA. The apparent redundancy of the genetic code hence facilitates a choice between an astronomical number of coding possibilities of a given amino-acid sequence and may thus facilitate the coordinated satisfaction of many constraints, in addition to translation efficiency, by the same sequence.

Acknowledgements

We thank the European Research Council for an 'ERC Ideas' grant, and the Ben May Foundation for continuous support.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136:** 927–935

Andersson SG, Kurland CG (1990) Codon preferences in free-living microorganisms. *Microbiol Rev* **54:** 198–210

Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D (2003) Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc Natl Acad Sci USA* **100:** 3889–3894

Bahir I, Fromer M, Prat Y, Linial M (2009) Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol Syst Biol* **5:** 311

Bailly-Bechet M, Vergassola M, Rocha E (2007) Causes for the intriguing presence of tRNAs in phages. *Genome Res* **17:** 1486–1495

Barbarese E, Koppel DE, Deutscher MP, Smith CL, Ainger K, Morgan F, Carson JH (1995) Protein translation components are colocalized in granules in oligodendrocytes. *J Cell Sci* **108** (Part 8)**:** 2781–2790

Bennetzen JL, Hall BD (1982) Codon selection in yeast. *J Biol Chem* **257:** 3026–3031

Bucciantini M, Giannoni E, Chiti F, Baroni F, Formigli L, Zurdo J, Taddei N, Ramponi G, Dobson CM, Stefani M (2002) Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* **416:** 507–511

Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129:** 897–907

Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y (2010) A role for codon order in translation dynamics. *Cell* **141:** 355–367

Carbone A (2008) Codon bias is a major factor explaining phage evolution in translationally biased hosts. *J Mol Evol* **66:** 210–223

Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* **7:** 98–108

Comeron JM (2004) Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics* **167:** 1293–1304

Cortazzo P, Cervenansky C, Marin M, Reiss C, Ehrlich R, Deana A (2002) Silent mutations affect *in vivo* protein folding in *Escherichia coli*. *Biochem Biophys Res Commun* **293:** 537–541

Crick FH (1966) Codon—anticodon pairing: the wobble hypothesis. *J Mol Biol* **19:** 548–555

Crombie T, Boyle JP, Coggins JR, Brown AJ (1994) The folding of the bifunctional TRP3 protein in yeast is influenced by a translational pause which lies in a region of structural divergence with *Escherichia coli* indoleglycerol-phosphate synthase. *Eur J Biochem* **226:** 657–664

Crombie T, Swaffield JC, Brown AJ (1992) Protein folding within the cell is influenced by controlled rates of polypeptide elongation. *J Mol Biol* **228:** 7–12

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* **14:** 1188–1190

de Smit MH, van Duin J (1994) Translational initiation on structured messengers. Another role for the Shine-Dalgarno interaction. *J Mol Biol* **235:** 173–184

de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C (2009) Global signatures of protein and mRNA expression levels. *Mol Biosyst* **5:** 1512–1526

Dekel E, Alon U (2005) Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436:** 588–592

Dittmar KA, Goodenbour JM, Pan T (2006) Tissue-specific differences in human transfer RNA expression. *PLoS Genet* **2:** e221

Dittmar KA, Mobley EM, Radek AJ, Pan T (2004) Exploring the regulation of tRNA distribution on the genomic scale. *J Mol Biol* **337:** 31–47

Dong H, Nilsson L, Kurland CG (1996) Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J Mol Biol* **260:** 649–663

dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* **32:** 5036–5044

dos Reis M, Wernisch L (2009) Estimating translational selection in eukaryotic genomes. *Mol Biol Evol* **26:** 451–461

Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134:** 341–352

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

Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and Arabidopsis. *Proc Natl Acad Sci USA* **96:** 4482–4487

Elf J, Nilsson D, Tenson T, Ehrenberg M (2003) Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science* **300:** 1718–1722

Farabaugh PJ, Bjork GR (1999) How translational accuracy influences reading frame maintenance. *EMBO J* **18:** 1427–1434

Gallie DR (1991) The cap and poly(A) tail function synergistically to regulate mRNA translational efficiency. *Genes Dev* **5:** 2108–2116

Gilchrist MA, Shah P, Zaretzki R (2009) Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. *Genetics* **183:** 1493–1505

Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res* **10:** 7055–7074

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res* **9:** r43–r74

Gregersen N (2006) Protein misfolding disorders: pathogenesis and intervention. *J Inherit Metab Dis* **29:** 456–470

Gsponer J, Futschik ME, Teichmann SA, Babu MM (2008) Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science* **322:** 1365–1368

Gu W, Zhou T, Wilke CO (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol* **6:** e1000664

Hamilton R, Watanabe CK, de Boer HA (1987) Compilation and comparison of the sequence context around the AUG startcodons in *Saccharomyces cerevisiae* mRNAs. *Nucleic Acids Res* **15:** 3581–3593

Hartl DL, Taubes CH (1998) Towards a theory of evolutionary adaptation. *Genetica* **102–103:** 525–533

Heger A, Ponting CP (2007) Variable strength of translational selection among 12 *Drosophila* species. *Genetics* **177:** 1337–1348

Hendrickson DG, Hogan DJ, McCullough HL, Myers JW, Herschlag D, Ferrell JE, Brown PO (2009) Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. *PLoS Biol* **7:** e1000238

Hense W, Anderson N, Hutter S, Stephan W, Parsch J, Carlini DB (2010) Experimentally increased codon bias in the *Drosophila* Adh gene leads to an increase in larval, but not adult, alcohol dehydrogenase activity. *Genetics* **184:** 547–555

Huang Y, Koonin EV, Lipman DJ, Przytycka TM (2009) Selection for minimization of translational frameshifting errors as a factor in the evolution of codon usage. *Nucleic Acids Res* **37:** 6799–6810

Ibba M, Soll D (2000) Aminoacyl-tRNA synthesis. *Annu Rev Biochem* **69:** 617–650

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol* **151:** 389–409

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2:** 13–34

Ikemura T, Ozeki H (1983) Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. *Cold Spring Harb Symp Quant Biol* **47** (Part 2)**:** 1087–1097

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324:** 218–223

Jackson RJ, Hellen CU, Pestova TV (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol* **11:** 113–127

Jacob WF, Santer M, Dahlberg AE (1987) A single base change in the Shine-Dalgarno region of 16S rRNA of *Escherichia coli* affects translation of many proteins. *Proc Natl Acad Sci USA* **84:** 4757–4761

Jia M, Li Y (2005) The relationship among gene expression, folding free energy and codon usage bias in *Escherichia coli*. *FEBS Lett* **579:** 5333–5337

Jiang H, Guan W, Pinney D, Wang W, Gu Z (2008) Relaxation of yeast mitochondrial functions after whole-genome duplication. *Genome Res* **18:** 1466–1471

Kahvejian A, Svitkin YV, Sukarieh R, M'Boutchou MN, Sonenberg N (2005) Mammalian poly(A)-binding protein is a eukaryotic translation initiation factor, which acts via multiple mechanisms. *Genes Dev* **19:** 104–113

Kaminska M, Havrylenko S, Decottignies P, Le Marechal P, Negrutskii B, Mirande M (2009) Dynamic organization of aminoacyl-tRNA synthetase complexes in the cytoplasm of human cells. *J Biol Chem* **284:** 13746–13754

Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T (2001) Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J Mol Evol* **53:** 290–298

Kanaya S, Yamada Y, Kudo Y, Ikemura T (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238:** 143–155

Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM (2007) A 'silent' polymorphism in the MDR1 gene changes substrate specificity. *Science* **315:** 525–528

Komar AA, Lesnik T, Reiss C (1999) Synonymous codon substitutions affect ribosome traffic and protein folding during *in vitro* translation. *FEBS Lett* **462:** 387–391

Konigsberg W, Godson GN (1983) Evidence for use of rare codons in the dnaG gene and other regulatory genes of *Escherichia coli*. *Proc Natl Acad Sci USA* **80:** 687–691

Kozak M (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44:** 283–292

Kozak M (2002) Pushing the limits of the scanning mechanism for initiation of translation. *Gene* **299:** 1–34

Kozak M (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361:** 13–37

Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324:** 255–258

Lackner DH, Beilharz TH, Marguerat S, Mata J, Watt S, Schubert F, Preiss T, Bahler J (2007) A network of multiple regulatory layers shapes gene expression in fission yeast. *Mol Cell* **26:** 145–155

Laurent JM, Vogel C, Kwon T, Craig SA, Boutz DR, Huse HK, Nozue K, Walia H, Whiteley M, Ronald PC, Marcotte EM (2010) Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics* **10:** 4209–4212

Lavner Y, Kotlar D (2005) Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene* **345:** 127–138

Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD (2003) A unification of mosaic structures in the human genome. *Hum Mol Genet* **12:** 2411–2415

Lithwick G, Margalit H (2003) Hierarchy of sequence-dependent features associated with prokaryotic translation. *Genome Res* **13:** 2665–2673

Loh PG, Song H (2010) Structural and mechanistic insights into translation termination. *Curr Opin Struct Biol* **20:** 98–103

Lu X, de la Pena L, Barker C, Camphausen K, Tofilon PJ (2006) Radiation-induced changes in gene expression involve recruitment of existing messenger RNAs to and away from polysomes. *Cancer Res* **66:** 1052–1061

Lucks JB, Nelson DR, Kudla GR, Plotkin JB (2008) Genome landscapes and bacteriophage codon usage. *PLoS Comput Biol* **4:** e1000001

Man O, Pilpel Y (2007) Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat Genet* **39:** 415–421

Meyerovich M, Mamou G, Ben-Yehuda S (2010) Visualizing high error levels during gene expression in living bacterial cells. *Proc Natl Acad Sci USA* **107:** 11543–11548

Moriyama EN, Powell JR (1997) Codon usage bias and tRNA abundance in *Drosophila*. *J Mol Evol* **45:** 514–523

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349

Navon S, Pilpel Y (2011) The role of codon selection in regulation of translation efficiency deduced from synthetic libraries. *Genome Biol* **12:** R12

Netzer N, Goodenbour JM, David A, Dittmar KA, Jones RB, Schneider JR, Boone D, Eves EM, Rosner MR, Gibbs JS, Embry A, Dolan B, Das S, Hickman HD, Berglund P, Bennink JR, Yewdell JW, Pan T (2009) Innate immune and chemically triggered oxidative stress modifies translational fidelity. *Nature* **462:** 522–526

Nie L, Wu G, Zhang W (2006) Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in *Desulfovibrio vulgaris*: a quantitative analysis. *Genetics* **174:** 2229–2243

Olsthoorn RC, Zoog S, van Duin J (1995) Coevolution of RNA helix stability and Shine-Dalgarno complementarity in a translational start region. *Mol Microbiol* **15:** 333–339

Pedersen S (1984) *Escherichia coli* ribosomes translate *in vivo* with variable rate. *EMBO J* **3:** 2895–2898

Percudani R, Pavesi A, Ottonello S (1997) Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J Mol Biol* **268:** 322–330

Precup J, Parker J (1987) Missense misreading of asparagine codons as a function of codon identity and context. *J Biol Chem* **262:** 11351–11355

Ran W, Higgs PG (2010) The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. *Mol Biol Evol* **27:** 2129–2140

Robbins-Pianka A, Rice MD, Weir MP (2010) The mRNA landscape at yeast translation initiation sites. *Bioinformatics* **26:** 2651–2655

Rodnina MV, Wintermeyer W (2001) Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanisms. *Annu Rev Biochem* **70:** 415–435

Schauder B, McCarthy JE (1989) The role of bases upstream of the Shine-Dalgarno region and in the coding sequence in the control of gene expression in *Escherichia coli*: translation and stability of mRNAs *in vivo*. *Gene* **78:** 59–72

Schrimpf SP, Weiss M, Reiter L, Ahrens CH, Jovanovic M, Malmstrom J, Brunner E, Mohanty S, Lercher MJ, Hunziker PE, Aebersold R, von Mering C, Hengartner MO (2009) Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol* **7:** e48

Sharp PM, Li WH (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res* **14:** 7737–7749

Sharp PM, Li WH (1987) The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15:** 1281–1295

Sharp PM, Rogers MS, McConnell DJ (1984) Selection pressures on codon usage in the complete genome of bacteriophage T7. *J Mol Evol* **21:** 150–160

Shields DC, Sharp PM, Higgins DG, Wright F (1988) 'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol Biol Evol* **5:** 704–716

Shine J, Dalgarno L (1974) The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci USA* **71:** 1342–1346

Sorensen MA (2001) Charging levels of four tRNA species in *Escherichia coli* Rel(+) and Rel(−) strains during amino acid starvation: a simple model for the effect of ppGpp on translational accuracy. *J Mol Biol* **307:** 785–798

Sorensen MA, Kurland CG, Pedersen S (1989) Codon usage determines translation rate in *Escherichia coli*. *J Mol Biol* **207:** 365–377

Spriggs KA, Bushell M, Willis AE (2010) Translational regulation of gene expression during conditions of cell stress. *Mol Cell* **40:** 228–237

Stansfield I, Jones KM, Herbert P, Lewendon A, Shaw WV, Tuite MF (1998) Missense translation errors in *Saccharomyces cerevisiae*. *J Mol Biol* **282:** 13–24

Stefani M, Dobson CM (2003) Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J Mol Med* **81:** 678–699

Steitz JA, Jakes K (1975) How ribosomes select initiator regions in mRNA: base pair formation between the 3′ terminus of 16S rRNA and the mRNA during initiation of protein synthesis in *Escherichia coli*. *Proc Natl Acad Sci USA* **72:** 4734–4738

Stenico M, Lloyd AT, Sharp PM (1994) Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Res* **22:** 2437–2446

Stoebel DM, Dean AM, Dykhuizen DE (2008) The cost of expression of *Escherichia coli* lac operon proteins is in the process, not in the products. *Genetics* **178:** 1653–1660

Stoletzki N, Eyre-Walker A (2007) Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol Biol Evol* **24:** 374–381

Takagi M, Absalon MJ, McLure KG, Kastan MB (2005) Regulation of p53 translation and induction after DNA damage by ribosomal protein L26 and nucleolin. *Cell* **123:** 49–63

Tarun Jr SZ, Sachs AB (1996) Association of the yeast poly(A) tail binding protein with translation initiation factor eIF-4G. *EMBO J* **15:** 7168–7177

Thanaraj TA, Argos P (1996a) Protein secondary structural types are differentially coded on messenger RNA. *Protein Sci* **5:** 1973–1983

Thanaraj TA, Argos P (1996b) Ribosome-mediated translational pause and protein domain organization. *Protein Sci* **5:** 1594–1612

Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS (2007) The stability effects of protein mutations appear to be universally distributed. *J Mol Biol* **369:** 1318–1332

Tsai CJ, Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM, Nussinov R (2008) Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *J Mol Biol* **383:** 281–291

Tsvetkov P, Asher G, Paz A, Reuven N, Sussman JL, Silman I, Shaul Y (2008) Operational definition of intrinsically unstructured protein sequences based on susceptibility to the 20S proteasome. *Proteins* **70:** 1357–1366

Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y (2010a) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141:** 344–354

Tuller T, Waldman YY, Kupiec M, Ruppin E (2010b) Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci USA* **107:** 3645–3650

Uversky VN, Gillespie JR, Fink AL (2000) Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins* **41:** 415–427

Varenne S, Buc J, Lloubes R, Lazdunski C (1984) Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J Mol Biol* **180:** 549–576

Vogel C, Abreu Rde S, Ko D, Le SY, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte EM, Penalva LO (2010) Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol* **6:** 400

Wang L, Wessler SR (2001) Role of mRNA secondary structure in translational repression of the maize transcriptional activator Lc(1,2). *Plant Physiol* **125:** 1380–1387

Warnecke T, Hurst LD (2010) GroEL dependency affects codon usage—support for a critical role of misfolding in gene evolution. *Mol Syst Biol* **6:** 340

Warnecke T, Weber CC, Hurst LD (2009) Why there is more to protein evolution than protein function: splicing, nucleosomes and dual-coding sequence. *Biochem Soc Trans* **37:** 756–761

Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, Minshull J, Gustafsson C (2009) Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS ONE* **4:** e7002

Wen JD, Lancaster L, Hodges C, Zeri AC, Yoshimura SH, Noller HF, Bustamante C, Tinoco I (2008) Following translation by single ribosomes one codon at a time. *Nature* **452:** 598–603

Wilcox C, Hu JS, Olson EN (1987) Acylation of proteins with myristic acid occurs cotranslationally. *Science* **238:** 1275–1278

Wright F (1990) The 'effective number of codons' used in a gene. *Gene* **87:** 23–29

Wu G, Nie L, Zhang W (2008) Integrative analyses of post-transcriptional regulation in the yeast *Saccharomyces cerevisiae* using transcriptomic and proteomic data. *Curr Microbiol* **57:** 18–22

Yang JR, Zhuang SM, Zhang J (2010) Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol Syst Biol* **6:** 421

Yanofsky C (1981) Attenuation in the control of expression of bacterial operons. *Nature* **289:** 751–758

Zaborske JM, Narasimhan J, Jiang L, Wek SA, Dittmar KA, Freimoser F, Pan T, Wek RC (2009) Genome-wide analysis of tRNA charging and activation of the eIF2 kinase Gcn2p. *J Biol Chem* **284:** 25254–25267

Zhang F, Saha S, Shabalina SA, Kashina A (2010) Differential arginylation of actin isoforms is regulated by coding sequence-dependent degradation. *Science* **329:** 1534–1537

Zhang G, Hubalewska M, Ignatova Z (2009) Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat Struct Mol Biol* **16:** 274–280

Zhang SP, Zubay G, Goldman E (1991) Low-usage codons in *Escherichia coli*, yeast, fruit fly and primates. *Gene* **105:** 61–72

Zhou T, Weems M, Wilke CO (2009) Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol* **26:** 1571–1580

4. Materials and Methods

4.1 Data sources

Expression profiling of human tRNAs and mRNAs in different cancerous cell types and physiological conditions

Custom-made arrays (Nimblegen) were supplied by our collaborator, Andres Lund form Biotech Research and Innovation Centre (BRIC) at the University of Copenhagen. The microarrays contain probes for 7000 protein-coding transcripts and 155 probes correspond to 206 tRNA genes. The various cell types from which RNA was hybridized onto the array are detailed in table 1.

tRNA gene copy number

The tRNA gene copy numbers of all analyzed species were downloaded from the Genomic tRNA Database (http://lowelab.ucsc.edu/GtRNAdb/), (Lowe and Eddy 1997).

Coding sequences

The coding sequences of *H. sapiens* and *M. musculus* were downloaded from the Consensus CDS (CCDS) project (ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/). The coding sequences of *C. elegans* were downloaded from Ensembl ftp site (http://www.ensembl.org) (WS210, release 59). The coding sequences of *D. melanogaster* were downloaded from FlyBase (http://flybase.org/) on November 2010. The coding sequences of *S. cerevisiae* were downloaded from SGD (Saccharomyces Genome Database, http://www.yeastgenome.org) on May 2008. The coding sequences of *S. pombe* and *Y. lipolitica* were downloaded from EMBL database (http://www.ebi.ac.uk/) on Oct 2009.

Classification of gene categories

Defined gene categories by biological process and cellular component were downloaded from the Gene Ontology project (http://www.geneontology.org/).

Cell Types	# of samples	# of replicates	Control cells
Primary Lymphoma	69		
Lymphoma cell line (HT)	1		
Reactive lymphnodes	10		
Normal B-cells	10		
Bladder cancer	83		
Bladder cell lines: 253JBV; 575A; CRL2169; HCV29; HT1197; HT1376; HU609; J82; RT4(USA); RT4(dk); SLT4; SW780; T24; UMUC14; UMUC9.	15		
Normal Bladder cells	8		
Colon carcinoma	44		
Colon cell lines: Colo205; DLD1; DLD1 TR7; HCT115; HCT116; HT29; LS174T; LS174T TR4; SW480; SW620.	10		
Colon adenoma	39		
Normal Colon cells (mucosa adjacent)	16		
Prostate cancer	28		
Prostate cell lines: BPH1-1; DuCaP-1; PNT1A-1; PSK1-1; VCaP.	5		
Prostate, adjacent to malignant cells	15		
Normal Prostate cells	11		
hESCs (human embryonic stem cells)		3	
hESCs - 1 day after using retinoic acid as differentiation-inducing agent		3	
hESCs - 3 days after using retinoic acid as differentiation-inducing agent		3	
hESCs - 5 days after using retinoic acid as differentiation-inducing agent		3	
Human fibroblasts (BJ/hTERT) over-expressing cMyc for 24 hours		3	3 replicates of cells transduced with control virus (pBabe)
Human fibroblasts (BJ/hTERT) over-expressing cMyc for 72 hours		3	
Human fibroblasts (BJ/hTERT) over-expressing RASV12 for 24 hours		3	
Human fibroblasts (BJ/hTERT) over-expressing RASV12 for 72 hours		3	
Human fibroblasts (BJ/hTERT) that have been serum-starved for 70 hours		3	3 replicates of assynchronized cells
Starved cells 30 minutes after re-addition of serum		3	
Starved cells 2 hours after re-addition of serum		3	
Starved cells 4 hours after re-addition of serum		3	

Table 1: **Cell types for which expression profiling of human tRNAs and mRNAs were measured.**

Chromatin modification

Density graphs of the H3K27ac modification were plotted using the Broad Histone (wgEncodeBroadHistone) Track at UCSC website (http://genome.ucsc.edu/). Specifically, this track displays maps of chromatin state generated by the Broad/MGH ENCODE group using ChIP-seq. In this track, densities are calculated as the number of sequenced tags overlapping a 25 bp window centered at that position, and are the results of pooled replicates.

4.2 Calculation of the variation in the human tRNA pool

For each tRNA type (i.e, anticodon) in a given sample we summed the expression of its corresponding individual genes. Then, for each sample, we divided the expression of each tRNA type by its averaged expression in either normal cells of the same tissue (for primary tumors and cancerous cell lines), or in the corresponding control (for cells triggered by various treatments) – see table 1. Finally, for a given cell types, we averaged the fold-changes in the tRNAs expression across its corresponding samples – see details in table 1.

4.3 Principal component analysis

Principal component analysis was performed using the MATLAB Statistics Toolbox.

4.4 Calculating translational efficiency by the tAI value

We calculated translation efficiencies of genes using the tRNA adaptation index (tAI) (dos Reis et al. 2004). Throughout this paper, we distinguish between translational efficiency of a gene, which corresponds to the original tAI measure, and translational efficiency of individual codons (originally defined by dos Reis et al. as the “absolute adaptiveness value”, W_i). Briefly, W_i defines the adaptiveness of an individual codon by the availability of the tRNAs that serve in translating it, incorporating both the fully-matched tRNA, as well as tRNAs that contribute to

translation through wobble rules (Crick 1966). Formally, the “absolute adaptiveness value” for the *i*–th codon is

$$W_i = \sum_{j=1}^{n_i} \left(1 - s_{ij}\right) tGCN_{ij}$$ (dos Reis et al. 2004)

where *n* is the number of tRNA isoacceptors that recognize the *i*-th codon, $tGCN_{ij}$ denotes the gene copy number of the *j*-th tRNA that recognizes the *i*-th codon, and s_{ij} correspond to the wobble interaction, or selective constraint on the efficiency of the pairing between codon *i* and anticodon *j*. As done in the original tAI formalism by dos Reis et al. the absolute adaptiveness value of codon *i* is further divided by the maximum W_i (termed W_{max}), obtaining the codon's relative adaptiveness value:

$$w_i = W_i / W_{\max}$$

The tAI value of a gene with L codons is then simply calculated as the geometric mean of the w_i's of its codons

$$tAI(g) = \sqrt[L]{\prod_{c=1}^{L} w_c}$$

Based on (Tuller et al. 2010) we interpret the codon adaptiveness values as representatives of translation speed of each codon.

5. Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome

(A paper describing this work was published in *Nucleic Acids Res*)

Abstract

Translation of a gene is assumed to be efficient if the supply of the tRNAs that translate it is high. Yet high-abundance tRNAs are often also at high demand since they correspond to preferred codons in genomes. Thus to fully model translational efficiency one must gauge the supply-to-demand ratio of the tRNAs that are required by the transcriptome at a given time. The tRNAs' supply is often approximated by their gene copy number in the genome. Yet neither the demand for each tRNA nor the extent to which its concentration changes across environmental conditions has been extensively examined. Here we compute changes in the codon usage of the transcriptome across different conditions in several organisms by inspecting conventional mRNA expression data. We find recurring dynamics of codon usage in the transcriptome in multiple stressful conditions. In particular, codons that are translated by rare tRNAs become over-represented in the transcriptome in response to stresses. These results raise the possibility that the tRNA pool might dynamically change upon stress to support efficient translation of stress-transcribed genes. Alternatively, stress genes may be typically translated with low efficiency, presumably due to lack of sufficient evolutionary optimization pressure on their codon usage.

Nucleic Acids Research Advance Access published August 31, 2012

Nucleic Acids Research, 2012, 1–11
doi:10.1093/nar/gks772

Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome

Hila Gingold, Orna Dahan and Yitzhak Pilpel*

Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel

Received March 19, 2012; Revised and Accepted July 23, 2012

ABSTRACT

Translation of a gene is assumed to be efficient if the supply of the tRNAs that translate it is high. Yet high-abundance tRNAs are often also at high demand since they correspond to preferred codons in genomes. Thus to fully model translational efficiency one must gauge the supply-to-demand ratio of the tRNAs that are required by the transcriptome at a given time. The tRNAs' supply is often approximated by their gene copy number in the genome. Yet neither the demand for each tRNA nor the extent to which its concentration changes across environmental conditions has been extensively examined. Here we compute changes in the codon usage of the transcriptome across different conditions in several organisms by inspecting conventional mRNA expression data. We find recurring dynamics of codon usage in the transcriptome in multiple stressful conditions. In particular, codons that are translated by rare tRNAs become over-represented in the transcriptome in response to stresses. These results raise the possibility that the tRNA pool might dynamically change upon stress to support efficient translation of stress-transcribed genes. Alternatively, stress genes may be typically translated with low efficiency, presumably due to lack of sufficient evolutionary optimization pressure on their codon usage.

INTRODUCTION

Organisms have evolved means to tune the translational efficiency of their genes to different desired levels. Facilitating this mode of regulation is the redundancy of the genetic code—synonymous codons are translated to the same amino acid, but their corresponding tRNAs might differ by their amounts in cells. Common measures of translational efficiency assess genes by either measuring the correlation between their codon usage pattern to that of selected 'elite' highly expressed genes (1,2), or by an explicit weight of the availability of tRNAs that translate them (3). One Such measure is the tRNA adaptation index, tAI (4), which deduces the abundance of the various tRNAs from their gene copy number (GCN) in the genome. The tAI measure predicts with reasonable accuracy both mRNA and protein levels (5,6).

Yet, recent studies suggest that models of translational efficiency should be more comprehensive [reviewed in (7); (8,9)]. In particular, the concentration of the various tRNAs can vary to different extents between conditions and tissues (9–13) and so is their base modification and amino acid loading (14–17). In addition codon usage was shown to vary between tissues (18). Toward a comprehensive model of translational efficiency we focus here on an unexplored important consideration—the supply-to-demand ratio of the different tRNAs that translate mRNAs.

Specifically, we explore the potential changes in the demand for each tRNA, namely the abundance of the corresponding codon(s) in the transcriptome at various biological conditions. We suggest a simple means to mine mRNA expression data in order to explore the dynamic codon usage across conditions and species. We find that under stress conditions the demand for tRNAs by certain codons increases, whereas the extent of representation of other codons in the transcriptome decreases. Interestingly the codons whose representation increases the most in the transcriptome upon stress correspond to tRNAs that are represented by the lowest GCNs in the genome in each of the examined species. This situation suggests two possible explanations: the supply for these codons increases too in stress, e.g. by increased production of the corresponding tRNAs, or that stress-related genes remain relatively poorly translated. We constructed a simulated evolution model that uses two minimalist assumptions: that stress genes are expressed infrequently during evolution, and that the supply of tRNAs in the cell is limited. Results from the simulation are consistent

*To whom correspondence should be addressed. Tel: +972 8 934 6058; Fax: +972 8 934 4108; Email: pilpel@weizmann.ac.il

with the hypothesis that limited codon adaptiveness of the stress genes results from lack of evolutionary constraint to optimize them and due to the limitation in the supply of tRNAs. Reassuringly the simulation reproduces codon usage differences observed in genomes. We thus suggest that the stress transcriptome remains poorly translated due to compromised translational efficiency of the stress genes.

MATERIALS AND METHODS

Data sources

The affymetrix platform of micro-arrays was used to allow absolute, i.e. non-relative, measurements of mRNA abundance. Normalized mRNA abundances for *Saccharomyces cerevisiae* following oxidative stress and DNA damage by methyl methanesulfonate (MMS) were downloaded from (19) and in heat shock, oxidative stress and osmotic stress from (20). Normalized mRNA abundances of *S. cerevisiae* during a 15 days process of wine fermentation (21) were downloaded from GEO (Gene Expression Omnibus, www.ncbi.nlm.nih.gov/geo), GEO accession: GSE8536. Normalized mRNA abundances of *Schizosaccharomyces pombe* during a short-term response to nitrogen starvation (22) were downloaded from ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/) under accession E-TABM-784. Transcriptome profiling data for *Caenorhabditis elegans* in response to oxidative stress (23) were downloaded from GEO under accession number GSE9301. The tRNA GCNs of species were downloaded from the Genomic tRNA Database (http://lowelab.ucsc.edu/GtRNAdb/), (24).

Generation of condition-dependent demand matrices

We created condition-dependent 'Codon-Expression' matrices for various environmental conditions. A 'Codon-Expression' matrix, is a $61 \times m$ matrix, which denotes the representation of the 61 codons in the transcriptome over m conditions. This matrix thus depicts the demand for each tRNA at each condition. The representation of codon i in the transcriptome at a given time/condition k is defined by

$$r_{ik} = \sum_{j=1}^{n} C_{ij} E_{jk}$$

where j is a gene, C_{ij} depicts the number of appearances of codon i in gene j, and E_{jk} indicates the mRNA abundance of gene j at condition or time point k. In the same manner, we generated 'Amino acid Expression' and 'Nucleotide Expression' matrices, which similarly depict the representation of the 20 amino acid in the translated transcriptome or the 4 nucleotides in the transcriptome at a given time/condition k.

Calculating translational efficiency by the tAI value

We calculated translation efficiencies of genes using the tAI (4). Throughout this article, we distinguish between translational efficiency of a gene, which corresponds to the original tAI measure, and translational efficiency of individual codons (originally defined by dos Reis *et al.* as the 'absolute adaptiveness value', W_i). Briefly, W_i defines the adaptiveness of an individual codon by the availability of the tRNAs that serve in translating it, incorporating both the fully matched tRNA, as well as tRNAs that contribute to translation through wobble rules (25–27). Formally, the 'absolute adaptiveness value' for the i-th codon is

$$W_i = \sum_{j=1}^{n_i} \left(1 - s_{ij}\right) tGCN_{ij} \text{ (ref. 4)}$$

where n is the number of tRNA isoacceptors that recognize the i-th codon, $tGCN_{ij}$ denotes the GCN of the j-th tRNA that recognizes the i-th codon, and s_{ij} correspond to the wobble interaction, or selective constraint on the efficiency of the pairing between codon i and anticodon j. As done in the original tAI formalism by dos Reis *et al.* the absolute adaptiveness value of codon i is further divided by the maximum W_i (termed $W_{\max}$), obtaining the codon's relative adaptiveness value:

$$w_i = W_i / W_{\max}$$

The tAI value of a gene with L codons is then simply calculated as the geometric mean of the w_i's of its codons

$$\text{tAI}(g) = \sqrt[L]{\prod_{c=1}^{L} w_c}$$

Based on (9) and (28) we interpret the codon adaptiveness values as representatives of translation speed of each codon.

Exploring the balance between drift and selection on codon usage by a computational simulation

We developed a computer evolutionary simulation of unicellular population of 1 000 000 haploid cells that cope with occasional stress periods. The genome of each cell consists of six archetypal genes, each of which is required during one or more of the simulated growth conditions (see Supplementary Material for details). We envisage a mapping function, such as the tAI score, that maps between sequence and expression level. During the simulation genes are mutated and as a consequence their expression level changes. At the beginning of the simulation, the six genes are equally scored with an initial value of expression level. The population then evolves at a fixed mutation rate of 0.001 mutations per genome per generation. Sequences are not represented explicitly in the simulation; instead genes are characterized by an expression level that implicitly corresponds to a genotype. Thus, 'mutated' expression levels at a given time step are computed by the previous step's expression levels multiplied by a random number drawn from an exponential probability distribution of changes in expression [as estimated before (29)].

We also ran the simulation with a mode in which the tRNA supply is not unlimited. This mode of the run sets a

constant maximal total expression level from the 'genome'. In this mode of limited supply of tRNAs, the expression of the i-th gene in each generation, lsExpression_{gi}, ('ls' stands for limited supply) is defined by

$$\text{lsExpression}_{gi} = \begin{cases} \text{Expression}_{gi} * \left(\frac{\text{MaxExpression}}{\sum_{i=1}^{n} \text{Expression}_{gi}}\right) & \text{if} \left(\sum_{i=1}^{n} \text{Expression}_{gi} > \text{MaxExpression} \right) \\ \text{Expression}_{gi} & \text{else} \end{cases}$$

where n is the number of genes in the cell, Expression_{gi} is the expression of the i-th gene in the mutated population, and MaxExpression is a constant maximal total expression level from the whole 'genome'.

The fitness of a given cell is determined by the arithmetic mean of expression of the m genes which are required in a given environmental condition

$$\lambda = \frac{\sum_{j=1}^{m} \text{Expr}_j}{m}$$

where Expr_j corresponds to either Expression_{gi} or lsExpression_{gi}, depending on the mode of simulation, i.e. if supply is limited or not.

Our model consists of drift and selection, thus we propagate individuals between consecutive generations $(t-1)$ to (t) in two stages. First, a population (whose size can be different from that of the population at generation $t-1$) is formed in which the i-th genotype population size is given by its size in the previous generation and its fitness by

$$x_{i(t)} \approx x_{i(t-1)} e^{\lambda_i}$$

Then, to keep a constant population size stochastic rescaling is applied that implements a Kimura-governed (30) allele sampling as the random drift step of the simulation.

RESULTS

The supply-to-demand balance in translational efficiency

The speed at which a codon is translated is expected to increase with the availability of its supply—amino acid-loaded tRNAs. Yet, if the codon is highly represented in the genome, and even more importantly, in transcriptome at a given condition, i.e. if the demand for the tRNA is high, then the codon's translational efficiency might be compromised. Translational efficiency should thus be modeled as a supply-to-demand process, i.e. the ratio between the availability of the tRNAs that translate a codon—the 'supply', to the extent of representation of that codon in the transcriptome at a given moment—the 'demand'. The supply component is effectively captured by the tAI of a codon (4), originally defined as W_i and termed codon 'adaptiveness value' (see 'Materials and Methods' section). The demand component is simply modeled as the representation of the i-th codon in the transcriptome, r_{ik}(see 'Materials and Methods' section).

Translational efficiency of the i-th codon is the supply-to-demand ratio:

$$W_i^{\text{SD}} = \frac{\text{Supply}_i}{\text{Demand}_i} = \frac{\text{codonAdaptivenessValue}_i}{\text{representation}_i} = \frac{W_i}{r_{ik}}$$

In the following sections we will focus on measuring the dynamics of codon usage across conditions in different species. Note that although in the original tAI model the supply is constant (modeled by copy number of the tRNA genes in the genome) it could in future be represented as a dynamic entity too, accounting for the possibility that tRNA levels may change across conditions and cell types (9–12,14–17).

mRNA expression data can be mined to deduce dynamics of cellular codon usage

Conventional mRNA expression micro-arrays are typically used to study transcription and more recently mRNA decay (19,31–37). Here we realized that micro-arrays may contain data pertinent to translational efficiency. In particular we mine expression micro-array data to compute changes in the representation of the various codons in the transcriptome in various growth conditions in several model organisms. Figure 1 shows as an example the change in usage of the six codons coding for arginine during response to a DNA damaging agent in the yeast *S. cerevisiae* (see 'Materials and Methods' section). The representation of some of the codons increases by as much as 25% in stress, whereas others decrease in the stressed transcriptome. These changes in codon representation might indicate a change in the demand for the various tRNAs, and potentially also a change in translational efficiency of some genes in stress. Interestingly the summed usage of all six arginine codons hardly changes, and as so is the usage of the four nucleotides.

We systemize our inspection of fluctuations of codons representation in the transcriptome. By multiplying the number of occurrences of each codon in each gene by the expression level of each gene in each condition we obtain a 'Codon-Expression' matrix which depicts the representation of each codon in the transcriptome in each condition (see 'Materials and Methods' section).

Similarly, for further control purposes, we create 'Amino acid Expression' and 'Nucleotide-Expression' matrices, which correspond, respectively, to the sum of codons for each amino acid or the representation of each nucleotide in the transcriptome at various conditions in a given species. The amino acid expression matrix allows us to ask whether changes at the codon expression matrix simply reflect changes in the relative appearance of the different amino acids at the translated transcriptome [as shown in some cases, c.f. (38)] whereas the nucleotide

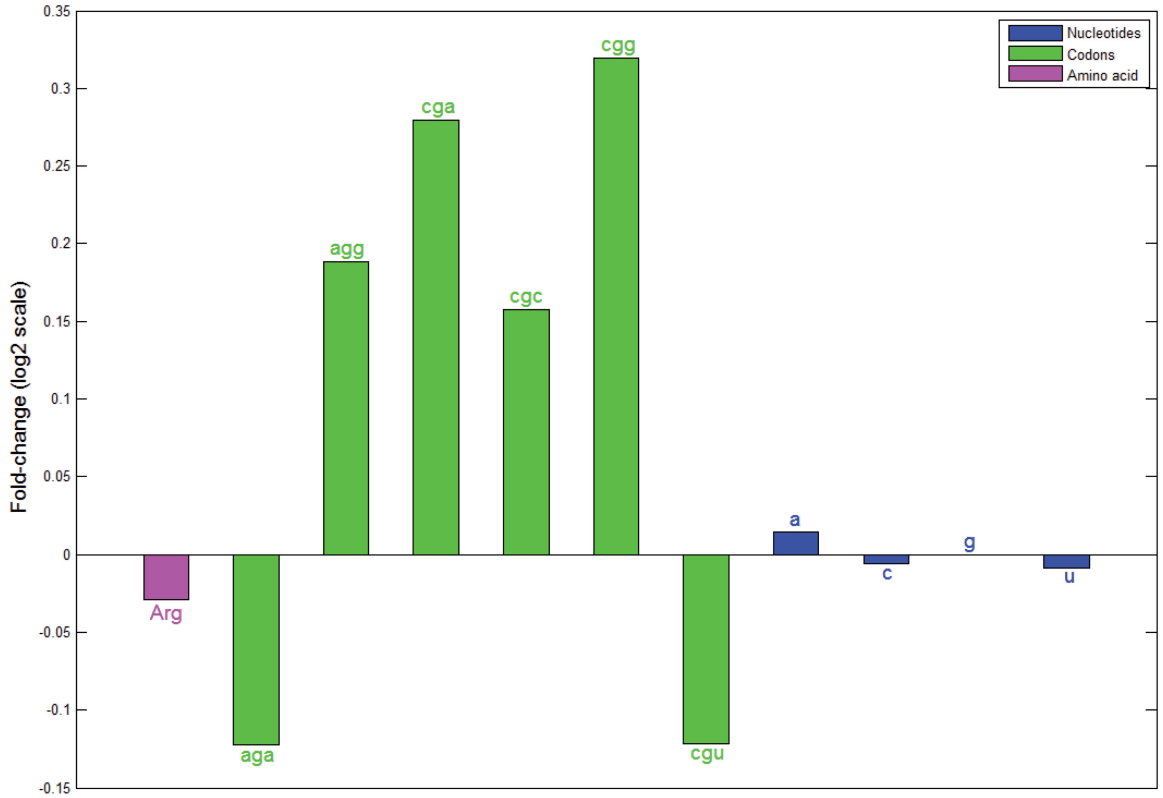


Figure 1. Variations in representation of amino acids, codons and their constituent nucleotides in the transcriptome. Illustration of the case of amino acid arginine: shown on the *y*-axis is the ($\log_2$) change in the representation of the six codons of this amino acid in the transcriptome during response to the chemical MMS. Four of the codons are increased in their representation following the stress compared with a reference condition, and two are diminished. This change is not accompanied by an appreciable change in the representation of arginine in the translated transcriptome, nor can it be reduced to a putative change in the amount of the constituent nucleotides which are shown to hardly vary during the process.

expression matrix weighs the putative changes in the nucleotide composition of the transcriptome (39).

The codon usage of the transcriptome varies upon stress

We analyzed mRNA expression data for the yeast *S. cerevisiae* in diversity of stressful environmental conditions, and also during recovery from stress (19–21). For each time point in each condition we computed the 'Codon-Expression' matrix (Figure 2). Qualitatively, the pattern of variation in codon usage is highly correlated between the various stresses and anti-correlated between the stresses and the stress-recovery experiments. The codons that are changed the most in stress show an increased representation of up to 40% relative to the reference condition (Figure 2). In contrast, changes in amino acid representation are very minor (Supplementary Figure S1), and it is often the case that codons for the same amino acid change in opposite directions. The representation of the four nucleotides is also relatively constant throughout conditions—fold-change values vary between 0.99–1.01 and 0.98–1.03 for codon position-independent and codon position-dependent usage of nucleotides, respectively (Supplementary Figure S1). Thus codon representation changes are not explained by a need to change the amino acid composition of the proteome and cannot be reduced to potential change in nucleotide availability. In turn this dynamics likely reflects changes in translational efficiency.

Low-efficiency codons are over-represented in the stress transcriptome

We were next interested to check whether changes in codon representation in stress, or during recovery from stress, are correlated with the abundance of the corresponding tRNAs. A common simplified proxy for tRNA availability is the copy number of the tRNA genes in the genome. tRNA GCNs correlate with tRNA concentrations, at least in non-stressful conditions (9,11,40,41). This correlation was recently corroborated in a study that examined in several mammals RNA Pol III occupancy in the vicinity of tRNA genes (13). We thus plotted for each codon its representation in the transcriptome at various conditions along with the GCN of the corresponding tRNA, and in addition a more refined measure of tRNA availability (called codon adaptiveness value, W_i, see 'Materials and Methods' section) that also incorporates contributions to the translation of a codon through wobble interaction (25) from non-perfectly matching tRNAs (Figure 2). Interestingly the codon change pattern during recovery from stress correlates positively with these two measures of tRNA availability, while the changes during stress showed negative correlation with the tRNA availability (Figure 3a, Supplementary

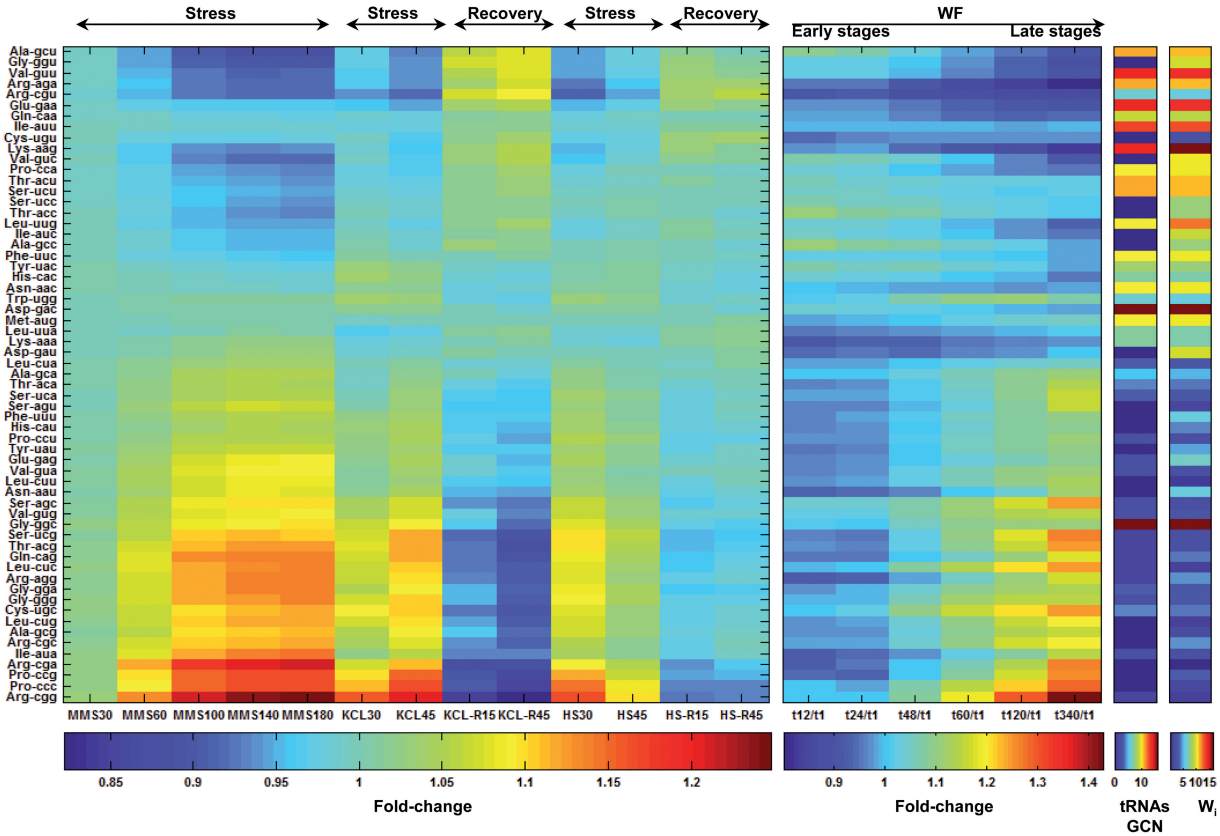


Figure 2. The fold-changes in codons representation in *S. cerevisiae*'s transcriptome across environmental changes. Left panel—fold-changes in the representation of the 61 codons in response to heat shock, osmotic and MMS stresses, and when potassium chloride (KCL) and heat-shock stresses were removed (time is denoted in minutes). Second panel from left—fold-changes in representation of the 61 codons in yeast's transcriptome throughout a 15-day wine fermentation. Columns marked with 'stress' denote the fold-changes at a given stress compared with time point zero, before the stress was applied, whereas 'recovery' marks columns which show the fold-changes after the stress, as cells were transferred to non-stressful conditions, compared with the stress conditions (refers to time point 45 of the respective stress). Columns marked with 'WF' represent the various time points (in hours) during the wine fermentation process. The matrices were normalized by dividing the fold-change values of individual codons at each time point to the total fold-change values across all codons. In addition the right most two columns depict for each codon the gene copy number of its fully matched tRNAs (tRNA's GCN), as well as its codon adaptiveness value. The codons in the matrices are sorted according to their fold-change increase in the stresses. As seen, codons that are increased in representation during either stress or fermentation have low tRNA GCN and low adaptiveness values.

Figure S2 and legends for numerical details). This result means that during stress the transcriptome's codon usage shifts from codons that correspond to the tRNAs that are represented by high gene counts toward codons that correspond to tRNAs that are typically encoded by few gene copies.

Stressful conditions induce a similar pattern of changes in codon usage in additional species

We next wanted to check whether the tendency to increase the representation of codons that correspond to low gene copy tRNAs in stressful conditions is shared in other species too. We analyzed the fission yeast *S. pombe* during a short-term response to nitrogen starvation (22) and the worm *C. elegans* in response to oxidative stress (23). The results clearly show a consistent trend—increased representation in the transcriptome upon stress of codons that correspond to tRNAs whose GCN is low—the Pearson correlation between the change in codon representation in the transcriptome in stress and the availability of the corresponding tRNA is −0.59 (P-value = 5.52×10^{-7}) and −0.8 (P-value = 1.02×10^{-14}), for *C. elegans* and *S. pombe*, respectively (Figure 3b).

In some cases a given codon may have high tRNA availability in one species and a low availability in another. We found out that such codons show increased representation in the stressed transcriptome only in the species where their corresponding tRNAs are at low level. For instance, codon GGA (Gly) has the highest W_i (codon adaptiveness value) in *C. elegans*, and relatively low value in *S. cerevisiae*, and reassuringly it shows decreased representation in the transcriptome of worm upon oxidative stress, whereas in yeast it is among the ten most elevated codons in stress. Thus the similarity of our trend across species does not simply result from a similar behavior of the same codon across species, but may rather reflect a commonality in which the codons that correspond to the rare tRNAs in each species respond similarly.

Distinct stress-specific genes induce a similar signature of changes in codon usage in various stresses

Our results in *S. cerevisiae* show a similar signature of codon usage change in stress across multiple stress types. This similarity could simply arise from a common set of general-stress genes that respond similarly to all stresses (32). In contrast we wanted to examine the possibility that distinct genes sets, that are each specific to each of the stresses, may also present the same trend. To examine this possibility we created distinct gene sets, each shows induction or repression of at least 2-fold in only one of the stress conditions. In addition we created a 'general stress gene set' that consists of genes that were either induced or repressed by at least 2-fold in all examined stresses. We characterized each of these gene sets by the change in the representation of each of the 61 codons in the various stress types (considering both the codon usage and the mRNAs levels see Materials and Methods for details). Then, we compared between the variations in codon representation of the gene sets across environmental changes (Figure 4).

Interestingly we found that all stress-specific gene sets show highly correlated pattern of change, and that each of them correlated negatively with tRNA availability in terms of codons' tAI values (Pearson correlations between −0.62 and −0.73; *P*-values << 0.05). In contrast, the general stress genes have a different signature that does not correlate significantly with tRNA availability (Pearson correlations between −0.22 and 0.13). This analysis indicates that in each stress type distinct genes converge upon the same pattern of increased representation of codons that correspond to low GCN tRNAs. On the other hand, the general stress genes are actually better adapted to the high gene copy tRNAs.

Why are not all genes codon-optimized? A potential effect of genetic drift and a limited tRNA supply

An intriguing question is why did not evolution optimize the codon usage of all genes in a genome? We hypothesize that at least three factors may account for this situation: (i) Drift versus selection balance: drift

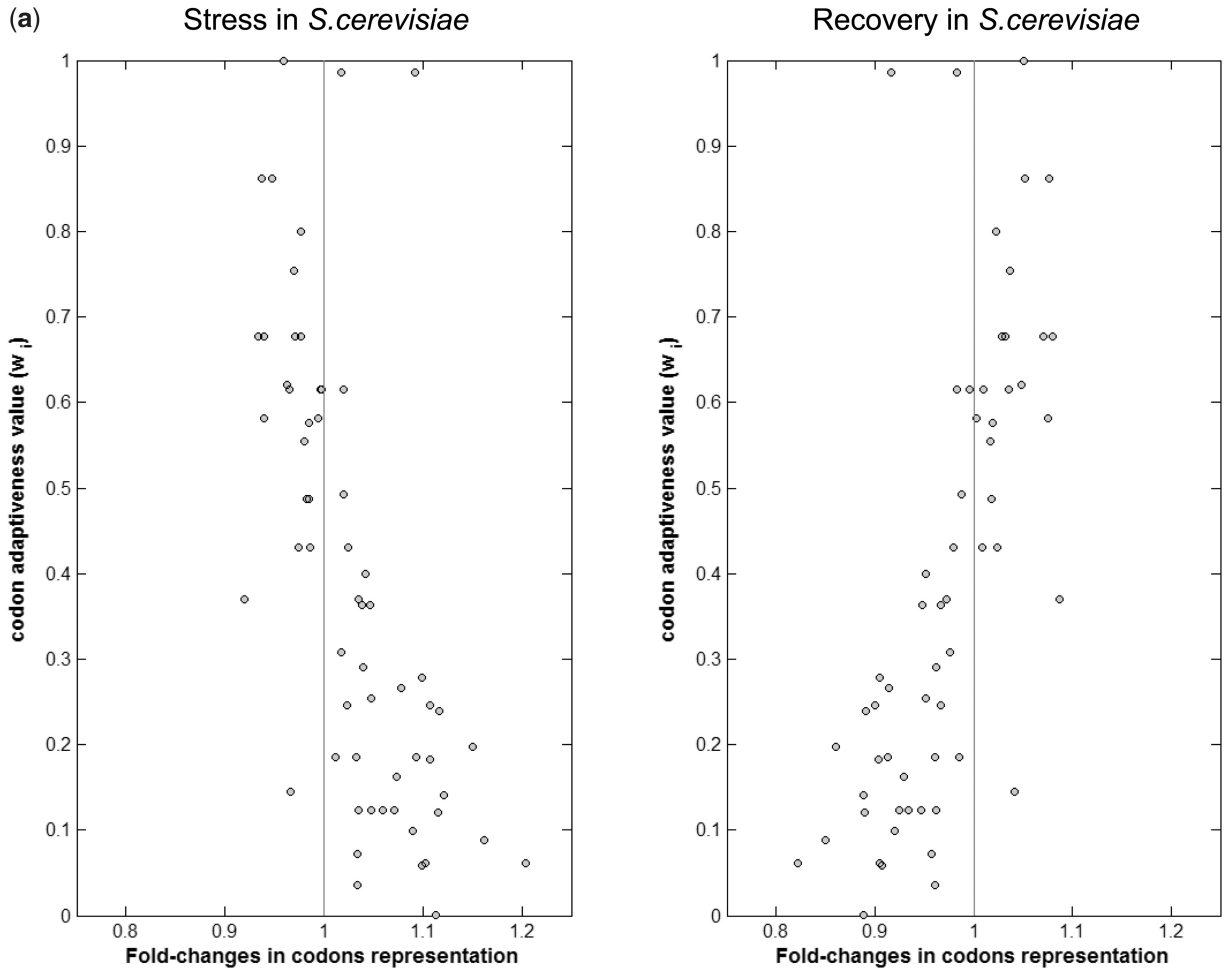


Figure 3. Correlation between variation in codons' usage and their translational efficiency across environmental changes. Each panel denotes the correlation between the adaptiveness values (W_i) of the 61 codons and the fold-changes in their representation in the transcriptome (**a**) Left panel—response to osmotic stress in *S. cerevisiae*; Pearson correlation = −0.66, *P*-value = 5.37×10^{-9}. Right panel—recovery from osmotic stress; Pearson correlation = 0.69, *P*-value = 1.01×10^{-9}. (**b**) Left panel—a short-term response to nitrogen starvation in *S. pombe*. Pearson correlation = −0.8, *P*-values = 1.02×10^{-14}; Right panel—response to oxidative stress in *C. elegans*, Pearson correlation = −0.59, *P*-value = 5.52×10^{-7}.

(continued)

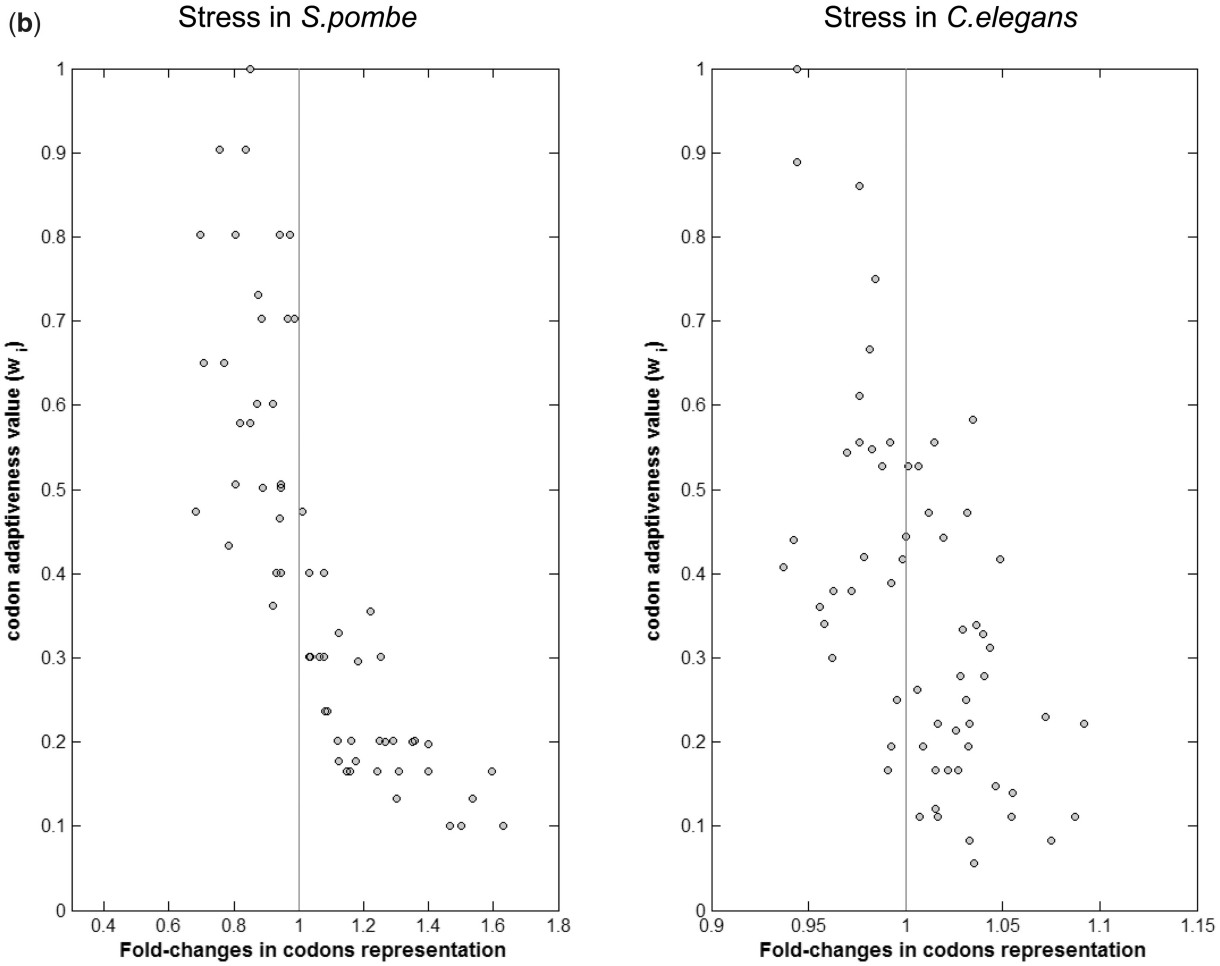


Figure 3. Continued.

may erode codon optimization of genes, thus counteracting the force of purifying selection and codon optimization in particular. Intuitively, while drift should constantly act on all genes at all times, selection should mainly act on a gene during evolutionary time periods and conditions in which it is needed. Thus genes that are expressed rarely during the life-history would experience compromised selection-to-drift ratio. (ii) Limitation in the supply: if the pool of amino acid-loaded tRNAs is not in great excess (see Discussion) and all genes were biased toward codons that correspond to abundant tRNAs then the demand for such tRNAs might be too high and translation would not be efficient even in genes in which high translational efficiency is most needed and thus expressed. Hence some genes may need to 'give way' to others. (iii) In addition, it is entirely possible that some genes may actually be needed in low level of expression and are thus deliberately not codon-optimized (42,43).

Focusing on testing the first two scenarios we constructed an evolutionary dynamics computer simulation to test the effect of drift-to-selection ratio and of demand-to-supply ratio on codon optimization levels of genes. Our simulated model consists of a constant-size population of 1 000 000 cells, each possess a genotype of 6 archetypal genes—a house-keeping gene which is expressed in every environment and growth condition, a 'good-life' gene which is expressed only at favorable growth conditions, three 'stress-specific' genes, each uniquely associated with one out of three different stress types, and a 'general-stress' gene, which is expressed in any of the stress types, but not during the favorable growth condition. The population experiences changes in the environmental conditions that could be either stressful or optimal. Stresses come in three different types. The different genes are either expressed or not depending on the prevailing conditions at a given moment. During cell doublings mutations are randomly seeded in the genome. These mutations are modeled here as changing codon usage and we refer to them by their effect on expression levels. The fitness of each cell in a given condition is determined by the expression level of all the genes that are needed at that condition, and growth rate is proportional to fitness (see 'Materials and Methods' section for further details). For example, under a particular stress the fitness would be affected by the expression level of the gene that is needed specifically at that condition, and also by the expression level of the 'house-keeping' and general-stress genes. Thus each gene is subject to the

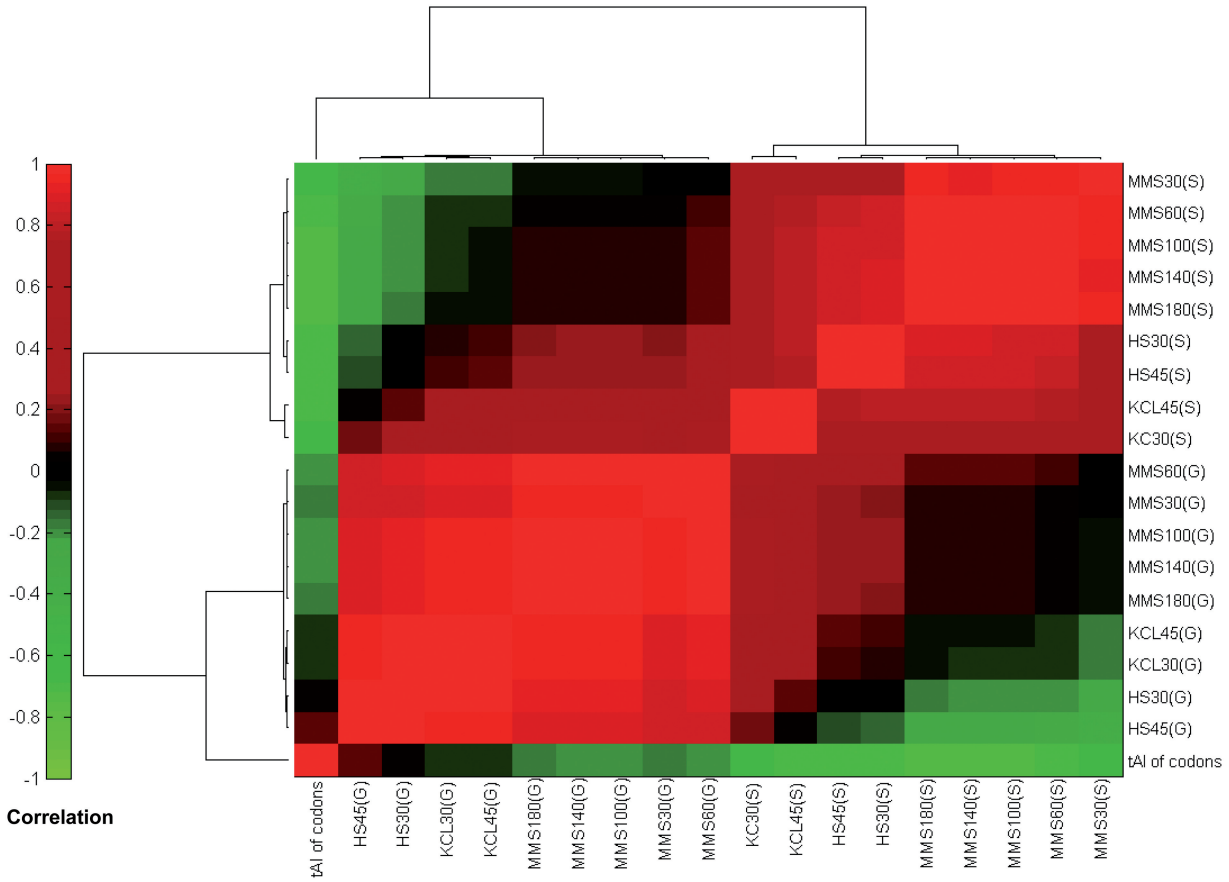


Figure 4. Clustering of codon usage profiles of general stress genes and stress-specific genes. This analysis compared between the codon usage profiles of various sets of genes that are characterized by their mRNA expression pattern across environmental changes. The fold-changes in the codon usage at each time point of a given stress were calculated separately for the stress specific and the general stress genes. In addition the codon-tAI values (W_i) are shown. Each column and row corresponds to one time point in a particular stress. The name of each codon set consists of the condition, followed by time point and a 'S' or 'G' designation indicating whether it was derived from the genes that were specifically changed at that stress or the general stress responsive genes, respectively. Hierarchical clustering was performed with 1-Pearson correlation as a distance metric and 'average linkage'.

effect of drift throughout the simulation, but selection is acting upon it only during times in which the environment is presenting the conditions that require it. We ran the simulation for 10 000 generations (see 'Materials and Methods' section) and followed the extent of expression level of the various genes. We ran the simulation in two modes. In the first mode, there was no bound on the total expression level from all the genes in the genome. This mode simulates a situation in which the tRNA supply was not limiting and only drift limits expression of genes. In the second mode of the simulation the supply of the tRNA is limited so that not all genes can be optimized simultaneously. Hence, whereas in the first mode only drift can compromise expression level of genes, in the second one it was a combination of drift and limited tRNA supply that could act together in limiting expression levels of genes. For each mode, we applied three different environmental regimes, in which the total duration of stressful conditions constitutes 20, 50 or 80% of the total evolutionary time.

The results of the simulation, under the two modes (Figure 5a) show that on average, the 'house-keeping' gene has the highest expression level, followed by either the good-life gene or general stress gene, depending on the fraction of evolutionary time that the population spent in stressful or non-stressful conditions. Regardless of the duration of stressful or non-stressful conditions, the stress-specific genes show the lowest expression level. The differences are observed when drift alone limits genes expression and it becomes even more pronounced when imposing a limitation on the tRNA supply.

How well does the simulation predict differences in codon optimization between house-keeping, stress-specific, general-stress and 'good-life' genes in the genome? To test the simulation we examined translational efficiency, by the tAI measure in the following *S. cerevisiae* gene sets, defined based on micro-array expression levels: (i) 'house-keeping genes', defined here as genes that maintain constant expression level in all conditions (maximal absolute change in expression of 15%); (ii) 'good-life' genes which are repressed under stressful conditions; (iii) general stress genes which are induced in every stress; and (iv) stress-specific genes which are induced in exactly one of the examined stresses (set (iii) and (iv) defined as mentioned above). Figure 5b shows the mean tAI ('Materials and Methods' section) values of the genes

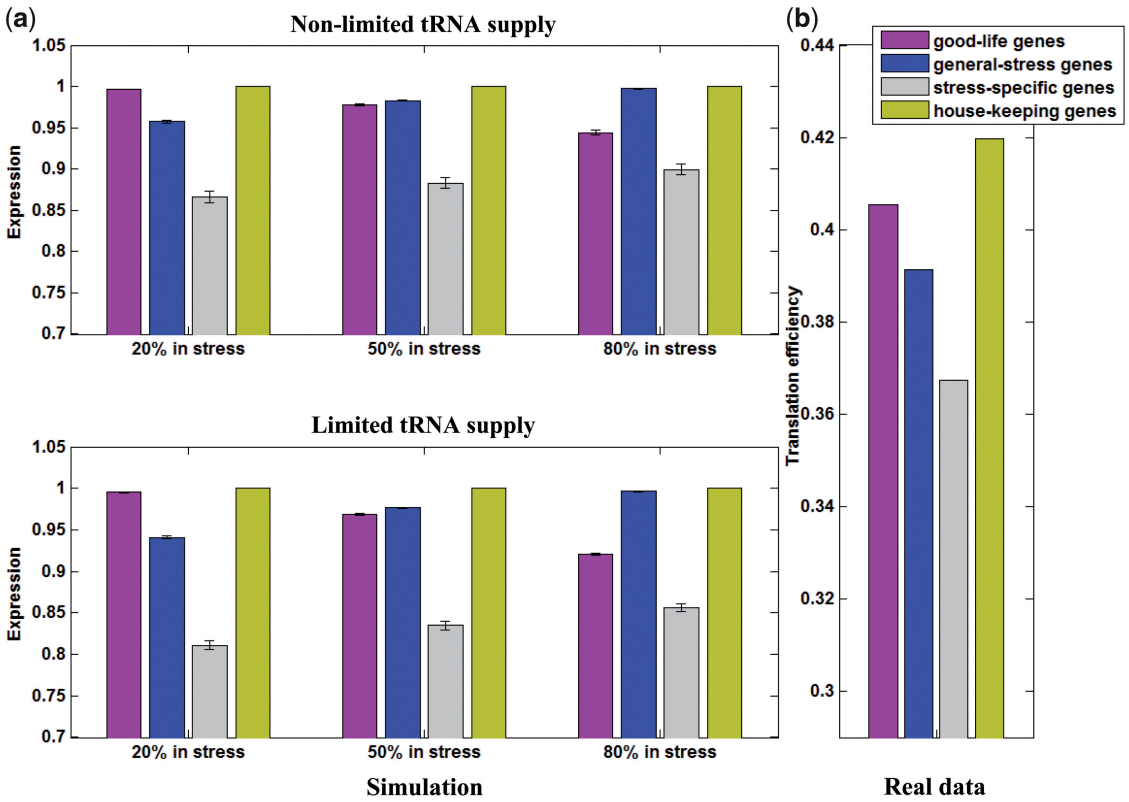


Figure 5. Translational efficiency of different gene sets: comparison of genome data and simulations. (**a**) Measurements of translational efficiency for environmental-dependent gene sets by computational simulation. We simulated population of cells, each possessing a genotype of six genes—a house-keeping gene (green) which is expressed constitutively, a 'good-life' gene (purple), which is expressed only at favorable growth conditions, three stress-specific genes, each expressed in only one of the stress conditions (represented by their average, colored in gray), and, a 'general stress' gene (blue), which is expressed in all of the stress types. Shown for each case is the averaged expression level across the entire simulated evolutionary time, each average, in turn, represents a mean of 50 independent runs of the simulation. The upper panel displays a simulation mode in which only drift limits expression of genes, whereas in the lower panel the tRNA supply too is limited so that not all genes can be optimized simultaneously. The three blocks of bars represent three different environmental regimes that differ in the percentage of evolutionary time in which the population was exposed to stress. (**b**) Measurements of averaged gene tAI values for various sets of genes from the *S. cerevisiae* genome. The bars show the mean tAI value for house keeping genes (green), 'good-life' genes, defined to be the stress-repressed genes (purple), general stress genes (blue) and stress-specific genes, i.e. the union of the gene sets that are specific to each of the stresses (gray).

in each of the four gene sets. The results show a clear rank with significant differences between the gene sets: house-keeping genes show the highest tAI values, followed by the 'good-life' genes, followed by the general stress genes, and ending with the stress-specific genes who show the lowest values (all differences are significant P-value < 0.05, t-test). These results are in good agreement with the simulation and they suggest that genes that are rarely used during the life-history would be poorly optimized and display low expression levels.

DISCUSSION

Traditional measures of translational efficiency either consider the codon usage pattern of the coding sequence alone, or additionally weigh the tRNA pool. Yet, even if the tRNA supply is explicitly taken into consideration, translational efficiency should be further evaluated by the actual consumption of tRNAs. Thus, we suggest here that translational efficiency should be thought as a demand versus supply model, in which the supply is given by the tRNAs availability, and the demand is captured by the representation of the 'consumers'—the codons—in the transcriptome. We show here that the demand varies across conditions, and we anticipate that future investigations will strengthen the notion that the supply too is not constant either in time or between cell types and tissues (10,13).

Our comprehensive model of translational efficiency is based on analysis of the supply-to-demand of the various tRNAs that translate each gene (Figure 6). Both demand and supply may be constant or change across conditions, cell types and developmental stages. Thus, our model of translation efficiency challenges the prevailing simplifying assumption that translation efficiency of a given gene is constant throughout organism life. In turn, our model implies for dynamic range of translation efficiency, suggesting that the interplay between tRNAs availability and codons representation play a role in shaping expression levels of individual genes throughout organism life.

Recently, the change in tRNA synthesis was measured across organs and species in several mammals (13) and was shown to be relatively constant within sets of tRNAs that share an anticodon. Such technology may

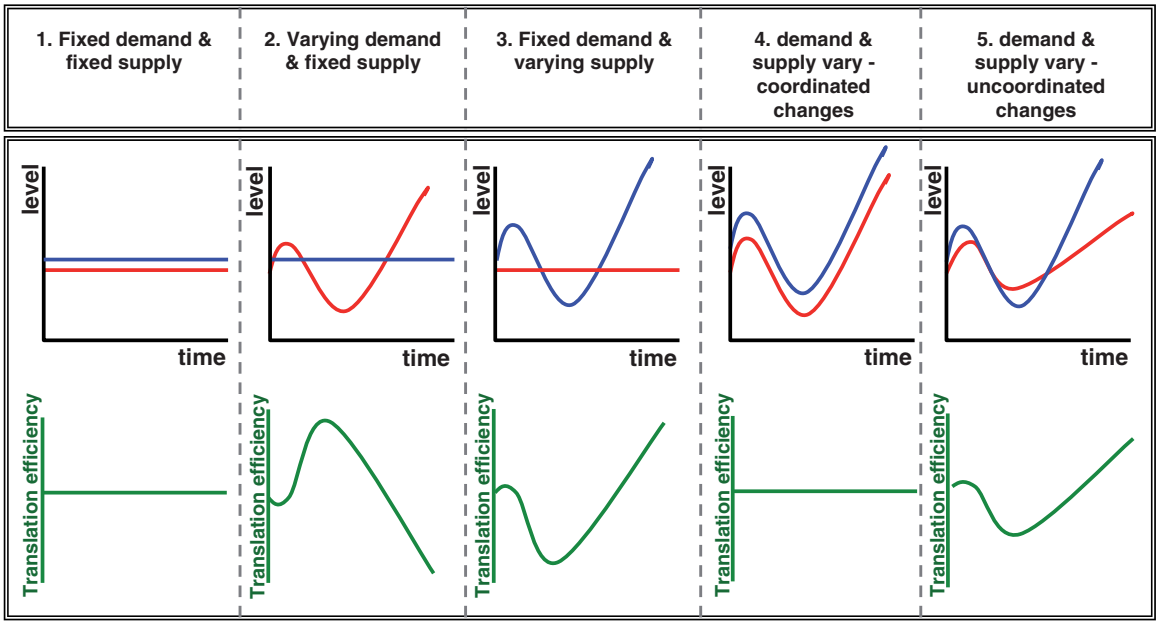


Figure 6. Putative dynamics of supply and demand for tRNAs and its implication on translational efficiency. The figure illustrates five potential modes of control of changes in the tRNAs supply (blue) and demand (red). In the various regimes the supply and demand might both be constant, or either of them, or both of them might change dynamically. Corresponding to each mode of control is the expected effect on translational efficiency (bottom, green).

reveal potential changes in the tRNA pool in micro-organisms when they respond to different conditions. A change or lack of a change in the tRNA supply would affect translation provided that the tRNA is in limiting amounts. Utilizing published estimations on the amount of tRNAs molecules and the translated portion of mRNAs in yeast, we calculated (Supplementary Material) a tRNAs-to-codons ratio for codons. Our calculations, although rough, suggest that tRNAs and their respective codons are present in the cell in the same order of magnitudes of copy numbers. This result implies that there is no appreciable surplus of tRNAs which could buffer changes in the demand without a change in the basal tRNA levels.

A main observation of this study is that the stress transcriptome is poorly adapted to the constant tRNA pool. The challenge is thus to explain this somewhat non-intuitive finding. Clearly a fitness advantage could be gained from better adapted stress codon usage. Yet, as we suggested above, a force that counteracts adaptation is genetic drift. Using a computational simulation, we demonstrate that the balance between drift and selection on an evolutionary time scale may explain the low adaptation of stress-specific genes. Real genomes too have been shown to follow such logic, whereby genes that are needed infrequently in the ecology or life-style of a given species remain poorly adapted to its tRNA pool (5,17,44,45). For these genes the eroding effect of drift prohibits optimal adaptation. This scenario is consistent with case no. 2 in Figure 6, which discusses the possibility that the tRNA pool does not change to match, or counteract a change in codon usage of the transcriptome. However, the simulation results of course do not exclude the possibility that the tRNA pool does change in a condition-dependent manner. Indeed our recent observation actually indicate in that direction (9) as they show that during diauxic shift in yeast the tRNA pool does change to a minor extent. Future experimental efforts along these lines would be needed to establish the possibility that the tRNA pool changes dynamically potentially to off-set changes at the demand.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1 and 2.

ACKNOWLEDGEMENTS

We thank the 'Ideas' program of the European Research Council and the Ben May Charitable Trust for grant support.

FUNDING

Funding for open access charge: European Research Council 'Starting' program.

Conflict of interest statement. None declared.

REFERENCES

1. Sharp,P.M. and Li,W.H. (1987) The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
2. Wright,F. (1990) The 'effective number of codons' used in a gene. *Gene*, **87**, 23–29.
3. Ikemura,T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the

respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J. Mol. Biol.*, **151**, 389–409.
4. dos Reis,M., Savva,R. and Wernisch,L. (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.*, **32**, 5036–5044.
5. Man,O. and Pilpel,Y. (2007) Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat. Genet.*, **39**, 415–421.
6. Tuller,T., Waldman,Y.Y., Kupiec,M. and Ruppin,E. (2010) Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl Acad. Sci. USA*, **107**, 3645–3650.
7. Gingold,H. and Pilpel,Y. (2011) Determinants of translation efficiency and accuracy. *Mol. Syst. Biol.*, **7**, 481.
8. Cannarozzi,G., Schraudolph,N.N., Faty,M., von Rohr,P., Friberg,M.T., Roth,A.C., Gonnet,P., Gonnet,G. and Barral,Y. (2010) A role for codon order in translation dynamics. *Cell*, **141**, 355–367.
9. Tuller,T., Carmi,A., Vestsigian,K., Navon,S., Dorfan,Y., Zaborske,J., Pan,T., Dahan,O., Furman,I. and Pilpel,Y. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, **141**, 344–354.
10. Dittmar,K.A., Goodenbour,J.M. and Pan,T. (2006) Tissue-specific differences in human transfer RNA expression. *PLoS Genet.*, **2**, e221.
11. Dong,H., Nilsson,L. and Kurland,C.G. (1996) Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *J. Mol. Biol.*, **260**, 649–663.
12. Heyman,T., Agoutin,B., Fix,C., Dirheimer,G. and Keith,G. (1994) Yeast serine isoacceptor tRNAs: variations of their content as a function of growth conditions and primary structure of the minor tRNA(Ser)GCU. *FEBS Lett.*, **347**, 143–146.
13. Kutter,C., Brown,G.D., Goncalves,A., Wilson,M.D., Watt,S., Brazma,A., White,R.J. and Odom,D.T. (2011) Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. *Nat. Genet.*, **43**, 948–955.
14. Sorensen,M.A. (2001) Charging levels of four tRNA species in Escherichia coli Rel(+) and Rel(-) strains during amino acid starvation: a simple model for the effect of ppGpp on translational accuracy. *J. Mol. Biol.*, **307**, 785–798.
15. Elf,J., Nilsson,D., Tenson,T. and Ehrenberg,M. (2003) Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science*, **300**, 1718–1722.
16. Begley,U., Dyavaiah,M., Patil,A., Rooney,J.P., DiRenzo,D., Young,C.M., Conklin,D.S., Zitomer,R.S. and Begley,T.J. (2007) Trm9-catalyzed tRNA modifications link translation to the DNA damage response. *Mol. Cell*, **28**, 860–870.
17. Zaborske,J.M., Narasimhan,J., Jiang,L., Wek,S.A., Dittmar,K.A., Freimoser,F., Pan,T. and Wek,R.C. (2009) Genome-wide analysis of tRNA charging and activation of the eIF2 kinase Gcn2p. *J. Biol. Chem.*, **284**, 25254–25267.
18. Plotkin,J.B., Robins,H. and Levine,A.J. (2004) Tissue-specific codon usage and the expression of human genes. *Proc. Natl Acad. Sci. USA*, **101**, 12588–12591.
19. Shalem,O., Dahan,O., Levo,M., Martinez,M.R., Furman,I., Segal,E. and Pilpel,Y. (2008) Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Mol. Syst. Biol.*, **4**, 223.
20. Mitchell,A., Romano,G.H., Groisman,B., Yona,A., Dekel,E., Kupiec,M., Dahan,O. and Pilpel,Y. (2009) Adaptive prediction of environmental changes by microorganisms. *Nature*, **460**, 220–224.
21. Marks,V.D., Ho Sui,S.J., Erasmus,D., van der Merwe,G.K., Brumm,J., Wasserman,W.W., Bryan,J. and van Vuuren,H.J. (2008) Dynamics of the yeast transcriptome during wine fermentation reveals a novel fermentation stress response. *FEMS Yeast Res.*, **8**, 35–52.
22. Kristell,C., Orzechowski Westholm,J., Olsson,I., Ronne,H., Komorowski,J. and Bjerling,P. (2010) Nitrogen depletion in the fission yeast Schizosaccharomyces pombe causes nucleosome loss in both promoters and coding regions of activated genes. *Genome Res.*, **20**, 361–371.
23. Park,S.K., Tedesco,P.M. and Johnson,T.E. (2009) Oxidative stress and longevity in Caenorhabditis elegans as mediated by SKN-1. *Aging Cell*, **8**, 258–269.
24. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
25. Crick,F.H. (1966) Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, **19**, 548–555.
26. Watanabe,K. and Osawa,S. (1995) tRNA sequences and variation in the genetic code. In: Söll,D. and RajBhandary,U. (eds), *tRNA: Structure, Biosynthesis and Function*. AMS Press, Washington, DC, pp. 225–250.
27. Yokoyama,S. and Nishimura,S. (1995) Modified nucleosides and codon recognition. In: Söll,D. and RajBhandary,U. (eds), *tRNA: Structure, Biosynthesis and Function*. AMS Press, Washington, DC, pp. 207–223.
28. Ingolia,N.T., Ghaemmaghami,S., Newman,J.R. and Weissman,J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.
29. Piganeau,G. and Eyre-Walker,A. (2003) Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. *Proc. Natl Acad. Sci. USA*, **100**, 10335–10340.
30. Crow,J.F. and Kimura,M. (1970) *An Introduction to Population Genetics Theory*. Harper & Row, New York, p. 406.
31. Causton,H.C., Ren,B., Koh,S.S., Harbison,C.T., Kanin,E., Jennings,E.G., Lee,T.I., True,H.L., Lander,E.S. and Young,R.A. (2001) Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell*, **12**, 323–337.
32. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
33. Wang,Y., Liu,C.L., Storey,J.D., Tibshirani,R.J., Herschlag,D. and Brown,P.O. (2002) Precision and functional specificity in mRNA decay. *Proc. Natl Acad. Sci. USA*, **99**, 5860–5865.
34. Bernstein,J.A., Lin,P.H., Cohen,S.N. and Lin-Chao,S. (2004) Global analysis of Escherichia coli RNA degradosome function using DNA microarrays. *Proc. Natl Acad. Sci. USA*, **101**, 2758–2763.
35. Molin,C., Jauhiainen,A., Warringer,J., Nerman,O. and Sunnerhagen,P. (2009) mRNA stability changes precede changes in steady-state mRNA amounts during hyperosmotic stress. *RNA*, **15**, 600–614.
36. Amorim,M.J., Cotobal,C., Duncan,C. and Mata,J. (2010) Global coordination of transcriptional control and mRNA decay during cellular differentiation. *Mol. Syst. Biol.*, **6**, 380.
37. Shalem,O., Groisman,B., Choder,M., Dahan,O. and Pilpel,Y. (2011) Transcriptome kinetics is governed by a genome-wide coupling of mRNA production and degradation: a role for RNA Pol II. *PLoS Genet.*, **7**, e1002273.
38. Mazel,D. and Marliere,P. (1989) Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. *Nature*, **341**, 245–248.
39. Kudla,G., Lipinski,L., Caffin,F., Helwak,A. and Zylicz,M. (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol*, **4**, e180.
40. Percudani,R., Pavesi,A. and Ottonello,S. (1997) Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *J. Mol. Biol.*, **268**, 322–330.
41. Kanaya,S., Yamada,Y., Kudo,Y. and Ikemura,T. (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, **238**, 143–155.
42. Konigsberg,W. and Godson,G.N. (1983) Evidence for use of rare codons in the dnaG gene and other regulatory genes of Escherichia coli. *Proc. Natl Acad. Sci. USA*, **80**, 687–691.
43. Zhang,S.P., Zubay,G. and Goldman,E. (1991) Low-usage codons in Escherichia coli, yeast, fruit fly and primates. *Gene*, **105**, 61–072.
44. Bahir,I., Fromer,M., Prat,Y. and Linial,M. (2009) Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol. Syst. Biol.*, **5**, 311.
45. Botzman,M. and Margalit,H. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biol.*, **12**, R109.

Supporting information

Changes in codon usage upon stress are not explained by changes in either amino acid usage or nucleotide usage

Having established that the codon usage changes dynamically during stress we wished to examine whether the change in the representation of a given codon can be explained by a corresponding change in the representation in the transcriptome of the nucleotides that constitute that codon, or alternatively by a change in the representation of its respective amino acid in the translated transcriptome. To examine these two alternative hypotheses we computed the nucleotide and amino acid expression matrices under the same stress conditions. The "Nucleotide expression matrix" is a 4xN matrix whose i,j-th element indicates the extent of appearance of nucleotide i in the transcriptome at condition or time point j. The "Amino acid expression" matrix is a 20xN matrix whose i,j-th element depicts representation of amino acid i at the translated transcriptome at condition or time point j. With the nucleotide expression matrix we ask whether changes at the codon expression matrix can be reduced to, and explained by, changes at the representation of the various nucleotides. Such changes may be related to putative changes in the nucleotide composition of the transcriptome.(1). Likewise, the amino acid expression matrix allows us to ask whether changes at the codon expression matrix simply reflect changes in the relative appearance of the different amino acids at the translated transcriptome, changes that my occur in specific amino-acid cases (2).

We detected only moderate fluctuations in the usage of amino acids upon stresses compared to the changes in the usage of individual codons (Figure S2). Is it possible that the changes in codon usage are simply derived from these changes in amino acid usage? For this purpose we calculated the partial correlations between fold-changes in the representation of individual codons upon stress and the translational efficiency (by the tAI measure) of these codons, while controlling for fold-changes in the usage of the respective amino acids. This analysis shows at most a negligible effect of variations in consumption of different amino acids on the preference of low-efficiency codons upon stress (all partial correlations are very close to the original correlation values). Using the "Nucleotide-Expression" matrix, we detected slight fluctuations in the GC content of the transcriptome upon different types of stress – fold-changes

values vary between 0.99-1.01 and 0.98-1.03 for codon position-independent and codon position-dependent usage of nucleotides, respectively (Figure S2).

Exploring the balance between drift and selection by a computational simulation

We developed a computer simulation of a simplified evolutionary process of unicellular population of a fixed size of 1,000,000 haploid cells for 10,000 generations. The genome of each cell consists of six genes – a house-keeping gene that is expressed in every environment and growth condition, a 'good-life' gene, corresponds to favorable growth conditions, three 'stress-specific' genes, which are uniquely associated with three different stress types, and a 'stress-generic' gene, which is essential for any stress type.

At the beginning of the simulation, the six genes are equally scored with initial arbitrary value of expression level that denotes optimal expression. The population then evolves while subjected to a fixed mutation rate, that is, the frequency of 0.001 substitutions per genome, in line with realistic values (3,4). Sequences are not represented explicitly in the simulation; instead genes are characterized by an expression level that implicitly corresponds to a genotype. Thus, "mutated" expression levels at a given time step are computed by the previous step's expression levels multiplied by a random number drawn from an exponential probability distribution of changes in expression (as estimated before (REF 5)).

We set the rate parameter λ to be 1.5, hence approximately eighty percent of the mutations are assumed to be deleterious. Running the simulation with less deleterious mutations ($\lambda = 1$), reproduces the results.

We ran the simulation in two modes. In the first, mutations affected the expression of genes, but there was no bound on the total expression level for all the genes in the genome. In the second mode of the simulation mutations affected expression as in the first mode, yet in addition the tRNAs supply is limited, so that not all genes can be optimized simultaneously. Practically, we forced a constant maximal total expression level from all genes. In this mode of limited supply of tRNAs, the expression of the *i*-th gene in each generation, $lsExpression_{gi}$, ("ls" stands for "limited supply") is defined by

$$lsExpression_{gi} = \begin{cases} Expression_{gi} * \left(\dfrac{MaxExpression}{\sum_{i=1}^{n} Expression_{gi}} \right) & if\ (\sum_{i=1}^{n} Expression_{gi} > MaxExpression) \\ Expression_{gi} & else \end{cases}$$

where n is the number of genes in the cell, $Expression_{gi}$ is the expression of the i-th gene in the mutated population, and $MaxExpression$ is a constant maximal total expression level from the whole "genome".

The evolving population of cells is exposed to occasional stress periods that come in three types, stress1, stress2 and stress3. Specifically, we applied three different regimes, in which the total duration of stressful conditions constitutes 20, 50 or 80 percent of the total evolutionary time.

Individual cells are selectively transferred for the next generation, as a function of their fitness. The fitness of a given cell is determined by a weight given to it according to the expression of its genes which are associated with the current environmental condition during which a distinct cell division event occurs. Specifically, the fitness in favorable growth conditions is a function of the expression values of the 'house-keeping' and 'good-life' genes, whereas the fitness during stress is determined as the averaged expression value of the 'house-keeping' gene, the relevant 'stress-specific' gene and the general stress gene. Practically, we measured the change in the fitness of individual cells as the absolute value of the difference of expression values of the condition-related genes from the optimal one. Having the fitness values for all the cells in the population, the simulation program selects cells for the next generation. Formally, the numeric change in the size of homogeneous population can be described as

$$\frac{dx}{dt} = \lambda x \left(1 - \frac{x}{K} \right)$$

where x denotes the population size, λ corresponds to the fitness, and K indicates the carrying capacity according to the logistic model. For a heterogeneous population consisting of two genotypes, the respective equations are

$$\frac{dx_1}{dt} = \lambda_1 x_1 \left(1 - \frac{x_1 + \alpha_{21} x_2}{K} \right) \text{ and } \frac{dx_2}{dt} = \lambda_2 x_2 \left(1 - \frac{x_2 + \alpha_{12} x_1}{K} \right)$$

where $\alpha_{21}x_2$ and $\alpha_{12}x_1$ describe the constraint enforced by the growth of genotype-2 subpopulation on the growth of genotype-1 subpopulation, and vice versa, respectively. Generalized to a higher number of sub-populations, the change in representation of genotype ***i*** in the population at time interval ***t*** can be described as

$$\frac{dx_i}{dt} = \lambda_i x_i \left(1 - \frac{x_i + \sum_{j \neq i} \alpha_{ji} x_j}{K} \right)$$

which reduces back to the one-population case if $\alpha_{ij} = 1$ for all *i*,*j* pairs

We propagate individuals between consecutive generations (t-1) to (t) in two stages. First, a population (whose size can be different from that of the population at generation t-1) is formed in which the *i*–th genotype population size is given by its size in the previous generation and its fitness by:

$$x_{i(t)} \approx x_{i(t-1)} e^{\lambda_i}$$

Then, to keep a constant population size stochastic rescaling is applied that implements a Kimura-governed (5) allele sampling.

Calculation of tRNAs-to-codons ratio

We performed a rough analysis that aimed to assess the relative abundance of tRNAs and codons in the cell. In particular, we examined the six rarest tRNAs in *S. cerevisiae*, each of which is encoded in the yeast genome by only one tRNA gene. These six rare tRNAs correspond to seven codons: CGG (Arg), CAG (Gln), ACG (Thr), UCG (Ser), AGG (Arg), CUU (Leu) and CUC (Leu). There is one-to-one correspondence between each of the first four codons and their tRNA; Codon AGG can be also translated by the fully-matched tRNA of AGA (6); the last two codons are translated by the same tRNA type, hence are counted together.

Estimates suggest that a yeast cell contains some 3.3 million tRNA molecules (BioNumbers database (7) and (8)). The copy number of molecules of each tRNA type is simply the fraction of its tRNA gene copy number out of the total gene copy number of all tRNA types multiplied by 3.3 million. As for codons, the number of codons of any type in the transcriptome is defined by the sum of appearances of a codon along all genes in the genome, multiplied by the average mRNA abundance in the cell (9)). To consider specifically the subset of codons that are actively translated, we consider the fraction of mRNAs which are occupied by at least one ribosome (=0.71, (10)).

The table below shows the ratio of the number of tRNA molecules to the corresponding codon copy number for the above selection of codons. As can be seen, the ratio is never larger or smaller than 10, suggesting that tRNA and their respective codons are estimated to be in similar amounts in the cell.

Codon	tRNA/codon abundence
Arg (agg)	0.22
Arg (cgg)	1.15
Gln (cag)	0.17
Ser (ucg)	0.24
Thr (acg)	0.26
Leu (cuc & cuu)	0.12

References

1. Kudla, G., Lipinski, L., Caffin, F., Helwak, A. and Zylicz, M. (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol*, **4**, e180.
2. Mazel, D. and Marliere, P. (1989) Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. *Nature*, **341**, 245-248.
3. Drake, J.W., Charlesworth, B., Charlesworth, D. and Crow, J.F. (1998) Rates of spontaneous mutation. *Genetics*, **148**, 1667-1686.
4. Joseph, S.B. and Hall, D.W. (2004) Spontaneous mutations in diploid Saccharomyces cerevisiae: more beneficial than expected. *Genetics*, **168**, 1817-1825.
5. J.F.Crow and Kimura, M. (1970) *An Introduction to Population Genetics Theory*. Harper & Row, New York.
6. Johansson, M.J., Esberg, A., Huang, B., Bjork, G.R. and Bystrom, A.S. (2008) Eukaryotic wobble uridine modifications promote a functionally redundant decoding system. *Mol Cell Biol*, **28**, 3301-3312.
7. Milo, R., Jorgensen, P., Moran, U., Weber, G. and Springer, M. (2010) BioNumbers--the database of key numbers in molecular and cell biology. *Nucleic Acids Res*, **38**, D750-753.
8. Waldron, C. and Lacroute, F. (1975) Effect of growth rate on the amounts of ribosomal and transfer ribonucleic acids in yeast. *J Bacteriol*, **122**, 855-865.
9. Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S. and Young, R.A. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717-728.
10. Arava, Y., Wang, Y., Storey, J.D., Liu, C.L., Brown, P.O. and Herschlag, D. (2003) Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, **100**, 3889-3894.
11. Shalem, O., Dahan, O., Levo, M., Martinez, M.R., Furman, I., Segal, E. and Pilpel, Y. (2008) Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Mol Syst Biol*, **4**, 223.

Figure S1: The fold-changes in representation of amino acids and nucleotide types in the transcriptome upon stress. (a) "Amino acid-Expression" matrix for diverse stress types, normalized as in Figure 2a. Each cell denotes the fold-change in the representation of a given amino acid in the transcriptome at a given time point upon specific stress, compared to its representation at time point zero. The amino acid labels are followed by numbers in parentheses, indicating the sum of gene copy number of all their corresponding tRNAs. (b) "Nucleotide-Expression" matrix for diverse stress types, normalized as in Figure 2a. Each cell denotes the fold-changes in the representation of a given nucleotide in the transcriptome at discrete time point (minutes) upon a specific stress, compared to its representation at the corresponding time point zero. A one letter label (a,c,g and u) refers to the total nucleotide representation, whereas specific codon position labels indicate the usage of a given nucleotide at each of the three positions of the codon.

Figure S2: Correlation between the codons adaptiveness values and the change in their representation in the transcriptome under stress. We calculated the Pearson correlation coefficient between 61-long vectors denoting fold-changes in the codon usage of the transcriptome in different time points (minutes) of diverse environmental conditions and the 61 codons' tAI values. A consistent negative correlation between the codons adaptiveness values (W_i) and their representation in the transcriptome in stress can be seen. The most negative correlations among the different time points in each of the examined stress types vary between -0.52 (oxidative stress) and -0.73 (MMS). Other than the correlation value for the first time point of the oxidative stress, all the correlations were found to be significant, with p-values spanning a range of 2.45×10^{-11} to 4.76×10^{-2}. The recovery from both heat-shock and the KCL stresses, (labeled 'R'), obtained by transferring the cells from the respective stressful conditions to normal growth conditions, is accompanied by sharp increase of the measured correlations between the codons' adaptiveness value (W_i) and their representation in the transcriptome, towards significant positive values (KCL: Pearson Correlation = 0.7, p-values = 4.43×10^{-10}; heat-shock: Pearson correlation = 0.67, p-values = 3.14×10^{-9}). We detected a similar pattern of change in the direction of the correlation, though with relatively moderate slope, for the oxidative stress, probably as a result of spontaneous recovery from the stress (11).

Figure S1

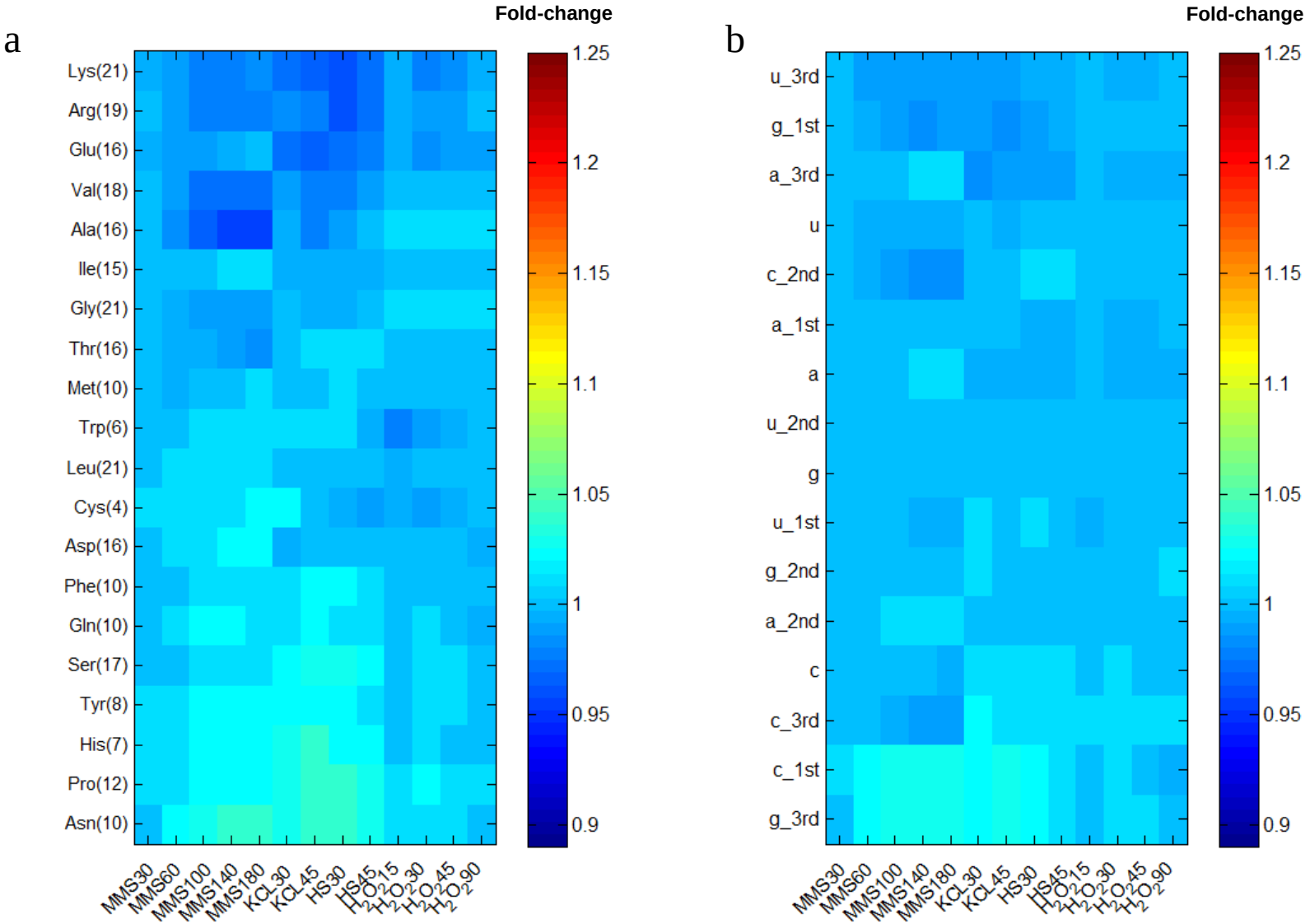
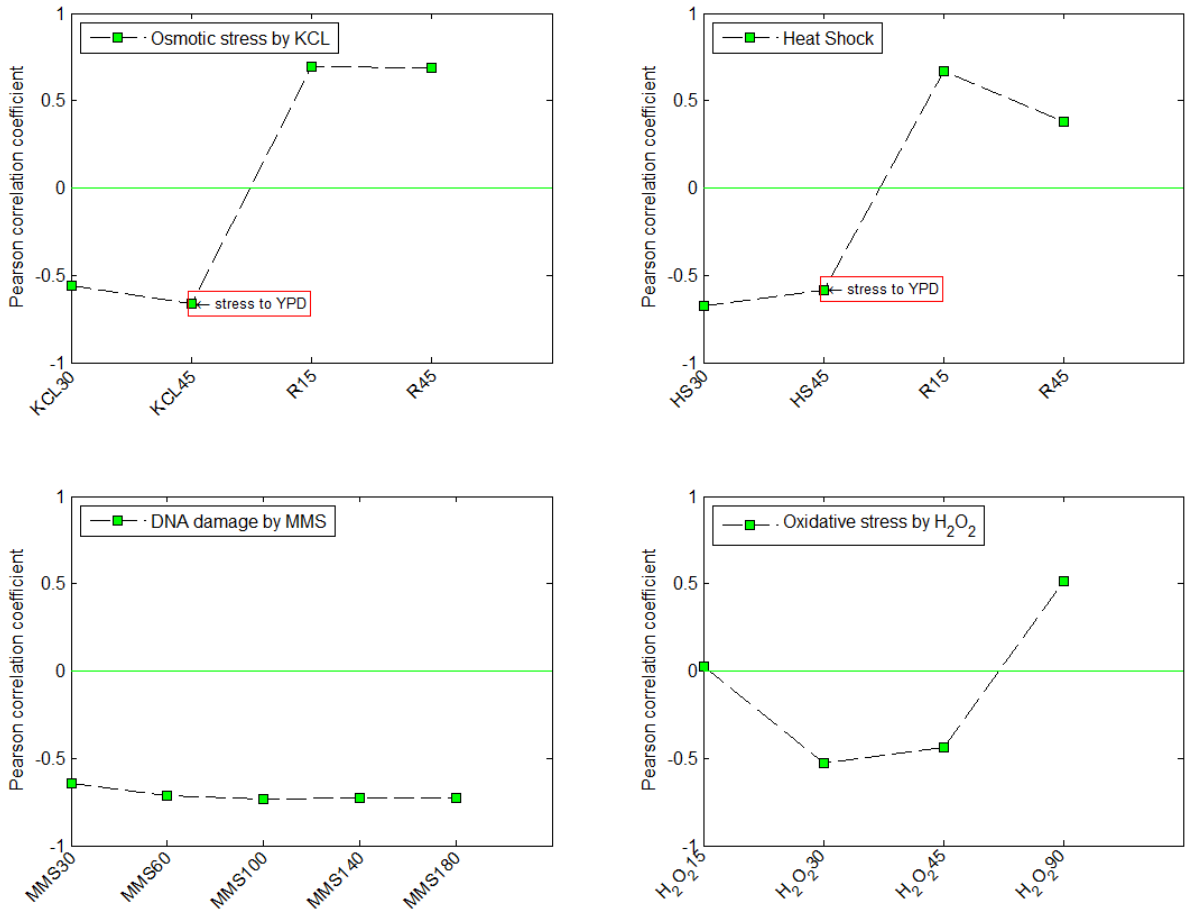

a
Fold-change
Lys(21)
Arg(19)
Glu(16)
Val(18)
Ala(16)
Ile(15)
Gly(21)
Thr(16)
Met(10)
Trp(6)
Leu(21)
Cys(4)
Asp(16)
Phe(10)
Gln(10)
Ser(17)
Tyr(8)
His(7)
Pro(12)
Asn(10)
1.25
1.2
1.15
1.1
1.05
1
0.95
0.9
MMS30
MMS60
MMS100
MMS140
MMS180
KCL30
KCL45
HS30
HS45
H2O215
H2O230
H2O245
H2O290
b
Fold-change
u_3rd
g_1st
a_3rd
u
c_2nd
a_1st
a
u_2nd
g
u_1st
g_2nd
a_2nd
c
c_3rd
c_1st
g_3rd
1.25
1.2
1.15
1.1
1.05
1
0.95
0.9
MMS30
MMS60
MMS100
MMS140
MMS180
KCL30
KCL45
HS30
HS45
H2O215
H2O230
H2O245
H2O290

Figure S2


Osmotic stress by KCL
stress to YPD
Pearson correlation coefficient
1
0.5
0
-0.5
-1
KCL30
KCL45
R15
R45
Heat Shock
stress to YPD
HS30
HS45
R15
R45
DNA damage by MMS
MMS30
MMS60
MMS100
MMS140
MMS180
Oxidative stress by H_2O_2
H_2O_215
H_2O_230
H_2O_245
H_2O_290

6. Cancerous processes may determine the cell fate by hijacking the translation machinery.

6.1 Introduction

The regulation of gene expression in cancer is of obvious immense importance and interest. While traditionally studies focus on deciphering changes in the *transcriptome* upon cancer, researchers are now increasingly interested in measuring *translation* and its changes in cancer. Originally, the interest in translation was mainly focused on initiation control (Mamane et al. 2006; Sonenberg and Hinnebusch 2009), and more recently translation elongation gains further attention (Hsieh et al. 2012). Particularly the role of the tRNA pool in proliferation and cancer is only beginning to be characterized and understood. In principle the tRNAs could affect the proliferation state of the cell, and, conversely they could also be affected by proliferation status of the cell. Small-RNA deep sequencing measurements (Yang et al. 2010a) and tRNA customized arrays (Pavon-Eternod et al. 2009) are beginning to provide data regarding changes in tRNA availability in cancer. For instance, by measuring the tRNA levels in several breast tumors using dedicated arrays, it was shown that the tRNA levels are selectively elevated, even up to 10 fold in the cancerous state compared to corresponding normal samples. Yet, which tRNAs display which types of changes and what is their effect on the cell is not known.

In their classic "The Hallmarks of Cancer", Hanahan and Weinberg (Hanahan and Weinberg 2000) state that "Our tissues also constrain cell multiplication by instructing cells to enter irreversibly into postmitotic, differentiated states, using diverse mechanisms that are incompletely understood; it is apparent that tumor cells use various strategies to avoid this terminal differentiation" and that "…cells may be induced to permanently relinquish their proliferative potential by being induced to enter into postmitotic states, usually associated with acquisition of specific differentiation-associated traits." Indeed, proliferation and differentiation are distinct cellular states; generally speaking, differentiated cells are less proliferative and proliferating cells are not terminally differentiated. This dichotomy is often reflected in a clear molecular profile: for instance, the transcription regulator Max can interact with c-Myc to induce proliferation or with an alternative partner, Mad, shifting the

balance towards differentiation (Hanahan and Weinberg 2000). While cancer provides a classical demonstration of the proliferation/differentiation dichotomy, such distinction was recently further illustrated also in a normal healthy organ – the adult mammalian liver (Klochendler et al. 2012). The researchers identified a small percentage of proliferative cells in this organ, and found them to have reduced levels of the liver differentiation marks (Klochendler et al. 2012). Moreover, an interesting distinction was found between the transcriptome of the dividing and differentiated cells of the organ. These studies reinforced the notion that the transcription program of differentiated and proliferative cells might be distinct and even negatively correlated. Interestingly, much less is known about the corresponding level of translation – Do cells feature distinct states of translation control when they proliferate or differentiate? Do different tRNAs for the same amino acid, prevail in differentiated or proliferative cells, and, if so, how are changes in the tRNA pool achieved? How does codon usage of the transcriptome change between these two cellular states? Does the proteome change when the tRNA pool changes? And perhaps above all, do potential changes in translation determine the proliferation/differentiation status of mammalian cells, and conversely does that cellular status affect gene translation?

6.2 Results

6.2.1. Recurring changes occur in the tRNA pool in cancer patients and in cell lines

In order to follow changes in the tRNA pool in cancer and related physiological processes we obtained from our collaborator, Andres Lund from the University of Copenhagen, experimental data, created with costumed microarrays, of the expression of 206 tRNA genes and ~7000 protein-coding transcripts in samples from 300 cancer patients, cell lines and normal tissues. The caner types consisted of diffuse large B-cell lymphoma, Bladder cancer, Colon cancer and Prostate cancer. The data presented expression level of 206 out of 516 human tRNA genes; these 206 genes represent 31 out of the 47 tRNA species in human, and are involved in the translation of 16 out of the 20 standard amino acids. We summed the expression of tRNA genes of the same species (anticodon) and calculated the changes in the expression of tRNA genes compared to their expression in the respective normal tissues.

Figure 1 shows the variation in the tRNA pool in 69 patients with diffuse large B-cell lymphoma, compared to the tRNA pool averaged over 10 normal B-cell samples. Overall, the tRNA pool changes reproducibly in different patients, where the expression of some tRNA genes is elevated in cancerous cells compared to normal B-cells, and the expression of others reduces.

Interestingly, tRNA types which translate synonymous codons may show opposite trend of change in expression – for instance, two tRNAs for Lysine, each of which exclusively translates one of the two codons for this amino acid show opposite behavior in cancer: the tRNA-Lys(CUU) is up-regulated in most patients while tRNA-Lys(UUU) is often down-regulated. Since different tRNA types of synonymous codons are loaded with amino acid by the same tRNA synthetase type, then if the expression of one of them increases while that of the others decreases, the former may not only gain from the direct elevation in its availability but may also experience an increase in accessibility to the enzyme, thus having a higher fraction of ready-to-translate tRNAs, due to changes in the proportions of charging tRNAs.

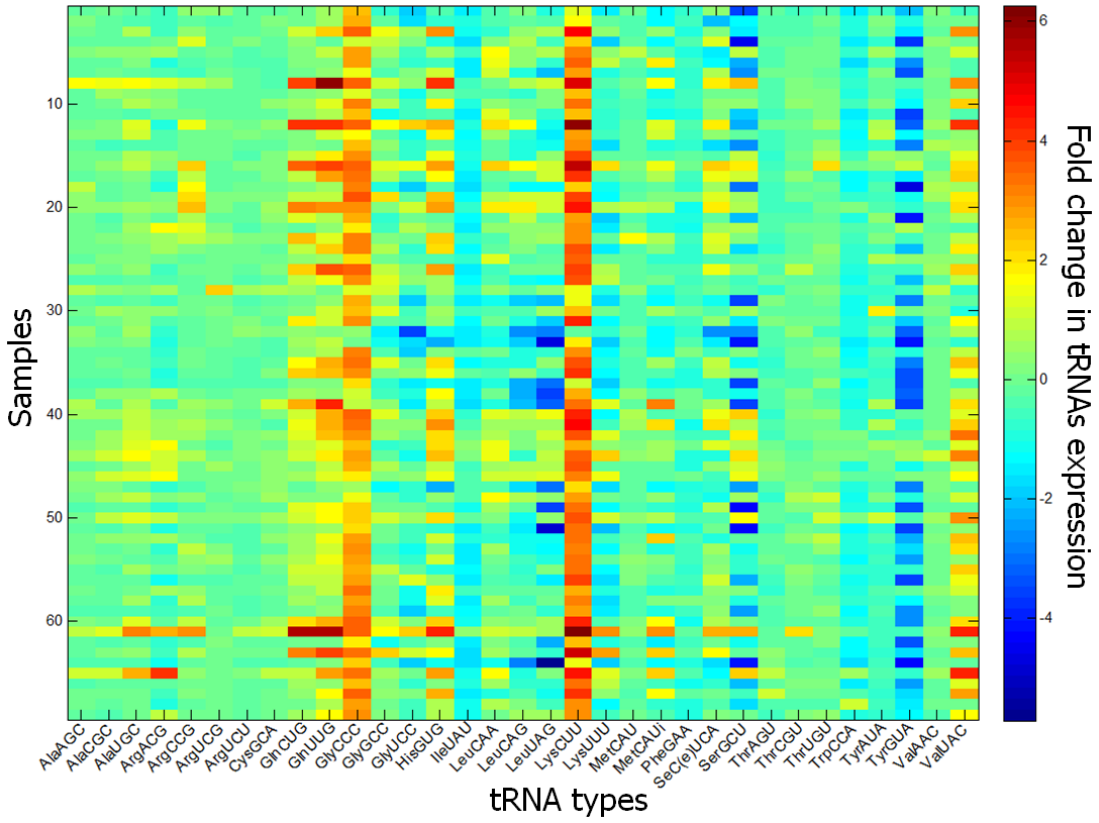


Figure 1: The tRNA pool changes reproducibly in cancer. Shown are the fold-changes in the expression of 33 human tRNA types in 69 patients with diffuse large B-cell lymphoma, compared to the expression of these tRNAs in normal B-cells, as was averaged across 10 different samples. The name of each tRNA consists of the amino-acid which it translates followed by the anticodon; 32 tRNA types translate standard amino acids, where one additional type translates Selenocysteine. The tRNA type which corresponds to Methionine is represented by its two types – elongator tRNA (MetCAT) and initiator tRNA (MetCATi). Fold-changes are shown in log2 scale.

6.2.2 The initiator-tRNA is over-expressed in naturally occurring cancer

One particular observation was obtained by comparing changes in the expression of the initiator and elongator tRNA for methionine (Figure 2). While we detected an induction of the initiator tRNA-Met in many of the patient samples, there was little if any variation in the expression of the elongator tRNA-Met. These results suggest that the elevation of the initiator tRNA is selected for in cancer due to a potential oncogenic effect. Indeed it was recently shown that over-expression of the initiator tRNA-Met significantly affects the tRNA expression profile and elevates cell proliferation in human epithelial cells (Pavon-Eternod et al. 2013). These findings suggest that perhaps more generally up-regulated tRNAs in our data might have oncogenic effects, while down-regulated ones might be tumor suppressive.

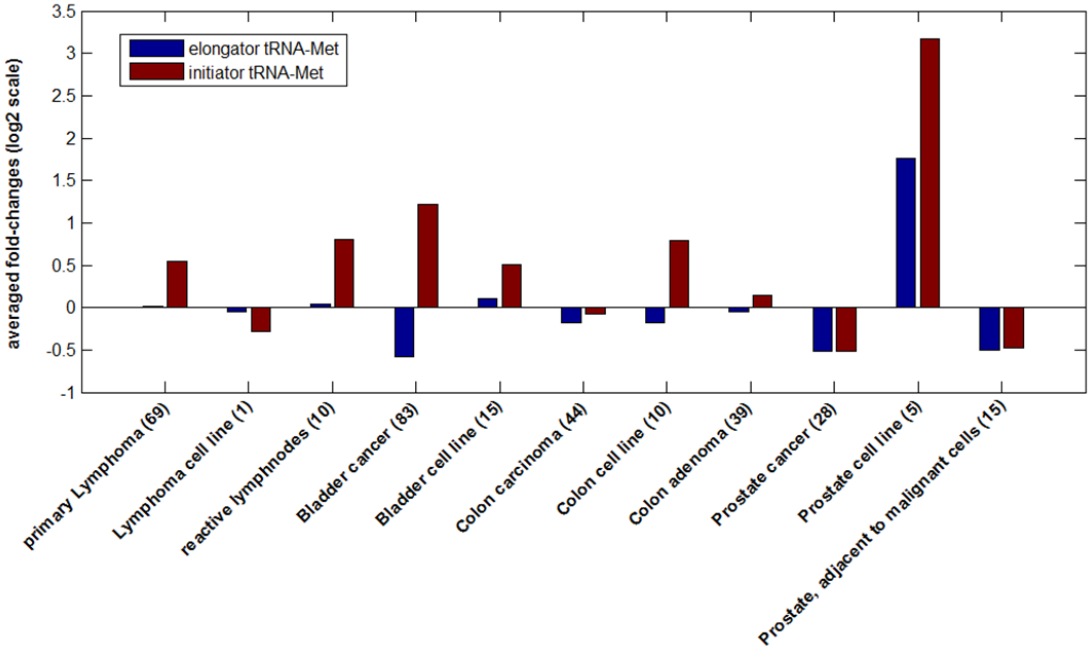


Figure 2: The initiator, but not the elongator tRNA-Met is induced in several cancer types. Shown are the averaged fold-changes in the expression the two tRNA-Met types in cell lines and primary tumors of four origins: B-cells, bladder, colon and prostate. Also shown are reactive lymphnodes of which some are also malignant; also shown are cells from patient with adenoma (colon) and non-malignant cells that are adjacent to prostate malignant cells. The numbers in parentheses denote the number of samples from each cell type. Fold-changes are shown in log2 scale, and were calculated compared to the averaged expression of the tRNAs in the corresponding normal cells.

6.2.3 Distinct cancer types show similar signature of variation in the tRNA pool

To further investigate the interplay between cancerous processes and the tRNA pool, we calculated the similarity of the variation in the tRNA pool among the cancerous samples and cancerous cell lines (Figure 3). The figure captures the similarity of change in the tRNA pool of samples that belong to the same type (main diagonal) and a comparison of the tRNA pools of samples that belongs to different type (off-diagonal).

In most of the examined cell types the pattern of change in the tRNA pool among different samples that belong to the same type is highly similar (as depicted by the high correlation values on the main diagonal). The correlations off-diagonal, between samples of different types vary and they appear to be governed by the status of the sample (namely whether it is derived from a primary tumor or from a cell line), rather than by the origin of the cells in the body. In particular, all the cell lines are

clustered together, and there are distinct from all primary tumors, that are also clustered together, irrespective of tissue of origin. This clustering into primary tumors and cell lines, and not by body origin is obtained if we assess similarity between samples using the tRNAs expression values, but not if we use other genes' expression as a means to classify samples. For comparison we also clustered the sample types based on mRNA coding genes rather than tRNA genes and observed a totally different classification, in which each primary tumor co-clustered with its matching cell line from the same tissue.

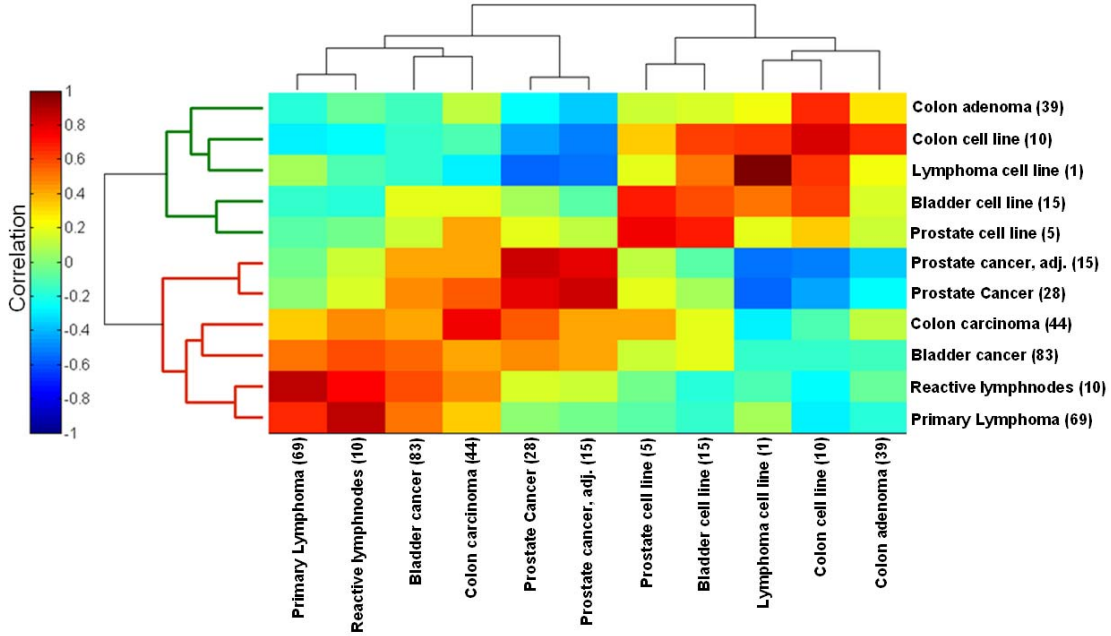


Figure 3: Clustering of samples based on tRNA expression partitions all cell lines separately from all primary tumors. This analysis compares between the averaged tRNAs expression profiles of nine different cell types. Cells are characterized by both their origins (B-cells, bladder, colon and prostate), and status – cancerous (tumor) cells, cell lines or other cell types. Each column and row corresponds to one cell type, where the numbers in parentheses denote the number of samples belonging to that cell type. Off-diagonal cells denote the correlation between the tRNAs expression profiles of two different cell types; the profile of a given cell type was defined by the averaged fold-changes in the tRNAs expression of the individual samples. The main diagonal, from top right to bottom left, depicts the median Pearson correlation among all possible pairwise comparisons of samples from the same cell type. Hierarchical clustering was performed with 1-Pearson Correlation as a distance metric.

6.2.4 The cancerous tRNA pool affects genes' translation efficiency in a differential manner

Identifying recurring changes in the representation of various tRNA types across different cancer types, we next aimed to deduce the implication of the variation in the tRNA pool on the translation efficiency of individual genes, and functional sets

of genes in the genome. For this purpose, we employed the tAI measure of translation efficiency of genes (dos Reis et al. 2004). Briefly, the tAI model assesses the translation efficiency of genes by the availability of the tRNAs that serve in translating it, incorporating both the fully-matched tRNA, as well as tRNAs that contribute to translation obeying wobble rules (Crick 1966). Typically, while applying the tAI model, the amount of the different tRNAs in cells is often deduced from the copy number of all the tRNA-coding genes in the genome. Formally, the "adaptiveness" of the *i*–th codon to the tRNA pool is defined by

$$W_i = \sum_{j=1}^{n_i} \left(1 - s_{ij}\right) tGCN_{ij}$$

where *n* is the number of tRNA isoacceptors that recognize the *i*-th codon, $tGCN_{ij}$ denotes the gene copy number of the *j*-th tRNA that recognizes the *i*-th codon, and s_{ij} correspond to the wobble interaction, namely selective constraint on the efficiency of the pairing between codon *i* and anticodon *j*. The adaptiveness value of codon *i* is further divided by the maximum W_i (termed W_{max}), obtaining the codon's relative adaptiveness value:

$$w_i = W_i / W_{\max}$$

The tAI value of a gene is then simply calculated as the geometric mean of its codon values

$$tAI(g) = \sqrt[L]{\prod_{k=1}^{L} w_{i_{kg}}}$$

where i_{kg} is the codon defined by the *k*th triplet in gene *g* and L is the length of the gene in codons (except the stop codon). In computing translation efficiency, and changes thereof, in cancer, we resorted to tRNA expression values from the arrays, instead of the constant tRNA gene copy number in the (non-cancerous) genome. Yet our array data do not indicate gene copy variation of all the human tRNA genes, but only for a partial set of 206 genes. As a result, we cannot compute the absolute tAI values of genes in cancer, yet instead we can calculate the changes in the predicted translation efficiency, due to the tRNAs for which we have probes, for every gene in cancerous cells compared to the respective normal tissues:

$$\text{Ratio}(tAI_{Cancer}, tAI_{Normal}) = \frac{tAI_{Cancer}}{tAI_{Normal}} = \prod_{i=1}^{37} \left(\frac{tEXP_{ij(Cancer)}}{tEXP_{ij(Normal)}} \right)^{f^i}$$

where i is one of 37 codons whose corresponding tRNA type is represented in the microarray, $tEXP_{ij}$ denotes the expression level of the j-th tRNA that recognizes the i-th codon, and f^i is the fraction of codon i in the gene. Note that most of the human codons are translated by a single tRNA type; for two Tyrosine codons that are translated by two tRNA types, we assess in our calculation the marginal contribution of each of the tRNA types to their translation.

In addition to computing tAI change for individual genes we also computed such scores for entire GO categories (we worked on all categories belonging to the "Biological Processes" classification, provided that they have at least 40 genes). We represented each category by the median change in the predicted translation efficiency of the genes that belong to it (figure 4).

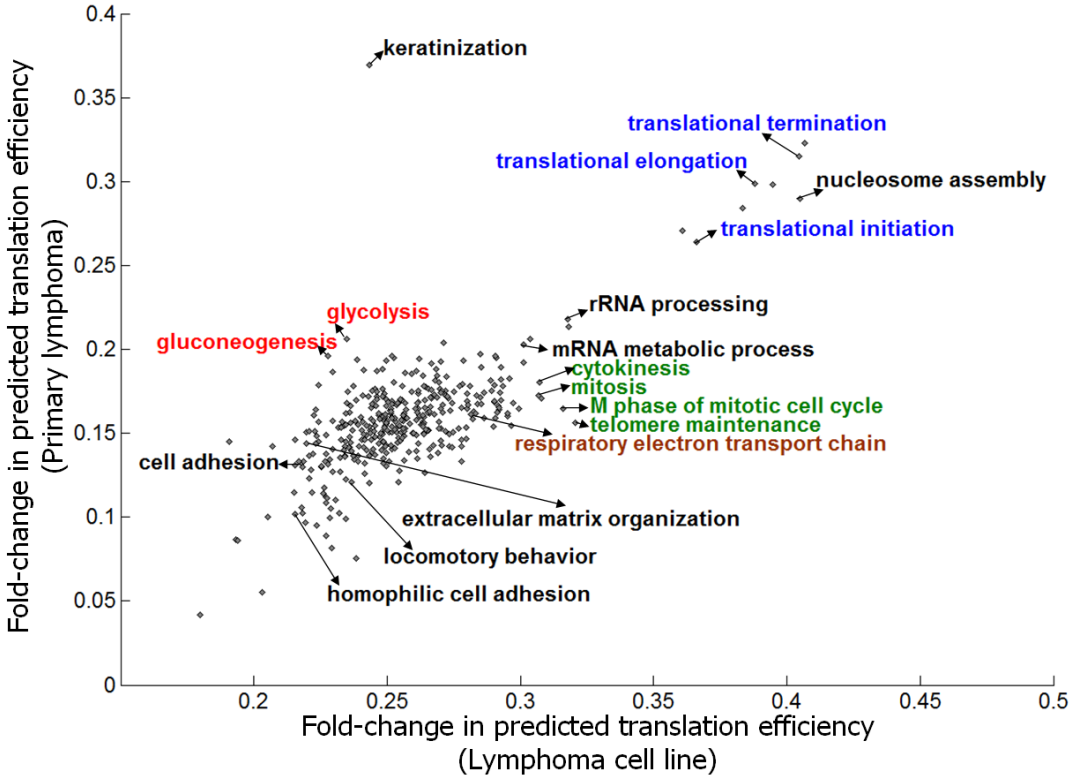


Figure 4: Similarity and difference in translation efficiency in primary Lymphoma and Lymphoma cell line. Each dot represents one GO term (by biological process categories) with at least 40 genes. The x-axis shows the median (log2) change in the predicted translation efficiency of the genes belonging to a given GO category in Lymphoma cell line (HT), where the y-axis shows median (log2) change in the predicted translation efficiency (by gene-tAI) of these genes in primary cancerous cell (averaged over samples of 69 patients).

Over all, translation efficiency changes quite reproducibly among different primary cancerous cells and cell line of same origin. In most of the examined cell types, genes belonging to GO categories of "translation", "mRNA metabolic processes" and "nucleosome organization" show elevation in their predicted translation efficiency in both primary cancerous cells form patients and cancerous cell lines, whereas genes belonging to GO categories such as "cell adhesion" and "extra cellular matrix organization" show reduction in their predicted translation efficiency in these two cell types (see demonstration in Figure 4 for the case of lymphoma). Interestingly, we observed differential changes in translation of distinct gene sets while comparing cells from patients to cancerous cell lines. Specifically, genes associated with "mitotic processes" and "telomere maintenance" are observed to have elevated predicted translation efficiency in cancerous cell lines compared to primary tumors. Primary tumors show striking elevation in the predicted translation efficiency of keratinization-related genes, consistent with evidences suggesting an active role for keratin in cancer cell invasion and metastatis (Karantza 2011).

6.2.5 An underlying modular design of the genome codon usage distinguishes between distinct biological processes.

How can changes in expression level of certain tRNA boost specifically the translation of certain genes while reducing others? Such a putative effect is probably only possible if distinct gene sets that are induced or repressed during proliferation would have a different codon usage in the normal, non-cancerous genome.

We thus turned to check whether the codon usage of mRNA-coding genes enables distinguishing between genes belonging to different biological processes. For this purpose, we first calculated the codon usage of all the human genes (in cases of two or more transcripts per gene, we calculate an averaged values). Then, for each GO term (by Biological Process) we calculated the average codon usage of the belonging genes. In this analysis each GO category resides in a 61-dimensional space of codon frequencies. Finally, to visualize the data we ran a Principal component analysis (PCA) on all 395 GO terms that contain at least 40 genes.

The result of the PCA is shown in figure 5. Much to our surprise the first two PCs were found to span a significant portion of the variance that exists in the entire set of 61 dimensions (40% and 18% by the first and second PC respectively). A

striking result was that especially along the first PC gene categories belonging to differentiation and proliferation categories were clearly distinct. This result indicates, in general, for differential codon usage of the various gene categories, and particularly for the opposing functions of differentiation and proliferation. Interestingly, genes associated with translation are located in proximity to the proliferation categories along the first component. In the vicinity of the proliferation-related categories in this PCA projection there are also GO categories associated with mRNA metabolic process, meiosis, nucleosome assembly and cell division (not marked). In the vicinity of the differentiation genes there are genes associated with pattern specification and cell adhesion. Thus, our results imply the differential codon usage of genes associated with the various process of the cell cycle and the Central Dogma, i.e. functions at the unicellular level, and the genes associated with multi-cellularity.

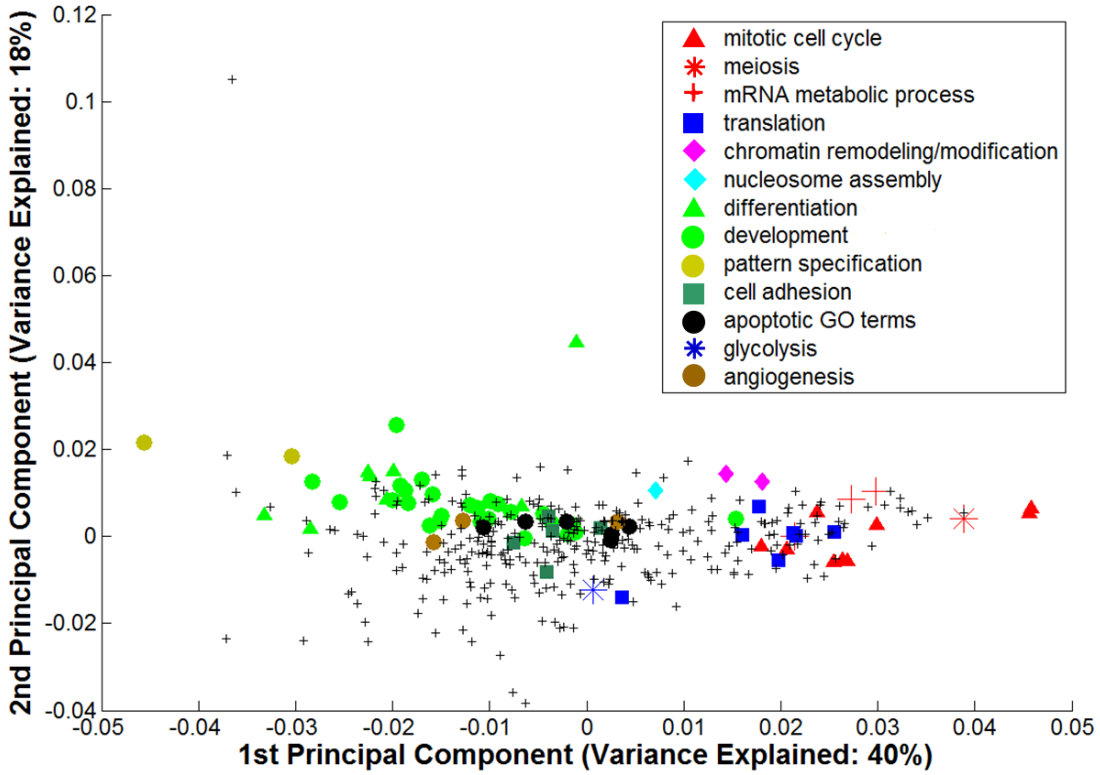


Figure 5: Differential codon usage of the various gene categories. Principal component analysis (PCA) for 395 GO terms, each containing at least 40 genes and are characterized by a 61 long vector corresponding to the average codon usage of its constituent genes. Each dot represents a GO term (by biological process). Several sets of GO terms are specified and marked : mitotic cell cycle (10 terms, including M and S phases, and M/G1, G2/M, and G1/S transitions); development (26 terms, including liver, brain, lung, heart, embryo and epidermis development); differentiation (8 terms, including cell differentiation, osteoblast differentiation, epithelial cell differentiation and neuron differentiation); translation (7 terms, including post-translational protein modification, translational initiation, elongation and termination); apoptotic processes (6 terms); nucleosome assembly; chromatin remodeling \ modification (2 terms); mRNA metabolic process (3 terms); glycolysis; angiogenesis (3 terms); cell adhesion (5 terms), and pattern specification (2 terms).

In order to ensure that the differential codon usage is not derived from differences in either the amino acid composition or the GC content between the genes that belong to the various GO categories, we ran PCA while normalizing codon usage to the amino acid usage, and also ran separate analyses for subsets of codons, grouped by the GC content of their nucleotides. Furthermore, we repeated the analysis while normalizing the codon usages to the usage of synonymous codons with identical GC content. Running all the above controls, we realized that the differential codon usage of "proliferation" and "differentiation" may be partially affected by the amino acid composition and GC content, but even after controlling for these potential confounding factors a clear and robust separation is still seen in the codon usage of the proliferation and differentiation genes. This is illustrated in Fig 6 which compares the codon usage of genes belonging to these two types of biological processes. Clearly for most amino acids there is a distinct preferred codon when that amino acid appears in genes related to differentiation or proliferation.

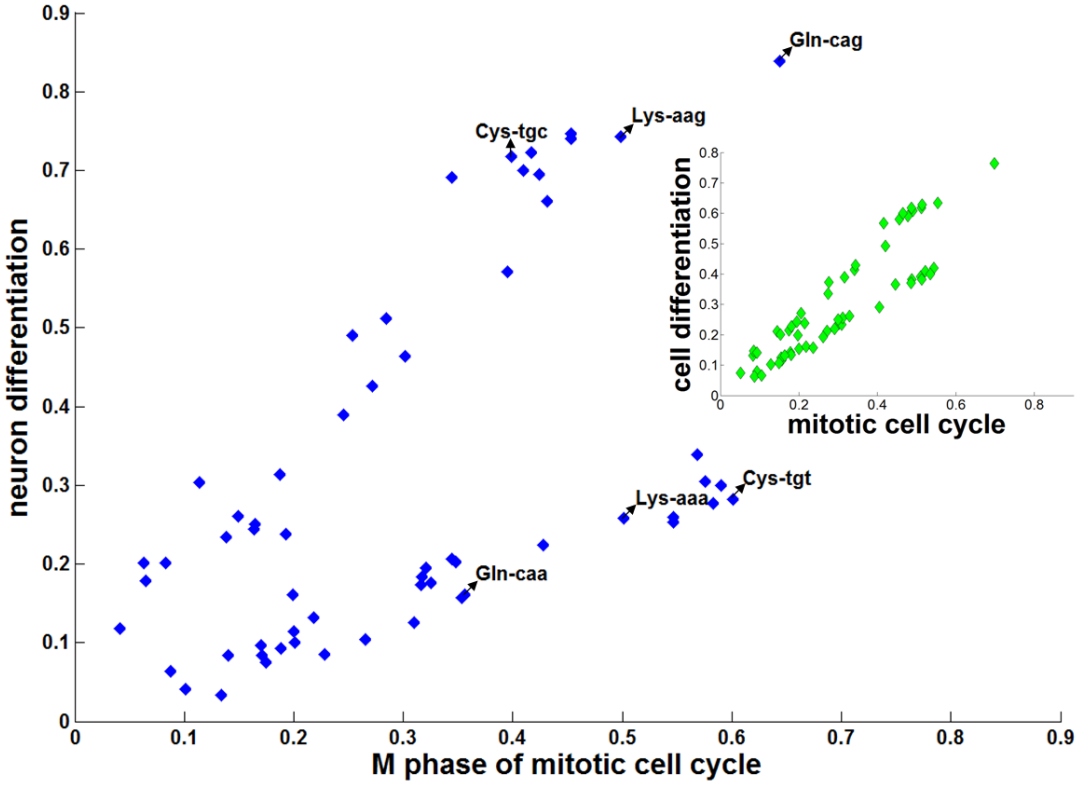


Figure 6: Differential codon usage of proliferation and differentiation related gene categories. Each blue dot represents the median codon usage of genes belong to GO terms of "M phase of mitotic cell cycle" (x-axis, 92 genes) or "neuron differentiation" (y-axis, 82 genes); the codon usage values are normalized to the amino acid usage. The same trend is still observed at a higher hierarchical level of the GO (inset), observed when grouping together genes belonging to the "cell differentiation" (506 genes) and "mitotic cell cycle" (301 genes) categories.

We observed a similar separation of proliferation and differentiation codon usage in other species too, including mouse and fly (Figure 7).

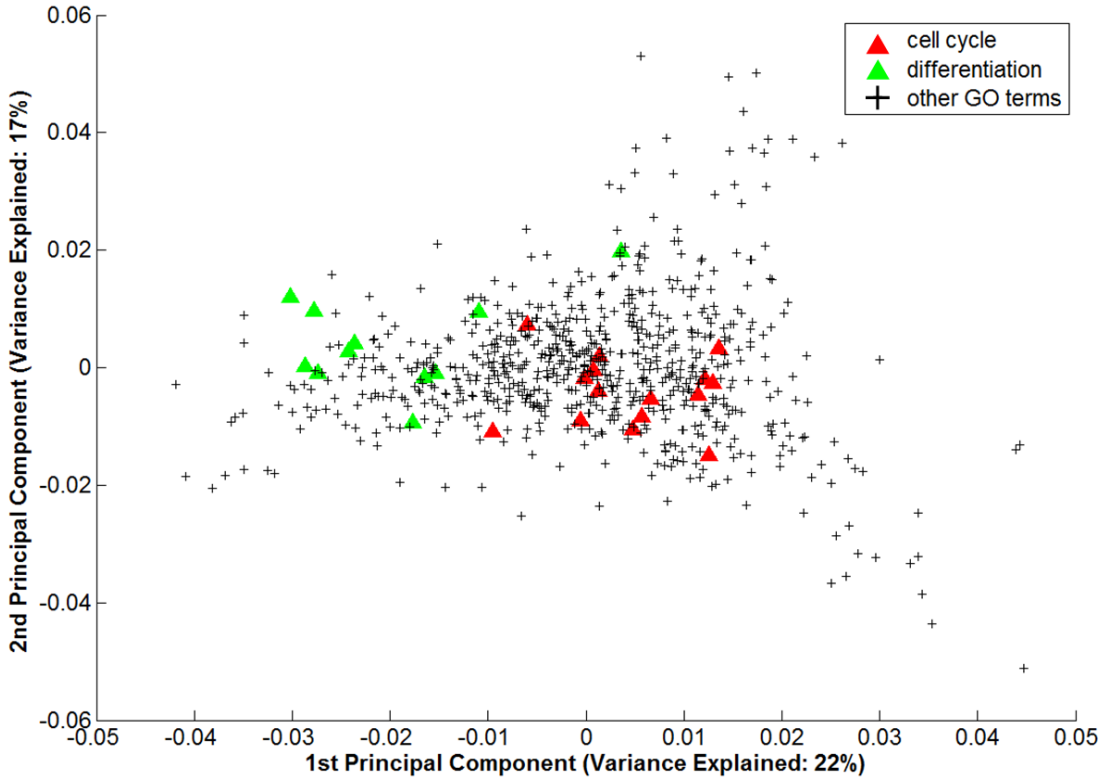
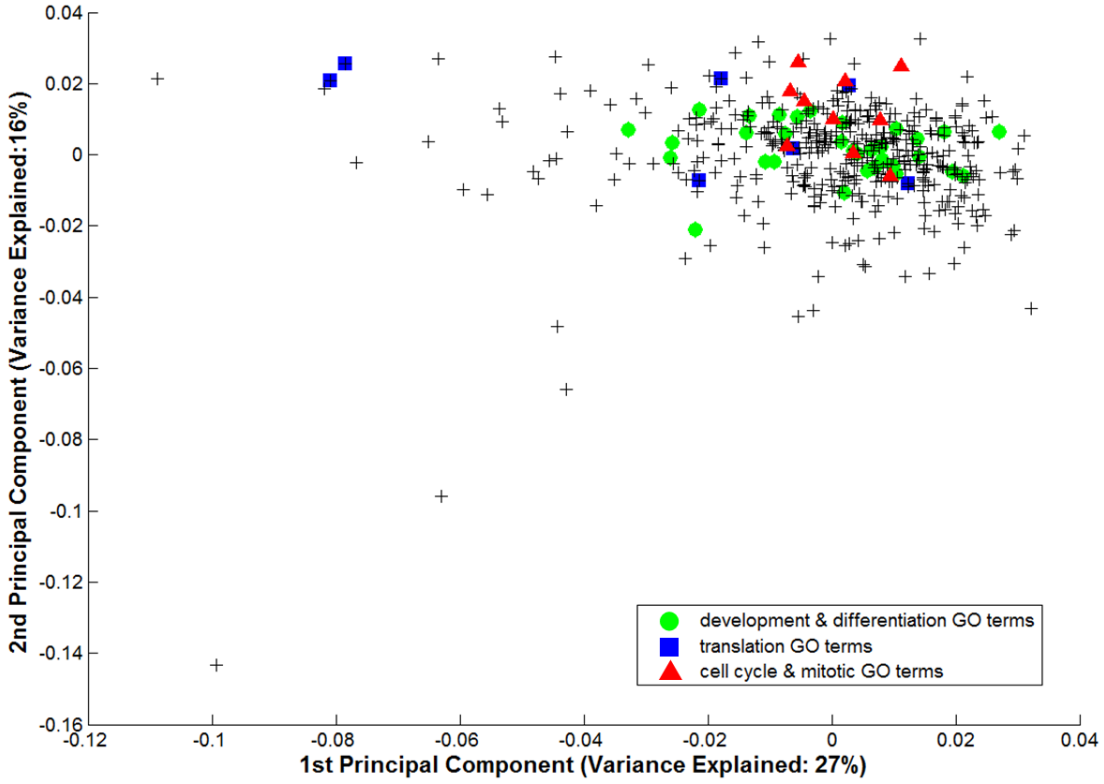


Figure 7: Differential codon usage of proliferation and differentiation genes in *D. melanogaster*. Principal component analysis (PCA) for 747 GO terms, each containing at least 8 genes and are characterized by a 61 long vector corresponding to the average codon usage of its constituent genes. Each dot represents a GO term (by biological process). Two sets of GO terms are specified and marked: cell cycle GO terms (15 terms), and differentiation (11 terms).

6.2.6 The codon usage modularity is associated with the existence of a proliferation-differentiation dichotomy in species

Our results reflect the classical dichotomy between proliferation and differentiation at the so-far not reported level of translation regulation. In order to examine whether the modularity of the genome codon usage is indeed associated with the interplay between proliferation and differentiation, we analyzed the codon usage of *C. elegans*. Fully developed adult *C. elegans* hermaphrodites consist of 959 non-dividing somatic cells. As somatic cells in the adult worm do not divide (to date there are only examples for endoreduplication of nuclear DNA within a few hypodermal cells – (Flemming et al. 2000)), there is a unique opportunity to examine the nature of

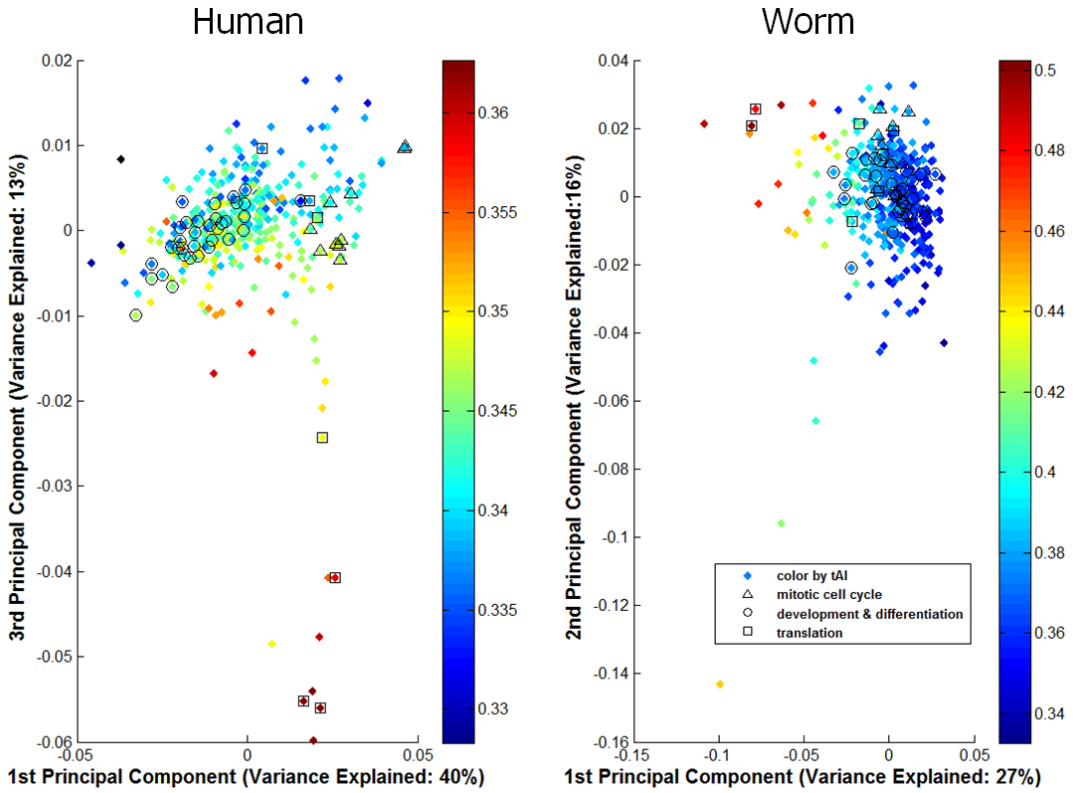
the codon usage in a species that does not feature a "switching" between proliferation and differentiation modes. We calculated the average codon usage of nematode's gene sets belonging to different GO terms, and ran a principal component analysis. Interestingly, contrary to the situation in human, mouse and fly, we observe no distinction of codon usage of proliferation and differentiation gene sets (Figure 8A). This result suggests that the strategy of distinction in codon usage is not used in a species in which proliferation and differentiation processes are not separated during life.



<u>Figure 8A</u>: Similarity in the codon usage of proliferation and differentiation genes in *C. elegans*. Principal component analysis (PCA) for 461 GO terms, each containing at least 5 genes and are characterized by a 61 long vector corresponding to the average codon usage of its constituent genes. Each dot represents a GO term (by biological process). Several sets of GO terms are specified and marked: cell cycle and mitotic GO terms (10 terms); development and differentiation (33 terms), and translation (7 terms).

What property of genes is spanned by the 1st PC in this species? We colored the various gene categories according to the median tAI value of the genes that belong to them (Figure 8B, left panel). Clearly the first PC corresponds to this measure of translation efficiency. It thus appears that in the worm expression level is the main biological attribute that affects codon usage. In contrast, in the vertebrates the first PC

appears to separate according to a different property – the uni-cellularity vs. multi-cellularity. Yet we noticed that in human, higher PCs, the 2nd and especially the 3rd do capture expression level, e.g. as approximated by tAI (Fig. 8B, right panel).



<u>Figure 8B</u>: Different biological attributes affect the codon usage *H. sapiens* (left panel) and *C. elegans* (right panel). Each plot shows Principal component analysis (PCA) representing GO terms (by biological process). Each dot is colored by the absolute translation efficiency (by the tAI measure). In both cases shown are the median translation efficiency values of the GO categories' constituent genes. While in the worm the first PC capture expression level (as measured by tAI), in human the first PC capture the distinction between uni-cellularity and multi-cellularlity, while the 2nd (not shown) and particularly the 3rd PC capture expression levels. GO terms of mitotic cell cycle are marked with triangle, GO terms associated with differentiation and developments are marked with circle, and translation-related GO terms are marked with square.

Further we examined the yeast *S. cerevisiae*. Fig 8C shows a PCA for the codon usage of the genes in the various GO categories in this species. Although yeast feature several developmental programs, including sporulation and pseudo-hyphal growth, we did not observe any specific codon usage for these genes. Instead, and in similarity to the worm, in this species the first PC appears to capture expression levels and in particular genes with high expression levels, such as the ribosomal proteins and the glycolytic enzymes that have a distinct codon usage, separated from the rest along the first PC (Fig. 8C)

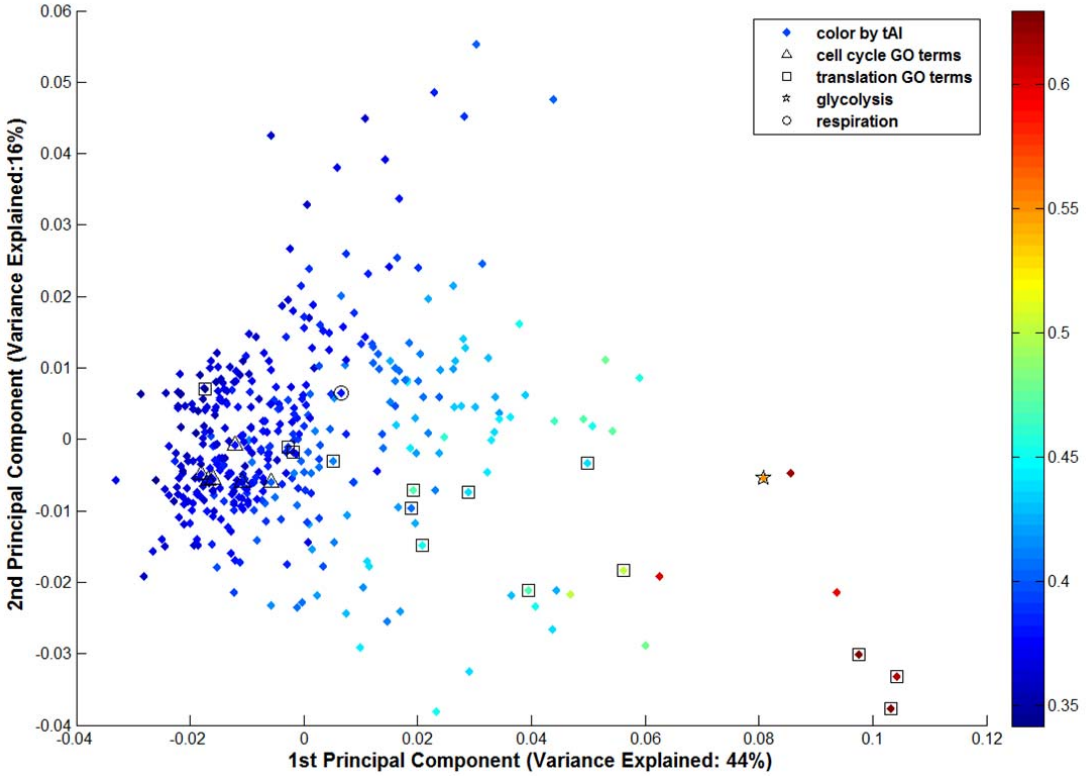
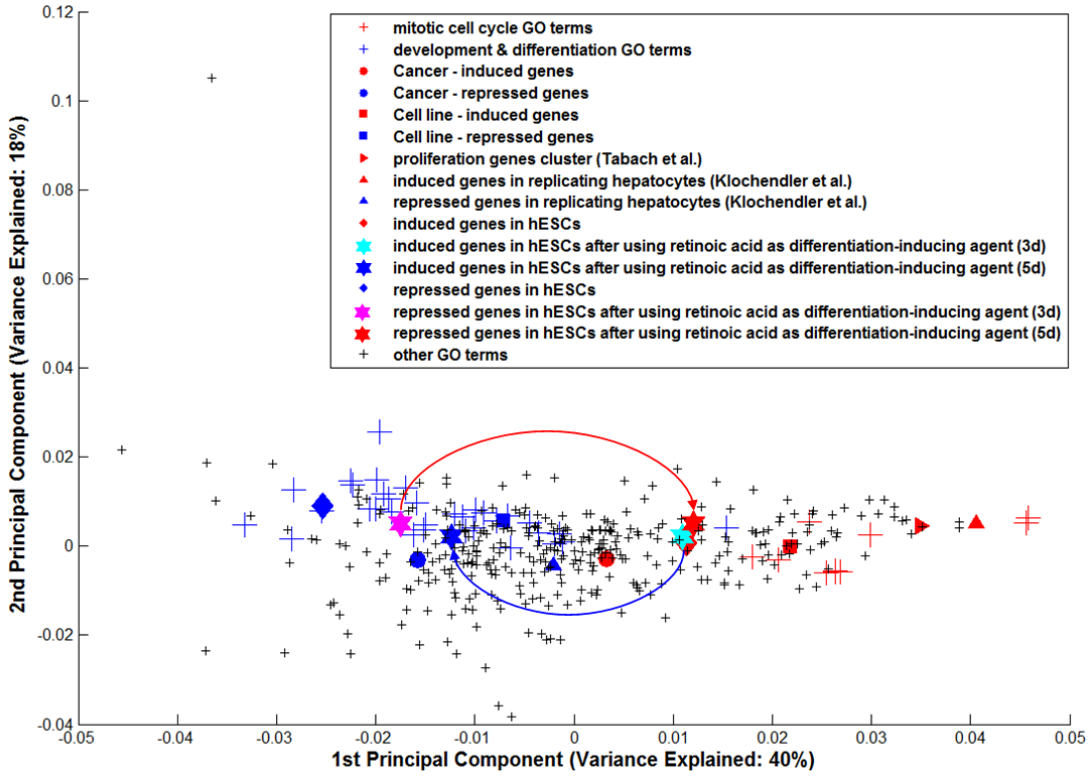


Figure 8C: The codon usage of highly expressed genes in yeast varies from that of rest of the genes. Principal component analysis (PCA) for 410 GO terms, each containing at least 10 genes and are characterized by a 61 long vector corresponding to the average codon usage of its constituent genes. Each dot represents a GO term (by biological process). Each dot is colored by the absolute translation efficiency (by the tAI measure). Shown are the median translation efficiency values of the GO categories' constituent genes. GO terms of cell cycle are marked with triangle, glycolysis is marked with pentagram, respiration is marked with circle, and translation-related GO terms are marked with square.

6.2.7 The modular design of the genome codon usage might allow cancer to hijack the translation machinery

We suggest here a so-far unrecognized separation between the codon usage of distinct biological processes. Yet, at any given time throughout organism's life, various combinations of biological processes occur simultaneously. An intriguing question is whether different environmental and conditional regimes are characterized with distinct codon usage patterns. To answer this question, we added to the PCA projection selected gene sets of the 100 most up-or down-regulated genes along the transition of human embryonic stem cells from proliferation to differentiation. The data was obtained while human embryonic stems cells were differentiated with retinoic acid and followed over a five days period. As can be seen (Figure 9), the

codon usage of the Transcriptome “migrates” along the first PC: the genes expressed at the early time points of the differentiation process have a codon usage that is reminiscent of the proliferation GO categories, yet the genes that are induced at 5 days into differentiation show a codon usage that is in the vicinity of the differentiation GO categories.



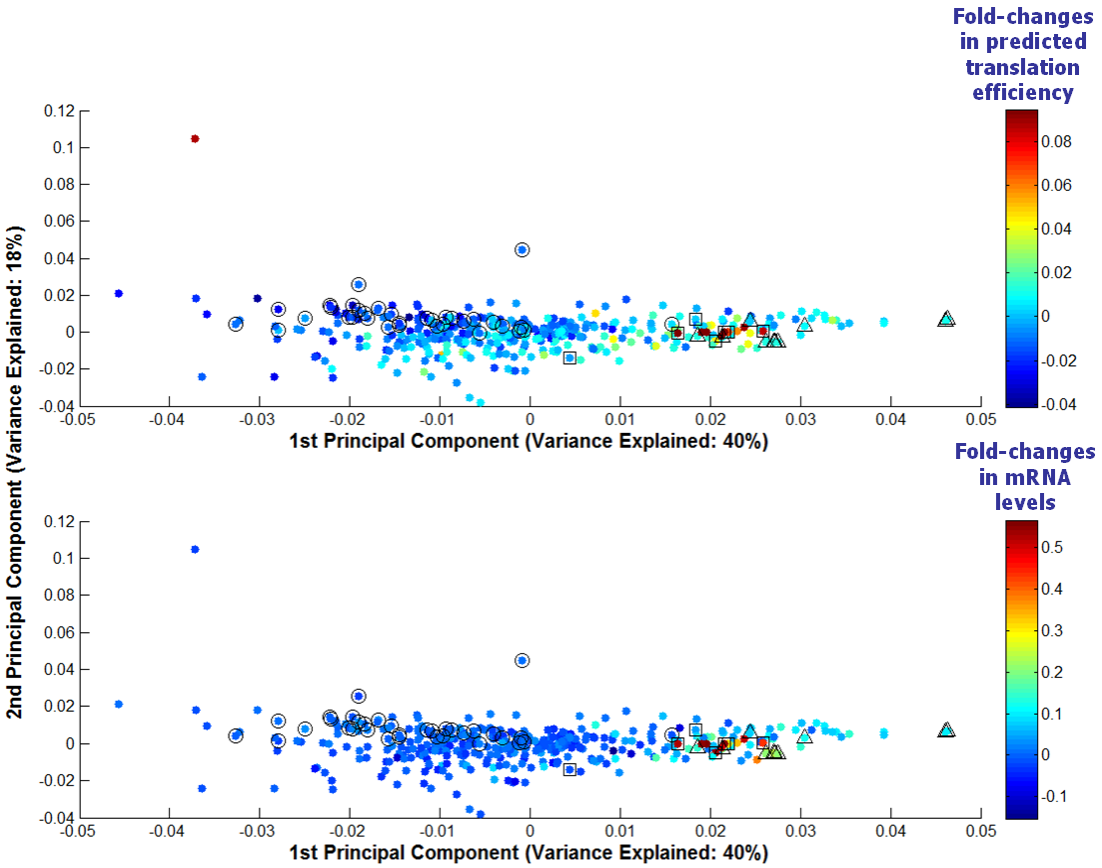
<u>Figure 9</u>: The codon usage of genes which are induced in cancer differs from that of genes which are repressed upon cancer. Principal component analysis (PCA) for 395 GO terms and 15 sets of induced or repressed genes in primary tumors, cancerous cell lines, human embryonic stem cells (hESCs), and hESCs after using retinoic acid as differentiation-inducing agent. The sets of induced and repressed genes in cancer contain 50 and 28 genes, respectively; the sets of induced and repressed genes in cancerous cell lines contain 83 and 58 genes, respectively. Published gene sets: human orthologs of up-and-down regulated mouse genes in replicating hepatocytes in vivo (Klochendler et al. 2012) – 584 and 383 genes, respectively; A gene expression profile in in-vitro transformed normal human fibroblasts, represented by 168 genes that have been termed "the proliferation cluster" (Tabach et al. 2005). Arrows indicate the changes in the sets of induced (blue arrow) and repressed (red arrow) genes – and the consequent change in the codon usage – along time course of 5 days in which hESCs undergo differentiation after using retinoic acid as differentiation-inducing agent.

In the previous section, we hint for genome-embedded distinction between uni-and multi-cellular functions. To our point of view, cancer may be considered as a shift from coordinated multi-cellular behavior towards a 'selfish' unicellular-like life style. Based on the mRNA levels from our arrays, we defined sets of induced and repressed genes for both primary tumors and cancerous cell lines. Specifically,

induced genes are genes which are among the 100 most up-regulated genes in at least two out of four examined cancer types (compared to normal cells from the corresponding tissue). The sets of repressed genes were determined in a similar manner, i.e. the 100 most down-regulated genes. Reassuringly the genes induced in cancer are localized in the vicinity of the proliferation and the Central Dogma processes genes, whereas the cancer down-regulated genes span the codon usage space region which is associated with multi-cellularity. Consistent with our results, published sets of proliferation-related genes reside in the Central Dogma vicinity of the codon usage space (Figure 9).

Ilustrating the association between the modular design of the genome codon usage and the prevailing life style of the cells, we next aimed to more directly examine the potential contribution of this correspondence. We realized that if the genome codon usage indeed serves as a platform for translation regulation, it requires a parallel involvement of the genomic tRNA pool. In order to understand the interplay between the cancerous tRNA pool and the differential codon usage of genes categories, we superimposed on the PCA plot the median change in the predicted translation efficiency of the genes belonging to each GO category. Figure 10 shows such analysis for the case of Colon carcinoma. As can be seen, the pattern of change in the predicted translation efficiency of different GO categories is associated with the distribution of genes categories along the PCA, as governed by their codon usage. This result indicates that cancer predominantly increase expression of tRNA whose codons are preferred among the proliferation genes and represses the expression of tRNAs whose codons are enriched among the differentiation genes. In parallel, we superimposed on the PCA the information about the change at the mRNA level of the genes in each GO category. Similarly to the case of translation efficiency, there is a correspondence between the fold-changes in mRNA levels in cancer and the differential codon usage of the various gene categories. All together, our results suggest that the same gene sets that are up-regulated at the mRNA level in cancerous cells are also expected to be translated more efficiently.

We repeated the analyses on all types of primary and cell line cancerous cells and found it to recur in all, yet the phenomenon was more pronounced among the cell lines. We hence suggest that the translation machinery may be recruited by cancerous cells, via dynamics in the tRNA pool, to support 'switching' in the cell lifestyle towards a proliferative mode.

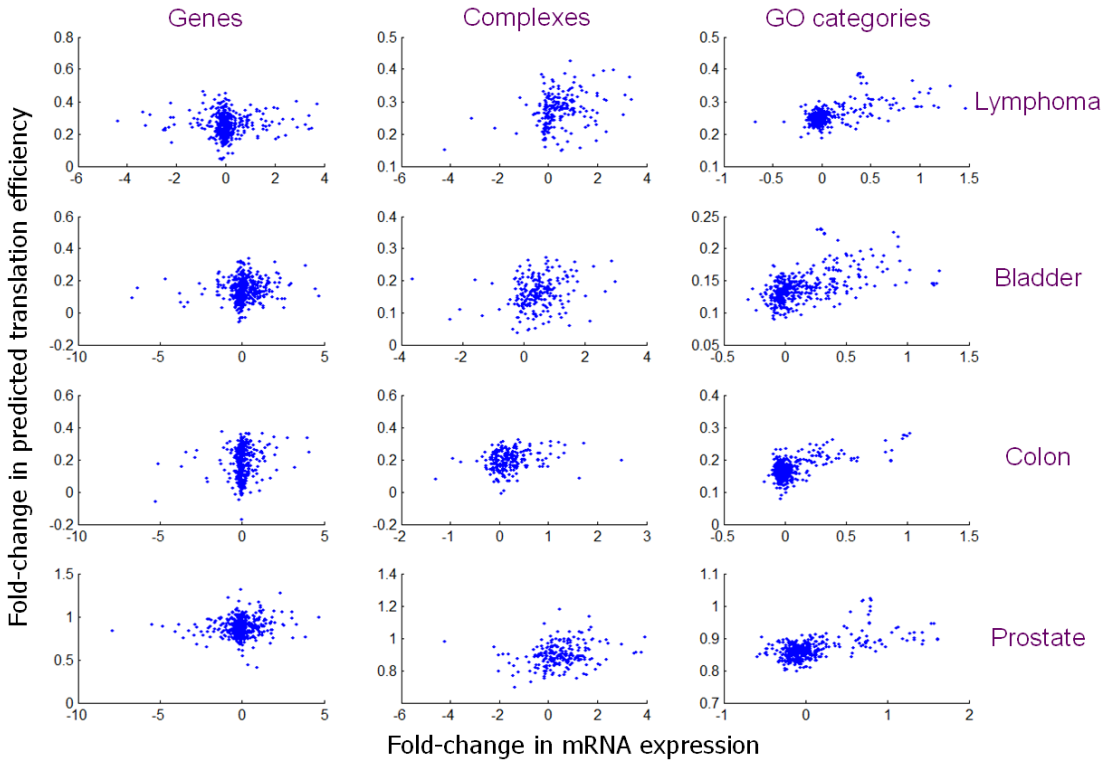


<u>Figure 10</u>: Changes in predicted translation efficiency and mRNA levels in cancer follow the codon usage signature of genes. Each plot shows Principal component analysis (PCA). Each dot represents a GO term (by biological process). GO terms of mitotic cell cycle are marked with triangle, GO terms associated with differentiation and developments are marked with circle, and translation-related GO terms are marked with square. Each dot is colored by either the (log2) change in the predicted translation efficiency (upper panel) or the (log2) change in mRNA levels (lower panel). In both cases shown are the median fold-changes of the categories' constituent genes; translation efficiency values were further normalized by dividing each individual score by the GOs averaged score. The data is this figure is derived from colon carcinoma patients; similar results are obtained for most other tumors and cell lines.

6.2.8 Adaptive changes in the cancerous tRNA pool may promote proliferation

We next wanted to formally check if predicted changes in translation efficiency correlate with changes in mRNA abundance. Particularly we asked whether the sets of genes that show predicted increase in translation also show an up-regulation at the mRNA levels in cancer. For that purpose we calculated the correlation between the fold-changes in translation efficiency to the fold-changes in mRNA abundance at three levels – the level of individual genes, the level of protein complexes - examining 223 complexes that are classified under the "Cellular

Component" GO categories, and the level of functional gene sets defined as GO categories, the “Biological Processes” classification (Figure 11). At the level of individual genes, we found no significant correlation between fold-changes in translation efficiency and fold-changes in the mRNA levels. We next examined 223 complexes, which at least two of their genes are represented in the microarrays. The fold-changes of each complex were defined by the median fold-change of the genes that it contains. Here we detected modest correlation between changes in translation efficiency and changes in mRNA level for the cell lines of all the four cancer types (diffuse large B-cell lymphoma – Pearson correlation = 0.31, p-value = 2.5 x 10^{-6}; Colon cancer – Pearson correlation = 0.24, p-value = 2.6 x 10^{-4}; Prostate cancer – Pearson correlation = 0.24, p-value = 3.0 x 10^{-4}; Bladder cancer – Pearson correlation = 0.23, p-value = 6.5 x 10^{-4}). Finally, we calculated the correlation at the level of GO categories (figure 11).



<u>Figure 11</u>: Changes at predicted translation efficiency correlate with mRNA changes in various cell line types. In each of the plots, each blue dot represents individual gene (left column), a complex (middle column) or GO term (right column). The x-axis shows the (log2) change in the mRNAs level, where the y-axis shows (log2) change in the predicted translation efficiency (by gene-tAI) of these genes. Fold-changes of individual genes were averaged over the respective samples number (diffuse large B-cell lymphoma – 1; Bladder cancer - 15; Colon cancer – 10; Prostate cancer – 5). For both complexes and GO terms, the fold-changes are given by the median fold-change of their constituent genes (each is represented in the array with at least 25 genes).

At this level of GO categories, we observed higher positive correlation between the variation in translation efficiency and the variation in mRNA level for all the examined cell lines (diffuse large B-cell lymphoma – Pearson correlation = 0.57; Colon cancer – Pearson correlation = 0.63; Prostate cancer – Pearson correlation = 0.57; Bladder cancer – Pearson correlation = 0.52. All associated p-values are lower than 10^{-300}. At the level of GO categories, we also detected strong positive correlation (Pearson coefficient = 0.73) between the variation in predicted translation efficiency and the variation in mRNA levels for samples from primary colon carcinoma, yet in the other primary tumors we did not observe significant correlations. Our results hence imply that the cancerous tRNA pool may contribute to proliferation in an adaptive manner, i.e., via promotion of proliferation-related biological processes.

6.2.9 Changes in the canceruos tRNA pool resemble short term changes that occur when normal cells proliferate and are revered from changes that occur during differentiation

To further examine the potential role of the tRNA pool in directing the status of cells, we analyzed expression data of cells upon different physiological conditions that involved proliferation and cell arrest upon starvation, in addition to differentiation introduced above. Interestingly, we detected a significant negative correlation (Pearson correlation coefficient = -0.55, p-value = 9.8×10^{-4}) between the variation in the tRNA pool of cells upon proliferation, (following over-expression of MYC), and the changes in the tRNA pool five days after induction of differentiation (by retinoic acid as a differentiating agent) – see Figure 12A.

We further examined the variation in the tRNA pool of hESCs along five days after induction of differentiation and compared it to the alternation in the tRNA pool of cells upon other physiological states (Figure 12B). Curiously, during the transition from earlier stages of differentiation to its latest time point, the tRNA pool of the differentiating cells gradually shifted from a "proliferation-like" tRNA pool, to a "cell arrest-like" tRNA pool. In particular, during the five days of the differentiation the tRNA pool switched from resembling the tRNA pool of proliferative cell that over-expressed the oncogene RAS for 24 hours, to being similar to the tRNA pool of cells upon arrest of proliferation due to serum starvation.

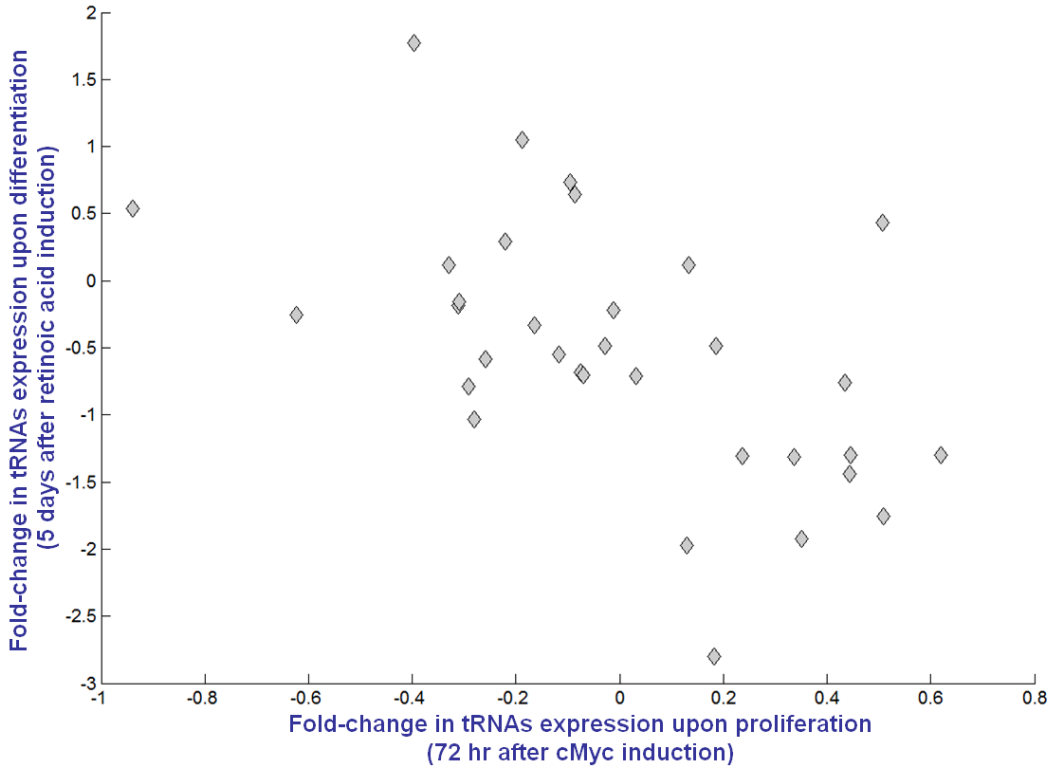


Figure 12A: Anti-correlated variation in the tRNA pool during proliferation and differentiation. Each dot represents a tRNA type. The x-axis shows the (log2) change in the tRNAs level of human fibroblasts transduced with cMyc for 72h (compared to cells transduced with control virus (pBabe)). The y-axis shows (log2) change in the tRNAs level of hESCs after using retinoic acid as differentiation-inducing agent.

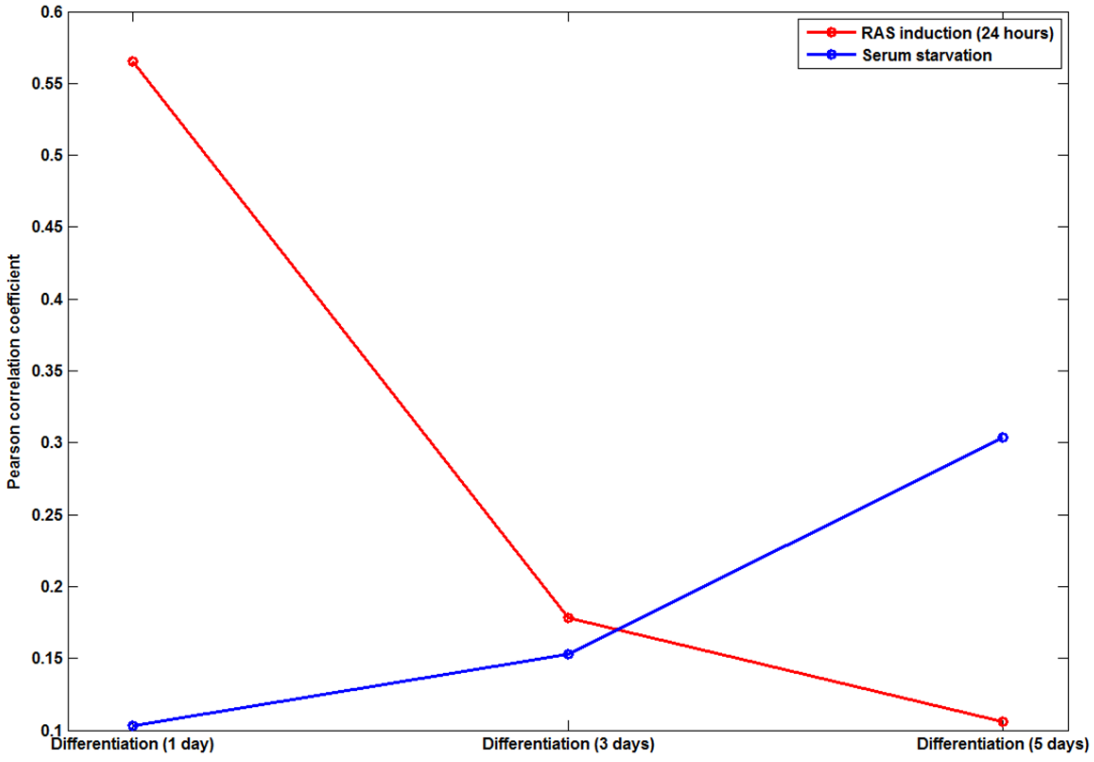
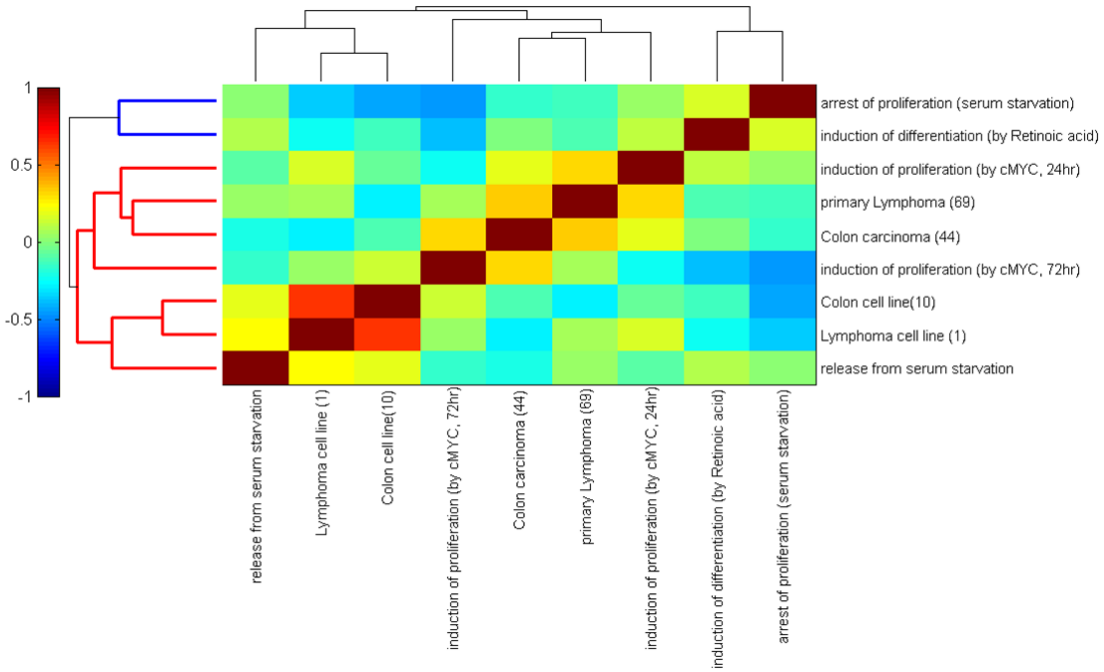


Figure 12B: The correlation between the tRNA pool at various days during differentiation of stem cells to the tRNA pools observed in proliferative cells (Ras over-expression) and in arresting (serum-starved) cells. Shown are Pearson correlation values. Comparisons were done between human embryonic stem cells after 1, 3 and 5 days of differentiation using retinoic acid as differentiation-inducing agent, and human fibroblasts over-expressing the oncogene RAS for 24 hours, as well as human fibroblasts that have been serum-starved for 70 hours.

Does the cancerous tRNA pool resemble the tRNA pools of proliferating cells? To answer this question, we clustered cancerous cells (primary cancer and cell line) and cells upon different physiological conditions based on the similarity in the variation in the cellular tRNA pool (Figure 13). As can be seen, the fold-changes in the tRNAs expression cluster cancerous cells together with proliferating cells, and apart from both differentiating cells and cells upon proliferation arrest.



<u>Figure 1</u>3: tRNAs expression profiles cluster together differentiated and starved cells and separately from cancerous and proliferating cells. This analysis compares between the averaged fold-changes in tRNAs expression of nine different cell types - samples of primary Lymphoma and Lymphoma-associated cell line (compared to the normal cells of the same tissue); Samples of Colon carcinoma and Colon carcinoma-associated cell line (compared to the normal cells of the same tissue); human fibroblasts over-expressing the oncogene cMyc for either 24h or 72 hr (compared to cells transduced with control virus); human embryonic stem cells (hESCs) after using retinoic acid as differentiation-inducing agent; Human fibroblasts that have been serum-starved for 70 hours, and such starved cells 4 hours after re-addition of serum. Hierarchical clustering was performed with 1-Pearson Correlation as a distance metric.

We then asked whether the pattern of similarity or dissimilarity observed at the tRNA level is recapitulated at the level of genes' translation efficiency. To answer this question, we analyzed the correlation between the predicted fold-changes in translation efficiency in cancerous cells to the predicted fold-changes in translation efficiency upon different physiological conditions that either induce proliferation and differentiation, or that are arrest proliferation, e.g. due to starvation (Figure 14).

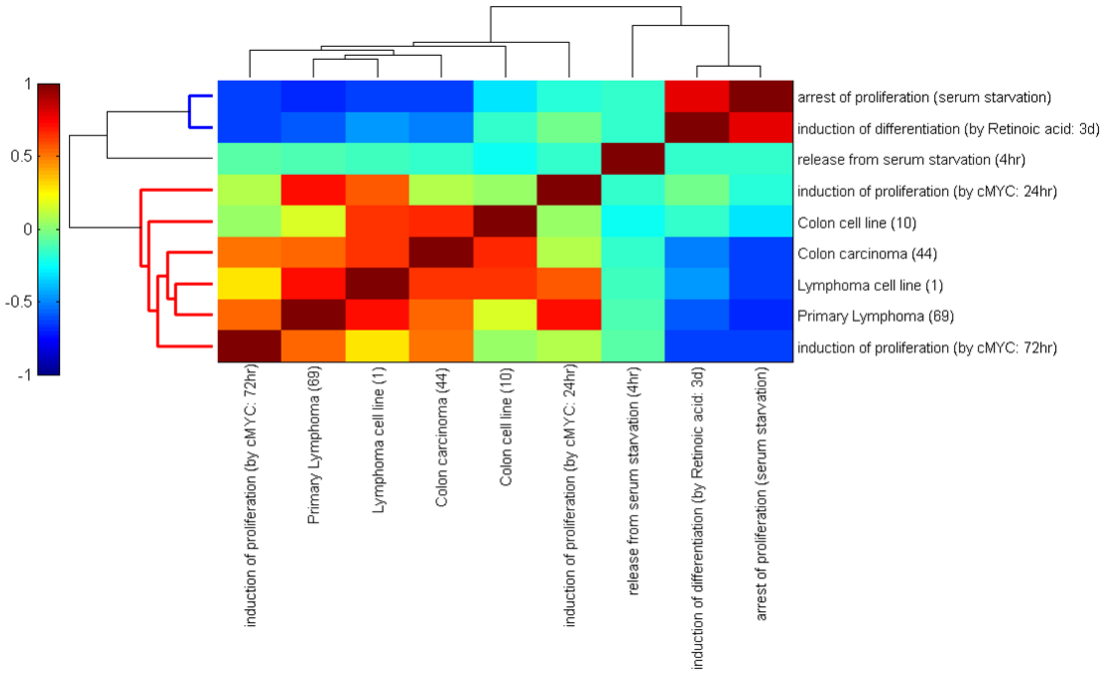


Figure 14: Changes in predicted translation efficiency cluster together differentiated and starved cells and separately from cancerous and proliferating cells. This analysis compares between the changes in the predicted translation efficiency of GO terms, which contain at least 40 genes, in various cell types. For a given GO term, the fold-changes were determined as the median fold-change of its constituent genes. Shown are nine different cell types - samples of primary Lymphoma and Lymphoma-associated cell line (compared to the normal cells of the same tissue); Samples of Colon carcinoma and Colon carcinoma-associated cell line (compared to the normal cells of the same tissue); human fibroblasts over-expressing the oncogene cMyc for either 24h or 72 hr (compared to cells transduced with control virus); human embryonic stem cells (hESCs) after using retinoic acid as differentiation-inducing agent; Human fibroblasts that have been serum-starved for 70 hours, and such starved cells 4 hours after re-addition of serum. Hierarchical clustering was performed with 1-Pearson Correlation as a distance metric.

The picture that emerges when examining similarity between samples by the extent of predicted changes in translation efficiency is that the various proliferative processes, cancerous or non-cancerous, cluster together, and away from the differentiation and cell arrest condition, which in turn co-cluster. Figure 15 illustrates the high similarity (Pearson correlation = 0.7) between the predicted changes in translation efficiency of distinct gene sets in cells induced for proliferation by over expression of cMyc after 24 hours and samples of patients with diffuse large B-cell lymphoma.

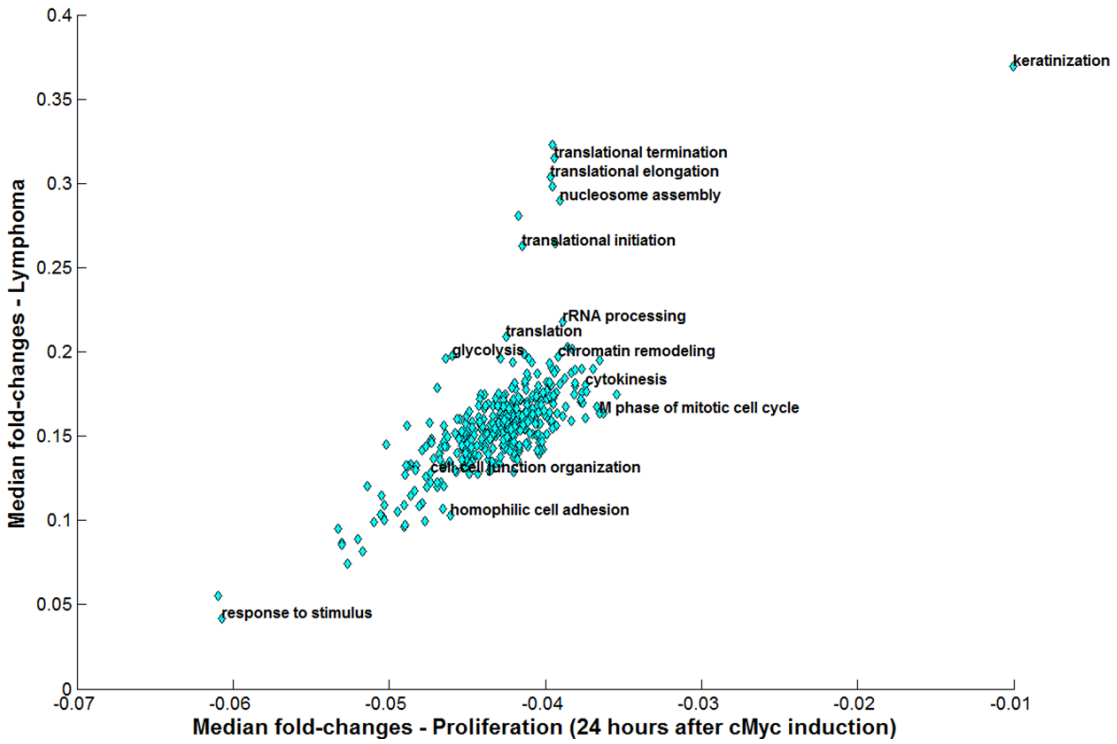


Figure 15: Functional categories change similarly their translation efficiency in cancer and upon non-cancerous proliferation. Each dot represents one GO category. The x-axis shows the median (log2) change in the predicted translation efficiency of the genes belonging to a given GO category in human fibroblasts over-expressing cMyc for 24 hours. The y-axis shows the median (log2) change in the predicted translation efficiency of the genes belonging to a given GO category in Lymphoma (averaged over samples of 69 patients).

Interestingly, the significant positive correlations, in terms of predicted translation efficiency, between cancerous cells and cells upon induction of proliferation are typically associated with similar changes in the transcriptome of the different cell types (Figure 16). For instance, the fold-changes in mRNA levels of cells transduced with cMyc after 24 hours are significantly correlated with diffuse large B-cell lymphoma (Pearson correlation = 0.32), and Bladder cancer (Pearson correlation = 0.26); the fold-changes in mRNA levels of cells transduced with cMyc after 72 hours are significantly correlated with Colon cancer (Pearson correlation = 0.77, see Figure 16). Consistent with this trend, changes in both predicted translation efficiency and mRNA levels in cancer are negatively correlated with those of cells upon serum starvation (Figure 16). All together, our results suggest a general program in which proliferation-related and differentiation-related conditions are using distinct types of tRNAs, and that the coding regions in turn use distinct and corresponding preferred codons for genes that belong to each of these types of conditions. This natural system appears to be hijacked by cancer which induces the

expression of the “proliferation tRNAs” while repressing the expression of the “differentiation tRNAs”. This change, along with a corresponding change at the mRNA level may contribute to malignancy of the cancerous proteome and transcriptome.

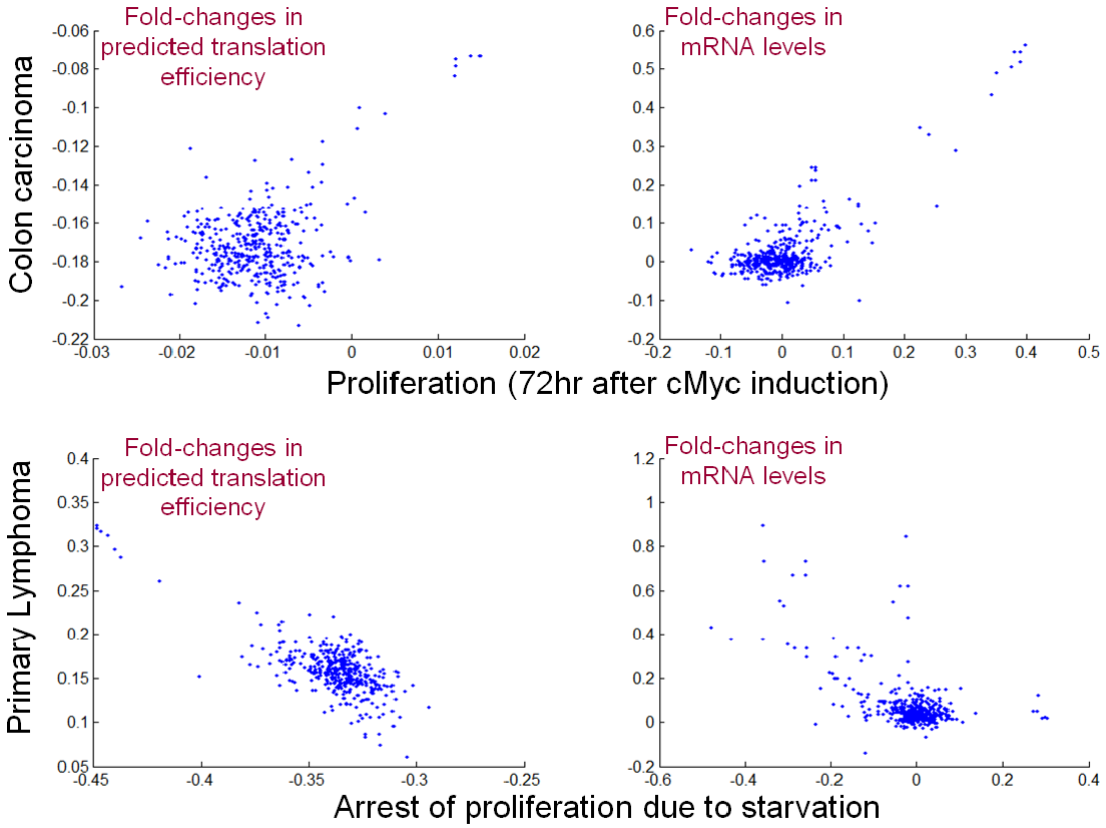


Figure 16: Changes in predicted translation efficiency and mRNA levels in cancer resemble those of proliferating cells. Each dot represents a GO term (by biological process), which contains at least 40 genes and is represented in the microarray by at least 25 genes. The x-and-y axes denote either the (log2) change in the predicted translation efficiency (left panels) or the (log2) change in mRNA levels (right panels). Both types of fold-changes are given by the median fold-change of their constituent genes; fold-changes of individual genes were averaged over sets of samples (Colon carcinoma – 44 samples; Primary Lymphoma – 69 samples), or over three biological replicates (for both over-expression of cMyc and serum starvation).

6.2.10 Potential regulation of tRNA expression at the level of histone epigenetics

In the previous sections we showed differential regulation of tRNA genes which are abundant in either proliferation or differentiation genes (henceforth termed as "pro-proliferation" and "pro-differentiation" tRNAs, respectively). In this section we ask what could be the mechanism responsible for the coordinated expression regulation of the various tRNAs, and what might account for the observed correlation

in expression that the pro-proliferation and pro-differentiation tRNAs show respectively with the proliferation-related and differentiation-related mRNAs. In particular we hypothesized that the epigenetic marks on histones in the vicinity of tRNA and mRNA coding genes might account for such coordination.

Recent studies indicated that a similar histone code operates on both mRNAs and tRNAs, i.e. that same histone modifications induce or repress transcription by both RNA polII and polIII (Barski et al. 2010; Oler et al. 2010). In the light of these studies, we examined whether the correspondence between up-regulated tRNAs and up-regulated mRNAs in cancerous cells is also reflected at the level of chromatin modification. For that purpose, we first defined sets of proliferation and differentiation mRNAs as the genes belonging to the GO categories of 'cell cycle' and 'cell differentiation', respectively. Second, we computed the averaged codon usage of each such gene set, and calculated the ratio between the proliferation and differentiation codon usages. Employing the codon usage ratio as a proxy of the demand for the various tRNA types, we next classified the tRNA types as "pro-proliferation tRNAs" (tRNAs whose corresponding codons are highly represented in proliferation-related genes compared to differentiation-related genes), "pro-differantiation tRNAs" (tRNAs whose corresponding codons are more abundant in differentiation-related genes relative to proliferation related-genes), and “other tRNAs” that do not belong to either category. 81 and 75 expressed tRNA genes belong correspondingly to the “pro-proliferation” and “pro-differentiation” tRNA sets (the distinction between expressed and silent tRNA genes is based on POLIII occupancy measurements (Oler et al. 2010)). Utilizing the data from the ENCODE project in human (Dunham et al. 2012), we extracted the read density of the H3K27ac and H3K9ac activating modifications in HeLa and Normal Human Lung Fibroblasts (NHLF) cells, and then plotted read density in the vicinity of the above mentioned distinct sets of mRNAs and tRNAs (Figure 17A). As can be seen, there is an association between the signature of the H3K27ac and H3K9ac activating modifications in a given set of mRNAs and in the corresponding set of tRNAs which are specifically involved in their translation. In particular the relatively high signal of H3K27ac and H3K9ac activating histone modifications in proliferation genes is accompanied with relatively high signal of these modifications in the "pro-proliferation" tRNA set, whereas both differentiation-related mRNAs and "pro-differentiation" tRNAs are characterized with relatively low read density of H3K27ac

and H3K9ac in HeLa cells. We repeated this analysis for NHLF cells (Normal Human Lung Fibroblasts), and did not observe such striking distinction (Figure 17A).

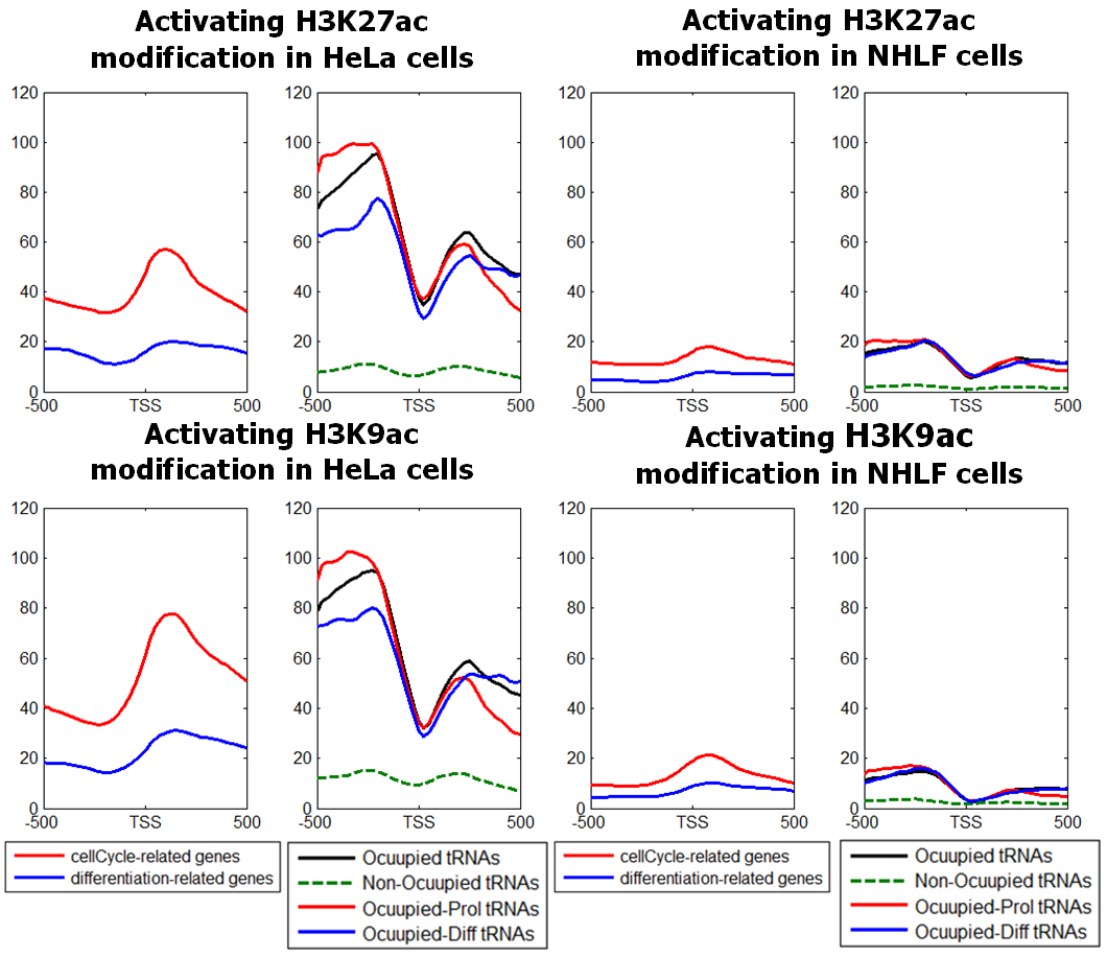


Figure 17A: Differential epigenetic marking on "pro-proliferation" and "pro-differentiation" tRNAs is associated with differential epigenetic signature of proliferation-and-differentiation mRNA-coding genes. The read density of the H3K27ac H3K9ac chromatin modifications is shown as a function of sequence positions for 1000 bps centered around the transcription start sites of mRNA and tRNA genes. The density of signal enrichment (for 25 base-pair intervals) was downloaded from UCSC and is based on ChIP-seq Signal from ENCODE. A given tRNA gene is defined as "occupied" if it is enriched with Pol III in at least one of the following cell types: Hela, human embryonic kidney HEK293T cells, human foreskin fibroblasts HFF cells, and Jurkat T cells (PMID:20418882). Shown are the averaged signals (y-axis) of the following gene sets: tRNAs that are not occupied by RNA polIII (180 genes; colored in green); all polIII-occupied tRNAs (299 genes; colored in black); occupied "pro-proliferation" tRNAs (81 genes; colored in red); occupied "pro-differentiation " tRNAs (75 genes; colored in blue); proliferation-related coding genes (372 genes; colored in red), and differentiation-related coding genes (494 genes; colored in blue).

This result implies that histone modification-based regulation might regulate distinctly the two sets of tRNAs, might change their supply in proliferating and differentiated cells, and might coordinate their availability with the demand – namely, mRNA expression. Such possibility is further sustained by the correspondence between the intensity of chromatin modifications and the distribution of gene

categories along the PCA, as governed by their codon usage (Figure 17B). Figure 17B demonstrates the extent of enrichment of the H3K27ac activating modification in Hela cells for genes belonging to different biological processes. Utilizing ChIP-seq Signal from the ENCODE data set, we scanned the human genes and looked for statistically significant signal enrichments which overlap the transcript and its 500bp up-and-down flanking regions; if more than one such region was found, we chose the one with the maximal signal. Then, for a given GO category, we calculated the frequency of genes with statistically significant signal enrichment, and the average score of the corresponding genes. As can be seen, the intensities of H3K27ac chromatin modification in HeLa cells follow the codon usage signature of genes.

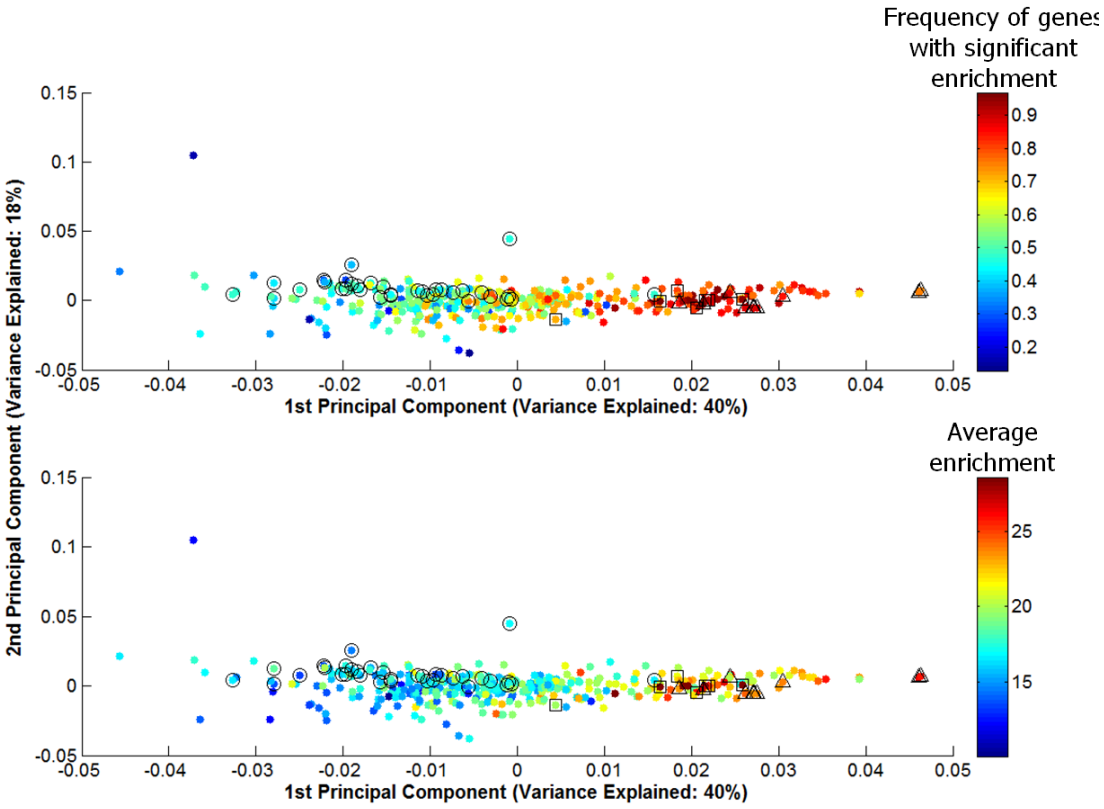


Figure 17B: The intensities of H3K27ac chromatin modification follow the codon usage signature of genes. Each plot shows the same Principal component analysis of codon usage as in above figures. Each dot represents a GO term (by biological process). GO terms of mitotic cell cycle are marked with triangle, GO terms associated with differentiation and developments are marked with circle, and translation-related GO terms are marked with square. Each dot is colored by either the frequency of genes with statistically significant signal enrichment in a given GO (upper panel), or by the average enrichment of the categories' constituent genes for which a significant enrichment was found (lower panel). Regions of statistically significant signal enrichment were downloaded from UCSC and is based on ChIP-seq Signal from ENCODE. The data in this figure is derived from HeLa cells.

6.3 Discussion

In this study we showed that coordinated changes in the tRNA pool and the codon usage of the transcriptome may serve as a translational regulatory strategy to promote "switching" in the status of cells, towards either proliferation or differentiation mode. We suggest that the feasibility of such mechanism is anchored by two main characters – the so far unrevealed distinct codon usage of distinct gene sets, and the dynamic nature of the tRNA pool. The potential causality relation between these two attributes should be further understood.

6.3.1 Dynamics in the tRNA pool upon physiological and cancerous processes

The traditional notion of constant tRNA pool throughout the life of the cell was recently challenged by evidence for variation in the availability of the different tRNA types in yeast in response to varying metabolic conditions (Tuller et al. 2010). Variation in the expression patterns of tRNAs was also observed while comparing between distinct cell types in human (Dittmar et al. 2006), and yet, for a given cell type of multi-cellular organism, there are no evidences for alternations in the tRNA pool.

In this study, we provide strong evidence for dynamics in the human tRNA pool upon proliferation and cancerous processes on one hand, and upon differentiation and cell arrest on the other hand, suggesting that the same gene in the same cell type may be translated differently in these different conditions. Specifically, we identified three major trends in the dynamics of the human tRNA pool.

First, we noticed reoccurring changes in the expression of tRNAs in primary tumors representing different cancer types, or cancerous cell lines from different origins. Such a global tissue-independent signature of variation in the tRNA pool may promote translation efficiency of genes involved in leading pathways of cancer, for instance - rapid cell division and elevated glycolysis (Vander Heiden et al. 2009). Considering the variation in the tRNA pool among human tissues (Dittmar et al. 2006), the internal partitioning within the cluster of primary tumors (figure 3) may be associated with variation between the tRNA pools of distinct normal tissues.

Second, we found that the variation in the tRNA pool of cancerous cells resembles that of normal cells upon induction of proliferation. This result suggests

that cancerous processes do not lead to a new state of the translation machinery, but rather "hijack" a naturally occurring state. Interestingly, clustering of various cell types based on their tRNA expression profiles indicate a clear distinction between the tRNA pools of cancerous and proliferating cells, and the tRNA pool of both differentiating cells and cells upon arrest of proliferation (figure 13). Furthermore, we show that the changes in the tRNA pool upon induction of proliferation are negatively correlated to the changes in expression of the tRNA pool that differentiation triggers. All together, our results reveal that the well-known proliferation-differentiation dichotomy is strikingly reflected at the level of the translation machinery.

Third, we observed distinguishable patterns of tRNA expression changes for primary tumors and cancerous cell lines, although both are expected to share physiological and molecular attributes. One may explain this result by the genetic differences between primary tumors and laboratory cell lines, which accumulate new mutations as they adapt to their artificial environment. Reasonable as it is, such an interpretation can explain the variation between a given primary tumor and its corresponding cell line, and yet cannot explain the similarity between the tRNA pools of distinct cell lines which grow in different environments. To our opinion, the cell line samples more accurately reflect the actual proliferative state of the examined cells - primary tumors might contain non-cancerous cells, leading to heterogeneous cellular, genetic and epigenetic composition, which might mask the actual proliferation rate in the cancerous cells. Consistent with this notion is the more pronounced coordination between the changes in mRNA levels and inferred changes in translation in cell lines compared to primary tumors. Together with the notion that cell lines typically proliferate much faster than cells in primary tumors (Dairkee et al. 2004) our results indicate that the tRNAs' expression might hold a unique information about the proliferation status of cells. Interestingly, we also observed striking differences between primary cancerous cells and cell line at the level of translation efficiency (figure 4). While genes associated with glycolysis and gluconeogenesis are among the genes which are expected to have the most increased translational efficiency in primary cancerous cells, respiratory genes are predicted to be translated more efficiently in cell lines. This difference might suggest that Warburg effect, the tendency of cancer to prefer fermentation over respiration (Vander Heiden et al. 2009), is less intensified in cell lines.

In this study we deduced the changes in tRNA expression based on microarray measurements. Yet, the life cycle of a tRNA molecule is complicated as it consists of transcription, sometimes splicing, further processing including base modification and charging with amino acid. The charging levels of isoaccepting tRNA species are also sensitive to environmental changes (Sorensen 2001; Elf et al. 2003). Presumably, the most ideal proxy of tRNAs availability is the actual concentration of amino acid-loaded tRNAs. Recent measurements (Zaborske et al. 2009) are beginning to supply estimates on availability of 'ready-to-translate' tRNAs; future studies in this direction would probably refine our estimation of the dynamics of the tRNA pool.

6.3.2 An underlying modular design of the genome codon usage distinguishes between distinct biological processes

Non-random utilization of synonymous codons, typically termed "codon bias", has shown to reflect translational selection in many species, including bacteria species (Lithwick and Margalit 2003), yeast species (Man and Pilpel 2007), *C. elegans*, *D.melanogaster* and *Arabidopsis thaliana* (Duret and Mouchiroud 1999; Duret 2000; Heger and Ponting 2007; Drummond and Wilke 2008). Specifically, highly expressed genes are characterized by over-representation of high-efficiency codons, i.e., codons which are translated in the cell by the most abundant tRNA types. The question of whether there is or there is no translational selection in human is still open.
Some studies found no evidence for translational selection in human (Kanaya et al. 2001; dos Reis et al. 2004), suggesting that synonymous codons in human are not selected to maximize translation efficiency (Lercher et al. 2003). Conversely, other studies do indicate weak, yet significant, translational selection in human, according to estimates of codon usage adaptation to the global tRNA pool (Comeron 2004; Lavner and Kotlar 2005).

We reveal that the translated portion of the human genome can be characterized by two prototypes of codon usage programs, which distinguish between gene associated with proliferation and gene sets associated with differentiation or cell arrest (figure 5). We further demonstrated that this dichotomy of the codon usage is associated with the actual physiological state of the cell – for instance, cancer up-regulated genes are localized in the vicinity of the codon usage space where the

proliferation and the Central Dogma processes genes are, whereas cancer down-regulated genes span the codon usage space region which is associated with differentiation and multi-cellularity. Interestingly, the relative deviation of induced genes from each of the proliferation and differentiation codon usage signatures may reflect the proliferative state of a given cell at a given time point - for instance, human embryonic stem cells which were triggered by retinoic acid induced more genes with differentiation-like codon usage after 5 days of induction, compared to the earlier time points of 1-and-3 days (figure 9). In this context, it would be of high interest to examine the codon usage signature of genes which are induced in primary tumors with different clinical progressivity. Particularly, metastases were not examined here and will serve as an interesting subject for further investigation.

In fact, each of the two main signatures of codon usage may consist of hidden layers of sub-clusters. For instance, we found that the codon usage of genes associated with negative regulation of angiogenesis deviates from the global codon usage of the genes involved in angiogenesis: while most angiogenesis genes have a codon usage that is typical of multi-cellularity genes, the genes that negatively regulate the process have a more "proliferation-like" codon usage signature. The modular design of the codon usage is not only limited to human - we observed similar proliferation-differentiation dichotomy in mouse and fly. Yet the worm – a post-mitotic animal, in which differentiation and proliferation are not un-coupled, does not show such separation. Furthermore, the dichotomy in codon usage is not necessarily related to proliferation versus differentiation modes, but may in general separate between genes associated with two mutually exclusive processes – for instance, we observed differential codon usage between the genes involved in yeast glycolysis and the genes involved in yeast respiration. Future analyses of the differences in codon usage while gauging the actual representation of codons in the transcriptome upon different conditions may further sharpen our understanding of the modularity in the genome codon usage.

A striking result is that the first principal component, along which the gene categories of differentiation and proliferation are clearly distinct, spans a significant portion (40%) of the variance that exists in the entire set of 61 PCs. Interestingly, we found no correspondence between the absolute translation efficiency of the genes and the distribution of the various GO categories along the first component. This result demonstrates one of the main concepts introduced in this thesis – translational

selection in human is not reflected in the simplistic terms of adaptation between the static codon usage to a constant tRNA pool, but rather in terms of varying adaptation between the actual codons representation in the transcriptome and a dynamic tRNA pool.

Our research show that the coding sequences of genes which are induced in cancer is biased toward the proliferation codon usage signature, whereas the codon usage of genes which are repressed in cancer fit the differentiation codon usage signature (figure 9). An intriguing question is then, which if any of these codon usage signatures is obeyed by either known oncogenes or tumor suppressors. Intriguingly, we found that the codon usage of few oncogenes, such as AKT3, KRAS, NRAS, BRAF and MDM2 significantly fits the proliferation signature and significantly differs from the differentiation signature. In this context it would be of interest to look at the so-far ignored synonymous mutations of known oncogenes or tumor suppressors, and to examine if such mutations alter the codon usages towards either the proliferation or the differentiation signature.

6.3.3 Cancerous processes may determine the cell fate by hijacking the translation machinery

Translation efficiency of genes is determined by the adaptation between the codon usage of the genes to the cellular tRNA pool. As the tRNA pool is considered to be constant, such adaptation– i.e., over-representation of high-efficiency codos - is assumed to be achieved along evolutionary time scale and to reflect translational selection. Here we challenge the most basic paradigm of the interplay between the tRNA pool and the codon usage – we show that the tRNA pool does vary upon both normal short range physiological and cancerous processes, hence the extent of adaptation between the codon usage of a given gene to the cellular tRNA pool changes accordingly along physiological time scale as short as few days.

We further found that the changes in the tRNA pool upon cancerous processes affect the translation efficiency of distinct gene set in a heterogeneous manner. Specifically, we show that the changes in the expression of the various tRNA types boost the predicted translation efficiency of proliferation-related genes whose mRNA levels are elevated upon cancer. This result has important implications to both fields of translation and cancer research. On the one hand, the result illustrates that

translational selection is dynamic and reversible on a short time scale. On the other hand, our result suggests that the interplay between the tRNA pool and the codon usage may serve as a regulatory mechanism to promote switching in cellular state from differentiation to proliferation. We further suggest that the distinction in codon usage and tRNA pool between the two states evolved as it lower the probability that differentiated cells will proliferate without control. We suggest a "double lock" scenario in which even if certain oncogenic mRNA-level increases in a differentiated cell it is less likely to be efficiently translated since the corresponding tRNAs that are needed to translate it efficiently are expected to be present at low level in this tissue.

We calculated here the predicted changes in translation efficiency of genes based on changes in the expression of the various tRNA types. Yet, translation efficiency is determined by both the availability of the tRNAs, and the strength of codon-anticodon pairing. Forth of the standard amino acids – Phe, Cys, Asn, Asp and His are encoded each by two different codons, which are translated in the human genome by the very same tRNA type via either fully-match or wobble interaction. Specifically, the affinity of each of the five corresponding tRNA types to its translated codons is determined by the same base modification. Interestingly, for each of these five amino acids, there is a clear preferred codon when it is required in either proliferation or differentiation genes, implying for a potential role of tRNA modification in tuning the dynamic adaptation between the codon usage and the tRNA pool. Moreover, such combination of strategies – change in the expression of many individual tRNA genes and change in the expression of one or few enzymes influencing many tRNA genes - may allow flexibility in the response time of the cell to emerging needs for changes in the tRNA pool as well as in controlling the extent and duration of such changes.

In this study we observed coordinated changes in the tRNA pool and the actual representation of codons in the transcriptome upon cancerous processes - cancer predominantly increase expression of tRNAs whose codons are preferred among genes which are induced in cancer and represses the expression of tRNAs whose codons are enriched among genes which are down-regulated in cancer. Furthermore, in some of the examined cancerous cell types, we observed a significant correspondence between the tRNAs and the mRNAs expression even when directly compared between the FC in the tRNAs expression and the variation in the representation of the codons at the transcriptome, achieved by multiplication of the

genes static codon usage by their actual expression values. For instance, we detected Pearson correlation of 0.42 (P-value = 0.02) between the change in the tRNAs expression and the change in the demand for them in lymphoma cell line. An intriguing question is then, whether there is a causality relation between the variation in the tRNA pool and the changes in representation of the various codons, and if so – what is the direction of such causality – can the tRNA pool sense the varying transcriptome, or does the elevation in the expression of the "pro-proliferation tRNAs" trigger the expression of the proliferation genes, which are enriched by their corresponding codons. Alternatively, the synchronization between elevation of "pro-proliferation" tRNAs and proliferation genes may be achieved by means of co-regulation. Our results, demonstrating higher enrichment of the activating H3K27ac and H3K9ac modifications in both "pro-proliferation" tRNAs and proliferation mRNA-coding genes compared to "pro-differentiation" tRNAs and differentiation mRNA-coding genes respectively, hint for such coordinated regulation. Similarly, additional regulatory elements, such as transcription factors, which are typically associated with POL II genes, may be involved in the regulation of the POL III genes.

Our preliminary evidences for transcriptional regulation of tRNA genes upon cancer do not exclude the possibility that chromosomal aberrations also play a role in shaping the cancerous tRNA pool. In that context, it would be of interest to look for co-amplification of adjacent proliferation genes and "pro-proliferation" tRNAs, or co-deletion of adjacent differentiation genes and "pro-differentiation" tRNAs.

While typically cancer is assumed to represent disorder that involves the transformation of normal cells into rapidly dividing cells, our results imply for deliberate rather than random re-organization of the tRNA pool and the translation machinery in malignant cells. Particularly, the cancerous tRNA pool seems to be specifically adapted to the codon usage of the Central Dogma functionalities and much less adapted to differentiation and development related genes, suggesting that cancer hijacks of the translation machinery to promote conversion from coordinated multi-cellular behavior of cells towards a 'selfish' unicellular-like life style.

7. A putative recycled pool of tRNA may boost translation efficiency

7.1 Introduction

The non-random usage of synonymous codons, typically termed "codon bias", reflects the action of two main evolutionary forces: selection for translational efficiency and mutational drift acting on coding and noncoding DNA (Akashi 1994; Berg and Silva 1997). The context of a consecutive pair of codons (representing the pair of codons located in the A and P ribosome sites when being translated) is also shown to be biased, as some codon pairs are used in coding sequences much more frequently than expected from the usage of the individual codons of these pairs, while some other codon pairs are observed much less frequently than expected. Codon pair biases are found to be directional, i.e. the bias associated with codon pair A-B is larger than the bias associated with codon pair B-A (Hatfield and Gutman, 1992). Codon pair bias was also found to be species-specific (Gutman and Hatfield 1989; Moura et al. 2007), and a weak correlation was reported between codon pair bias and codon bias (Gutman and Hatfield 1989; Buchan et al. 2006). Codon-pair bias is mainly assumed to be related to translation accuracy, as it has a direct impact on missense, nonsense and frameshifting errors (Precup and Parker 1987; Parker 1989). It was suggested that in eubacteria and archeae, codon-pair context is mainly determined by constraints imposed by the translational machinery, while in eukaryotes the emergence of DNA methylation and tri-nucleotide repeats influenced codon-pair context (Moura et al. 2007).

While translation fidelity is associated with non-random arrangement of pairs of adjacent codons, traditional measures of translation efficiency only consider the global codon usage of a gene, ignoring the order of the codons along it. Yet, a recent study suggests that the order of the high- and low-efficiency codons along the genes is of prime importance. Analysis of multiple genomes revealed a trend in which the first approximately 30–50 codons in genes preferentially correspond to more rare tRNAs (Tuller et al. 2010). Such genic sections form ‘low efficiency ramps’, which might deliberately attenuate the ribosome during early elongation; It was proposed that such attenuation enables a jam-free flow of ribosomes once they passed that region, thus reducing the probability of ribosome fall-off.

In my thesis I challenge an implicit assumption of traditional models of translation efficiency that all codons utilize the same global tRNA pool. Instead, I

hypothesize that codons at the ribosome A-site can utilize recycled tRNAs from the codons that were just translated, and are hence exposed to local pools of tRNAs that may either promote or disrupt translation elongation efficiency and accuracy. Such an hypothesis hence predicts a bias in both the composition and order of codons in synonymous codons pairs which are separated from each other by one or more codons. Consistent with my hypothesis, a recent observation (Cannarozzi et al. 2010) showed that in subsequent occurrences of the same amino acids, genes tend to deliberately use codons that are translated by the same cognate tRNA. Similar to the ramp design, this trend was shown to be predominantly obeyed by rapidly induced genes, hinting that this is another means to boost translation efficiency.

7.2 Results

Along my research, I examined potential temporal and condition-dependent deviations from the simple view of a constant demand of the tRNA pool. In parallel I challenge the prevailing notion of one global tRNA pool even at a given condition and time. Instead I raise the hypothesis that codons at the ribosome A-site are exposed to local pools of recycled tRNA, which may be contributed from either translation on the very same ribosome, or from translation on neighboring ribosomes. I term the hypothetical local concentration of a recycled tRNA molecules in the vicinity of a given codon as "local tRNA pool", and suggest that it (a) may elevate translation efficiency where subsequent occurrences of the same amino acids are encoded by repetitive codons, and (b) may promote both translation efficiency and accuracy if near-cognate tRNAs of a given codon are depleted from its local tRNA pool. A real validation of the local tRNA pool hypothesis is currently impossible, as it requires measuring the actual location of individual tRNAs along the transcript. Yet, I suggest that sequence organization rules that are consistent with the hypothesis may indicate the existence of local tRNA pools.

7.2.1 Repetitive codons pairs are favored in subsequent occurrences of the same amino acids in S. cerevisiae

I measured the observed frequency of repetitive pairs out of the total codon pairs of identical amino acids in *S. cerevisiae*. The analysis was done separately for highly and lowly expressed genes, which are defined here as the 1000 genes which hold the highest average mRNA expression levels in the yeast cell-cycle expression data (Cho et al. 1998) and the 1000 genes which hold the lowest average levels in that data, respectively. Specifically, the highly expressed genes are enriched with constantly-expressed genes. For instance, 206 of the 1000 higly-expressed genes belong to the "Translation" GO category (P-value = 4.29e-26). The observed frequency was calculated for codon pairs with distance of up to 100 codons (the distance between 2 consecutive codons is defined to be 1). The observed frequencies are compared to the expected frequency, which is estimated by:

Expected frequency of repetitive pairs =

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\left(fCj\right)^{2}\right)\times fAAi$$

Where *i* is an amino acid (AA), *j* is a repetitive combinations of AA *i*, *f*C*j* is the frequency of the corresponding codon out of AA *i* total number, and *f*AA*i* is the frequency of AA *i* out of all the total analyzed amino acids (1-box tRNAs and stop codons were excluded). Figure 18 shows that repetitive codon pairs are preferred in subsequent occurrences of the same amino acids. The results show observable preference of repetitive codon pairs for repetitive amino acids with small distance (2-3), as well as a very strong periodicity of 12 codons, with preferred distance between repetitive pairs being 12, 24, 36, 48, and 60. In contrast lowly expressed genes show little deviation from the expected frequency of repetitive codon pairs.

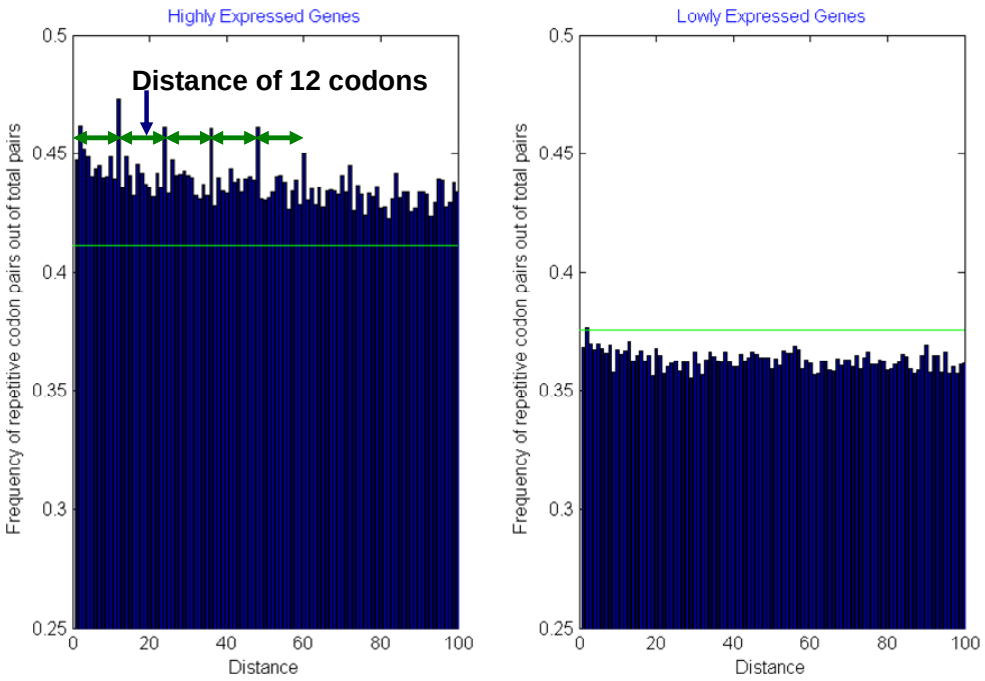


Figure 18: **Preference of repetitive codon pairs in highly expressed genes of yeast.** The blue bars represent the observed frequency of repetitive codon pairs out of the total pairs of identical amino acids. The Green line in each plot is the expected frequency of each pair (depends on both codon and amino-acid usage).

Since it is widely accepted that selection on synonymous sites acts to promote translation efficiency (Drummond and Wilke 2008; Tuller et al. 2010), it is crucial to validate that the signal is not a by-product of biased representation of synonymous codons in highly expressed genes. To this end, I repeated a codon-shuffle 100 times, in which I shuffled synonymous codons in each of the highly expressed genes, while keeping the original amino acid sequence (see figure 19a). The control results suggest

that the apparent preference of repetitive codon pairs is not derived from the usage of preferred individual codons in highly expressed genes.

A recent Cell paper, that was published during my work, similarly indicates, by both computational and experimental analysis, that in subsequent occurrences of the same amino acids, genes tend to deliberately use codons which are translated by the same tRNA isoacceptor (Cannarozzi et al. 2010). Yet, this publication discusses codon pairs that are translated by the same tRNA types, and does not distinguish between heterogenic pairs and repetitive pairs of synonymous codons. In our research, we focus on preference of repetitive codon pairs, and reveal species-specific and codon-specific signatures of repetitive codon pairs, each characterized by well defined periodicity.

7.2.2 Species-specific signature of repetitive codons pairs

We expand our analysis to additional organisms, and detected no significant preference of repetitive codon pairs for either *E. coli* or *S. pombe*. However, we did detected preferences of repetitive codon pairs in highly expressed genes of *C. elegans*. Interestingly, similarly to the case of *S. cerevisiae*, we observed a periodicity in the elevated usage of repetitive codon pairs, though with smaller interval. Preference of identical codon pairs in highly expressed genes of *C. elegans* is associated with distances 3, 6, 9, 12, 15 etc. Implementing the codon-shuffle control, we noticed that the stronger signal is for distance 3 (figure 19b). This result may imply that in *C. elegans* the repetitive pattern in codon usage is affected by preference of individual codons.

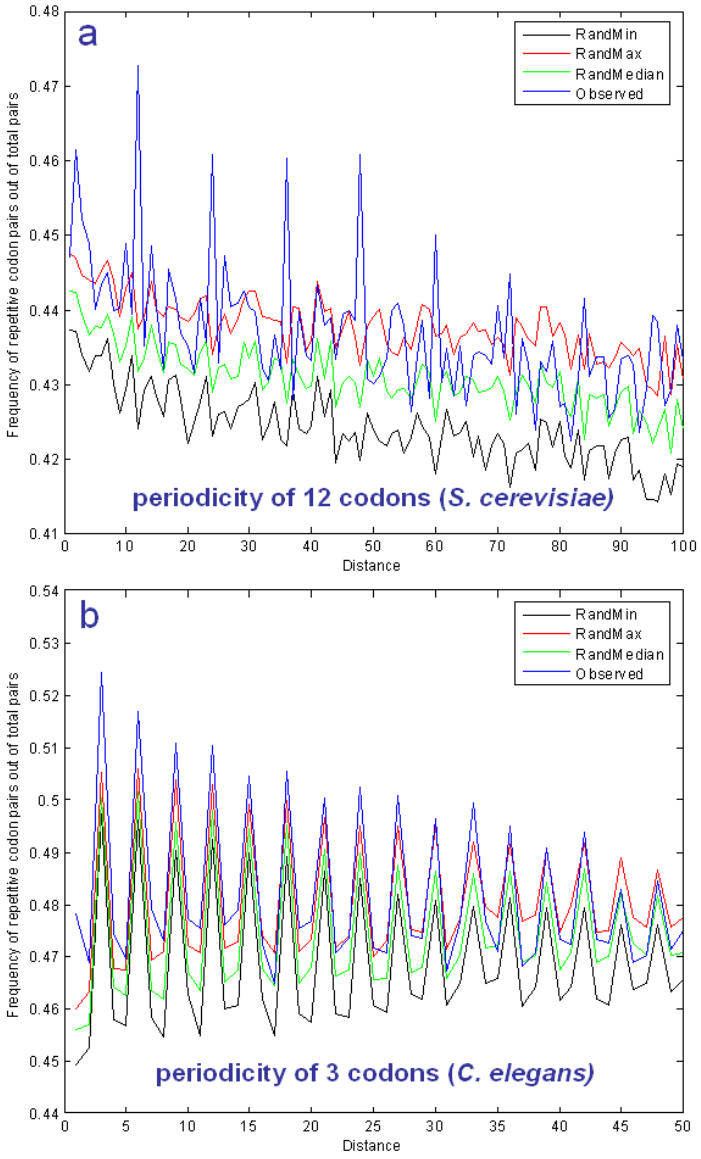


Figure 19: The pattern of preference of repetitive codon pairs varies between different organisms. The blue line is the observed frequencies of repetitive codon pairs out of the total identical amino acids codon pairs. The black and red lines represent the minimum and maximum computed frequencies respectively, whereas the green line is the median value. (a) Periodicity of 12 codons in yeast. (b) Periodicity of 3 codons in nematode.

7.2.3 Preference of repetitive codons pairs is associated with specific amino acids

We next checked whether the observable preference of repetitive codon pairs is equally obeyed by all the 18 amino acids which are translated by two or more synonymous codons. Interestingly, we found that the signal is uniquely associated with specific amino acids. The observed periodicity of 12 codons in highly expressed genes of *S. cerevisiae* is pronounced in Thr (ACC and ACT codons) - demonstrated in figure 20, Ser (TCC and AGT codons), and Asn (AAC codon). The case of AGT (Ser) is of prime interest. This codon is a low-efficiency one, as it is translated via wobbling interaction by tRNA isoacceptor which is transcribed by only two tRNA genes. Repetitive occurrences of such a low-efficiency codon suggest that this sequence organization rule is intriguing on its own, and is not a side effect of selection towards global translation efficiency. The observed periodicity of 3 codons in highly expressed genes of *C. elegans* is derived from repetitive codon pairs of Pro (CCA) and Gly (GGA).

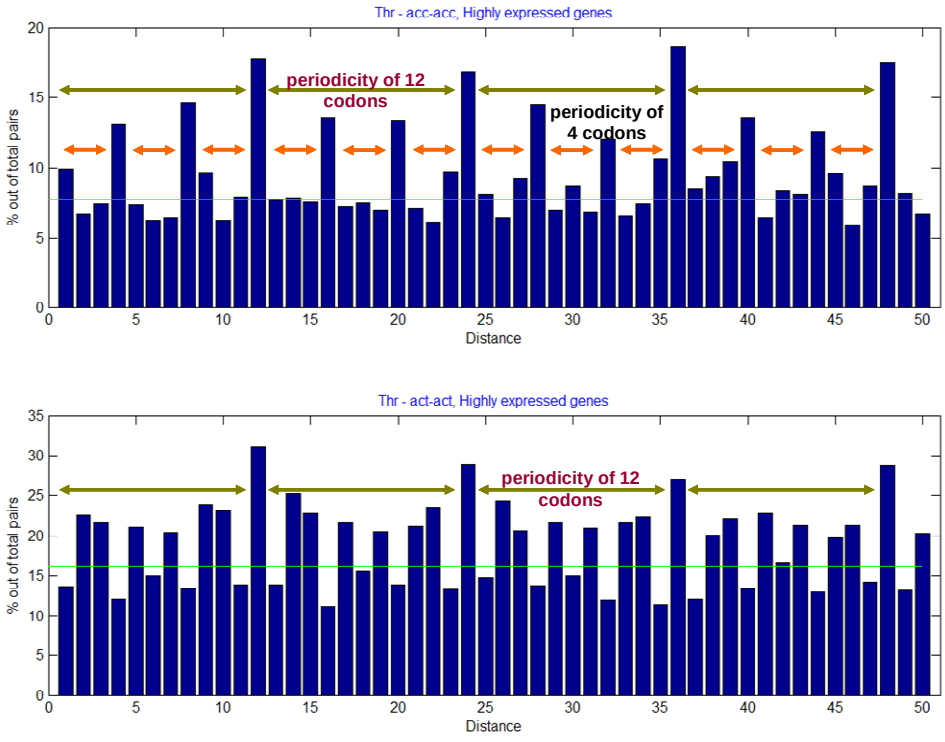


Figure 20: **A periodicity in the elevated usage of repetitive codon pairs in subsequent occurrences of Throenine amino acid.** The blue bars represent the observed frequency of ACC-ACC and ACT-ACT repetitive codon pairs out of the total pairs of codons encoding Thr amino acid. The Green line in each plot is the expected frequency of each pair (depends on the frequencies of the four Thr codons)

7.3 Discussion

In this study, I hypothesize that translated codons at the ribosome A-site can utilize tRNAs from two sources – the global cellular tRNA pool, which is equally accessible to all the translated codons, and a local tRNA pool, which consist of recycled tRNAs from the codons that were just translated. For a given codon, the best hypothetical tRNA local pool is a homogenic pool composed of its fully-matched tRNAs. On the other hand, the worst pool comprises its near-cognate tRNAs, as the initial binding of near-cognate tRNA causes delays of varying duration to the observed rate of translation (Rodnina and Wintermeyer 2001) and may result in misincorporation of the wrong amino acid into the protein. However, the combinatorial space is wide and the nature of the local tRNA pool depends on the identity of the codns that were recently translated in the vicinity of the codon. Under this assumption we suggest that translational selection may not be exclusively reflected by the global codon usage which correspond to the global tRNA pool, but may also be recognized while examining the order of the codons along the transcript. Such order can determine the nature of the local tRNA pool available for each of the gene's codons.

Consistent with our hypothesis, we found that repetitive codons pairs are favored in subsequent occurrences of the same amino acids in highly expressed genes of *S. cerevisiae* and *C. elegans.* We speculated that such sequence organization pattern promotes local pools of required tRNAs, which may boost the translation elongation efficiency of the genes. During our research, a similar study (Cannarozzi et al. 2010) appeared which indicated that in subsequent occurrences of the same amino acids highly expressed genes tend to use codons that are translated by the same tRNA type. In general, the putative utilization of recycled tRNAs requires re-charging of the tRNAs with the corresponding amino-acids. However, since the affinity of the very same tRNA type to distinct synonymous codons is not equal and is dependent on specific base modifications, the adjustment of recycled tRNA from the translation of one codon type to another may require further involvement of tRNA modification proteins. In that aspect, our own observation which specifically shows over-representation of identical codons in subsequent occurrences of the same amino, rather than preference of codons which are translated by the same tRNA but may be

different from each other, is more adequate to the hypothesis of local tRNA pool – recycle-tRNAs which are used for the translation of identical codons need no further wobble-related base modification, thus are are more compatible with their 'ready-to-translate' form. Intriguingly, we observed periodicity in the pattern of co-occurrences of repetitive codons. This periodicity seems to be both amino acid- and species-specific. It would be of great interest to find out whether such periodicity is associated with secondary and/or tertiary structural elements of the protein.

Similarly to Cannarozzi's paper, we also detected preffered combinations of codons which are translated by the same tRNA in subsequent occurrences of the same amino acid. Yet, we realized that not only the composition of the codon pairs is of interest, but also their order. The reasoning behind this notion is that the relations between the codons are not necessarily symmetric. For instance, the genetic code contains 13 pairs of synonymous codons in which the first two nucleotides (marked with XY henceforth) are identical, and the third nucleotide is either A and G. According to the wobble rules (Crick 1966), the fully-matched tRNA of the XYA codon can translate the XYG codon, but the fully-matched tRNA of the XYG codon cannot translate the XYA codon. We examined such codon pairs in *S. cerevisiae* and indeed observed directional patterns in utilization of heterogenic codons in subsequent occurrences of the same amino. Specifically, in highly expressed genes we noticed modest and yet significant preference of XYA-XYG codon pairs upon the opposite combination of XYG-XYA pairs (data not shown).

We assume that the nature of the putative local tRNA pool may not be only associated with translation efficiency, but also with translation accuracy. Particularly, for a given codon, we looked for depletion of near-cognate tRNAs from its local tRNAs pool. We scanned codon pairs of heterogenic amino-acids, but found no evidences for a global avoidance of subsequent occurrences of codon pairs in which the cognate tRNA of the first codon is considered as near-cognate tRNA of the second one. Yet, deliberate depletion of near-cognate tRNAs from local tRNA pools of a codon may still exists in specific regions where mis-incorporations are most likely to disrupt protein functions.

From a kinetic point of view, my hypothesis is not trivial. First, it requires that the diffusion of the recycled tRNA will be slow enough compared to the rate of translation elongation. Alternatively, if simple diffusion might be too fast compared to rate of elongation then to coordinate tRNA availability with translation one needs to

invoke ‘local translation factories’ nearby the ribosome, which will supply the re-charging services to the recycled tRNA and limit their diffusion away from the ribosome. Studies indicating the capacity of aminoacyl–tRNA synthetases to interact with the ribosome (Kaminska et al. 2009) and reporting on colocalization of protein translation components (Barbarese et al. 1995) may serve as supporting evidence.

8. Codon choice may reflect a potential balance between "Efficiency" and "Accuracy"

8.1 Introduction

In the stochastic search for the right tRNA, the ribosome might incorrectly bind a tRNA with a one base mismatch relative to the codon, often termed 'near-cognate tRNA' (tRNAs with more than one base-mismatch relative to the codon do not bind, or are highly unlikely to do so) (Rodnina and Wintermeyer 2001). If a near-cognate tRNA binds to the A-site of the ribosome, the wrong amino acid might be incorporated, creating a 'missense translational error'. The frequency of such translation errors *in vivo* was estimated to be 10^{-5} in yeast cells (Stansfield et al. 1998), but more recent measurements in *B. subtilis* showed a surprisingly high rate of 10^{-2} (Meyerovich et al. 2010). Many studies measured the frequency of specific amino acid substitutions, using various methods and reporter systems (Edelmann and Gallant 1977; Khazaie et al. 1984; Parker and Holtz 1984; Toth et al. 1988; Cornut and Willson 1991). A more recent paper (Kramer and Farabaugh 2007) quantified the misreading errors caused by binding of Lys(UUU)-tRNA to mutated active site of firefly luciferase, in which the essential lysine codon was replaced by various codons which are considered as near-cognate substrates of the Lys(UUU)-tRNA. Interestingly, the researchers revealed a varying range of misreading frequencies, which have shown to be associated with a competition between the corresponding cognate tRNAs and the Lys(UUU) near-cognate tRNA. Missense errors that might disrupt protein function impose metabolic costs of wasted synthesis; if the loss of function is accompanied with improper folding, the damage might be even more pronounced. The misfolded protein may interact with other cellular components, causing protein aggregation (Bucciantini et al. 2002), disruption of membrane integrity (Stefani and Dobson 2003) and it may ultimately result in cell dysfunction and disease—reviewed in (Gregersen 2006).

Many species, including *E. coli*, yeast, worm, fly, mouse and human, show translational selection in favor of optimal codons - in terms of translation efficiency - at sites where misincorporations are most likely to disrupt protein functions (Akashi 1994; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008). Selection for translation accuracy was shown to be predominantly associated with prevention of misfolding and its potential toxic consequences (Drummond and Wilke 2008; Zhou et

al. 2009; Warnecke and Hurst 2010; Yang et al. 2010b). Further, in the context of translation accuracy, selection pressures on synonymous sites also appear to act against frame-shifting errors (Farabaugh and Bjork 1999), and to reduce the cost of nonsense errors (Gilchrist et al. 2009).

Translation can thus be thought of in terms of a competition process between the cognate and near-cognate tRNAs for a given codon, where the higher the concentration of correct tRNAs, the lower the probability of binding the wrong ones. While most of the studies in this field focus on the availability of the cognate tRNAs, few mathematical models suggest that the competition from near-cognates, and not the availability of cognate aa-tRNAs, is the most important factor which determines the translation rate (Zouridis and Hatzimanikatis 2008). In particular it has also been shown that the codons with highest near-cognate competition in *E. Coli* overlap only partially with the rarest codons (Fluitt et al. 2007). In my study I investigated the so far unrevealed balance between selection on translation efficiency and translation accuracy, as is reflected in choice between synonymous codons, towards either the most-efficient codon or the one which is exposed to the smallest pool of near-cognate tRNAs.

8.2 Results

A basic paradigm of translation efficiency is that highly expressed genes and highly conserved regions of genes are enriched with high-efficiency codons, i.e., codons which are translated by large pools of cognate tRNA molecules. In addition, translation efficiency measures which explicitly consider the tRNA pool (Ikemura 1981; Ikemura and Ozeki 1983; dos Reis et al. 2004), are gauging not only the tRNAs availability, but also the different types of codon-anticodon pairing. Typically, translation by a fully-matched tRNA is assumed to occur faster than translation via wobble interaction.

However, during the translation process, the ribosome might incorrectly bind a tRNA with one base-mismatch relative to the codon, often termed "near-cognate tRNA". The definition of near-cognate tRNA is wide, and may also refer to near-cognate recognition between codons and the fully-matched tRNAs of their synonymous ones (specifically in cases where mismatch between the 3rd position of the codon and the 1st position of the anticodon cannot form wobble interaction). However, in this section we restricted the definition of near-cognate relation to reflect potential pairing between codon of a given amino-acid and tRNAs correspond to different amino-acid. Having two types of cognate and near-cognate tRNAs, both recognize the very same codon, but carry the right and wrong amino acid, respectively, the elongation rate might not only depend on the availability of cognate tRNAs, but also on the near-cognate tRNAs concentration. Hence, we propose that translation efficiency should also reflect the competition between these tRNA types.

While typically, the choice between synonymous codons is mainly associated with the abundance of cognate tRNAs, we speculated that the ratio between cognate and near-cognate tRNAs also play a role in shaping codon preferences. In order to examine the potential effect of the cognate—to—near-cognate tRNAs ratio, we focused on codon pairs of the type XYT-XYC. Our definition of XYT-XYC codons refers to pairs of codons which consist of identical nucleotides in their first and second positions (X and Y could be any nucleotide, but same in the two codon), but differ from each other in their third nucleotide, which is "T" in one codon and "C" in the other. There are 16 such codon pairs, and according to the Genetics Code they are all synonymous. In accordance with the wobble rules (Crick 1966) applied to such XYT-XYC pairs, the fully-matched tRNA of each codon may also translate the other

one, via wobble interaction. In most cases, only one of the two fully-matched tRNA types of the XYT-XYC codons exists in the genome, codons of the other are translated by tRNA that is present. Whereas both the codons of a given XYT-XYC pair are typically served by the same cognate tRNA type, the abundance of their near-cognate tRNAs is often varying. Defining "x" and "y" as any nucleotide which is not "X" and "Y", respectively – the XYT codon may incorrectly bind the fully match tRNAs of either the xYT or the XyT codons, where the XYC codon may incorrectly bind the fully match tRNAs of either the xYC or the XyC codons. This scenario allows us to directly examine the potential impact of the cognate—to—near-cognate tRNAs ratio on choice between synonymous codons.

While the global representation of the XYT and XYC codons in the ORFome may be related to GC content, it is well known that highly expressed genes and highly conserved regions of genes may deviate from the global pattern of codon preferences. Although the XYT and XYC codons are translated by the very same tRNA type, they do differ from each other in their pairing type with that tRNA, i.e. – fully match interaction or wobble interaction. Hence, a naïve assumption would suggest, that if the cognate tRNAs are indeed the major determinant of codon choice, then, for the XYT and XYC codon pairs, the most expressed and the most conserved genes will show over-representation of the codon which its fully matched tRNA exists in the genome, as this pairing type is assumed to be favored upon the wobble interaction in terms of pairing strength.

To examine this notion, for each synonymous XYT-XYC codon pair, we plot the ratio between the usage of the codon which is translated by its fully-matched tRNAs to the usage of the second codon, which is also translated by the same tRNA type, but via wobble interaction (Figure 21). We examined how this relative usage of the two codons changes as a function of the expression or conservation of genes. This analysis was done for *S. cerevisae*, whose genome contains only one of the two fully-matched tRNA types of each of the XYT-XYC codon pairs (either for the XYT or XYC codon). As can be seen, there is a striking change in codon preferences while comparing highly-expressed genes to the rest of the genes. However, only in nine out of the 16 examined pairs (Aspartic acid, Asparagine, Histidine, Tyrosine, Proline, Alanine, Serine, Arginine and Phenylalanine amino acids), codon preference is in the favor of the codon which is translated by the fully-matched tRNA. In five of these cases, there is a shift in the codon preference – at low or intermediate levels of

expression codon preference is in the favor of the codon which is translated via wobble interaction, while highly expressed genes show preferred usage of the fully-matched codon. In parallel, we checked the pattern of codon preference as function of the genes conservation level. In most cases, qualitatively, the pattern of codon preferences is similar to that observed along the span of gene expression, and yet, some considerable differences can be seen. For instance, for each of the amino acids Asparagine, Aspartic acid, Histidine and Tyrosine, the elevation in the usage of the codon that is translated via fully-match interaction is much higher for the most highly expressed genes compared to that of the most highly-conserved genes. Since the group of highly-expressed genes overlaps that of highly-conserved genes, the actual gap in change in preferences is even higher.

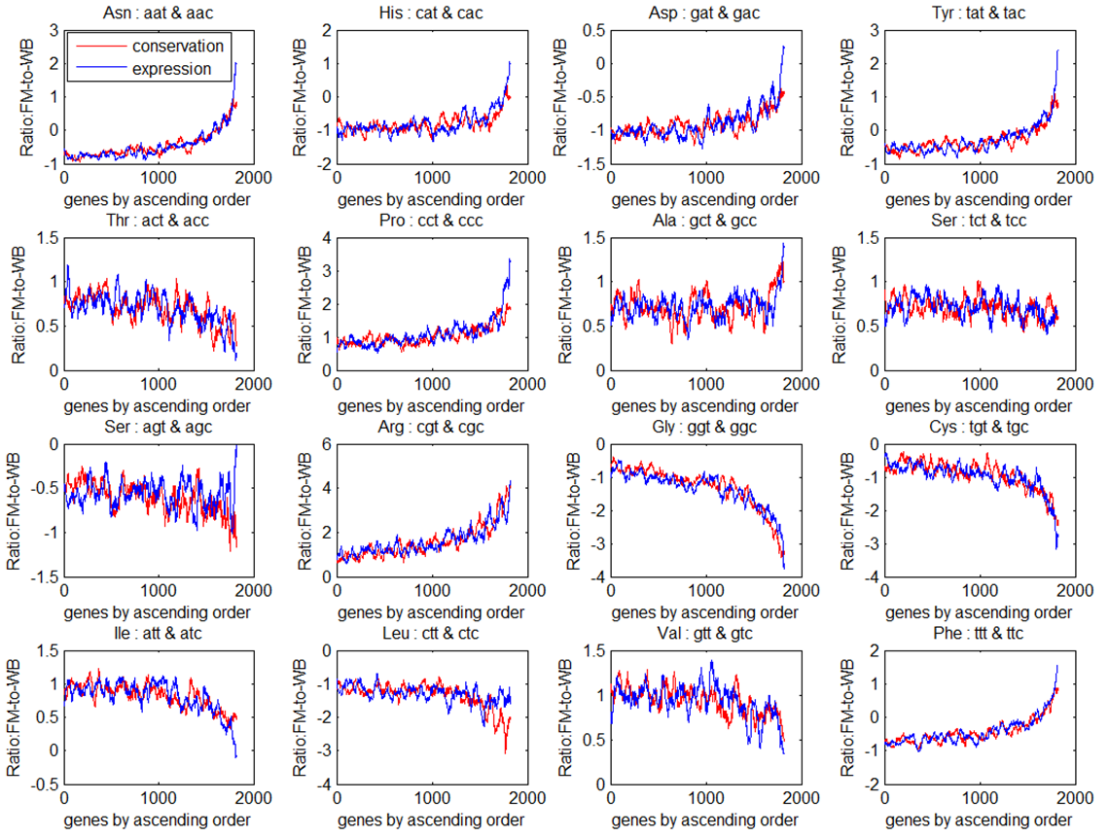


Figure 21: Codon preferences vary as function of gene expression and gene conservation. Each plot display the log 2 ratio of usages for synonymous XYT-XYC codon pairs, in which both codons are translated by the same tRNA type, via either the fully-matched (FM) tRNA or, via wobble (WB) interaction, by the non-fully matched tRNA. Any dot along the x-axis represents a window of 50 genes, where the genes are arranged in ascending order according to their expression level (blue line) or conservation level (red line; conservation was assessed by aligning *S. cerevisiae* genes to their orthologs in additional yeast species). The y-axis denotes the usages ratio in log2 scale. Along the y-axis, values higher than 0 denote for preference of the codon which is translated via fully-matched (FM) interaction, while negative values indicate a preference of the codon which is translated via wobble interaction.

We hypothesize that such differential codon preferences of highly-expressed genes and highly-conserved genes may reflect trade-offs between speed and fidelity

considerations. The four above mentioned amino acids, Asparagine, Aspartic acid, Histidine and Tyrosine, are encoded by pairs of XYT-XYC codons, in which the nucleotide in the second position is "A" (figure 22). Thus, any cognate tRNA which translates these amino acids, has one mismatch with one of the codons of the three other amino acids, and may incorrectly bind to it. For instance, the fully matched tRNA of the AAC (Asn) codon, is a near-cognate tRNA of each of the CAC (His), GAC (Asp) and TAC (Tyr) codons. We term such quartet of codons as "Competition box".

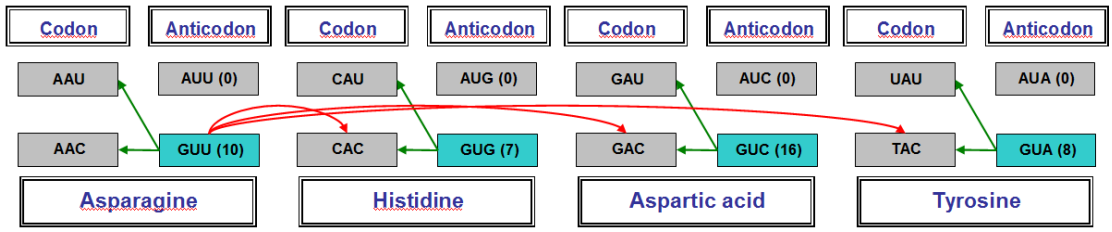


Figure 22: An example of a "competition box" - XYC-XYT pairs of four amino acids in *S. cerevisiae* and their corresponding tRNAs. The numbers in parentheses denote the gene copy numbers of the tRNAs; the arrows represent relation between tRNAs and codons –green arrows represent cognate pairing, by either full-match or via wobble interaction, while red arrows indicate potential pairing between codons and their near-cognate tRNAs (tRNAs with a one base mismatch relative to the codon, which might incorrectly bind it, leading to misincorparation of wrong amino acid).

Since only the fully-matched tRNAs of the XYC codons of these four amino acids exist in the yeast genome, these codons may be favorable over their synonymous XYT codons in term of translation efficiency (defined here as function of codon-anticodon pairing strength). Yet, the XYC codons may be relatively inferior in terms of translation fidelity, i.e., the exposure to higher probabilities of missense errors (due to potential pairing with near-cognate tRNAs). For instance, Aspartic acid is encoded by two codons – GAC and GAU. These two codons are translated by 16 genes of the GUC anticodon, by either fully-match interaction or wobble interaction (for the GAC and GAU codons, respectively). Yet, these codons differ from each other by the number of their near-cognate tRNA genes in the genome which are associated with the Asn-His-Asp-Tyr "competition box" – there are such 25 near-cognate tRNA genes for the GAC codon (10(GUU) + 7(GUG) +8 (GUA)), but none for the GAU codons (see figure 22). For the clarity of reading, in each such codon pair, we termed the codon which is translated via fully-match interaction as "high-

speed" codon, whereas the codon which is exposed to lower number of near-cognate tRNAs is defined as "high-fidelity" codon.

Focusing on the above "Competition box", we checked the codon preferences in 111 occurrences of Serine/Threonine protein kinases active site (obeying the signature PS00108 in the PROSITE databse) in yeast genes. This signature contains at least three representatives of the Asparagine, Histidine, Aspartic acid and Tyrosine amino acids, where Aspartic acid is the active residue. For three out of these four amino acids, we found the usage of the XYT (high-fidelity) codon dominates over that of the XYC (high-speed) codon, where the strongest preference was observed for Aspartic acid (Asp - 84 vs. 27 appearances; His - 67 vs. 31 appearances; Asn - 67 vs. 44 appearances; Tyr - 7 vs. 6 appearances). The codon preference of Asparagine, Histidine, and Aspartic acid are significantly different from the expected representation of the codons based on their frequency outside the Serine/Threonine kinase active site in these genes (p-value < 0.05, χ^2 test). This choice of a codon thus demonstrates that selection was made to increase fidelity at the expense of speed.

We also found that the preferred usage of codons exposed to low competition from near-cognate tRNAs is not only common in highly conserved genes, but it is also predominantly obeyed by the most conserved regions within these genes. Threonine, Serine, Alanine and Proline amino acid are each encoded by four codons, two of them in each amino acid are XYT-XYC codons. The second nucleotide of each of these amino acid's codons is "C". Thus, the fully-matched tRNA of a given codon of these amino acids, is a near-cognate tRNA of the codons of the other three amino acids which contain the same nucleotide (i.e. “C” or “T”) in their third position. In *Y.lipolitca*, all the fully-matched tRNAs of the XYT codons of these four amino acids are represented in the genome with at least 20 copies of the same gene, while none of the fully-matched tRNAs of the corresponding XYC codons exists in that genome. Thus, the XYT codon in the XYC-XYT pairs of each of these four amino acids may be associated with high-speed translation, whereas the XYC codon may be exposed to a lower level of tRNAs competition. Figure 23 shows the codon preferences associated with these four synonymous codons pairs for different gene sets and different positions within genes. Over all, the XYC codon, which is associated with high fidelity, is preferred upon the XYT codon, which is assumed to be translated with higher speed (considering codon-anticodon pairing types). This trend become

stronger when focusing on a set of conserved genes, and interestingly, among different positions of these genes, there is a striking signal of preference of the XYC codon in the most conserved positions, but there is no such signal at all for non-conserved positions.

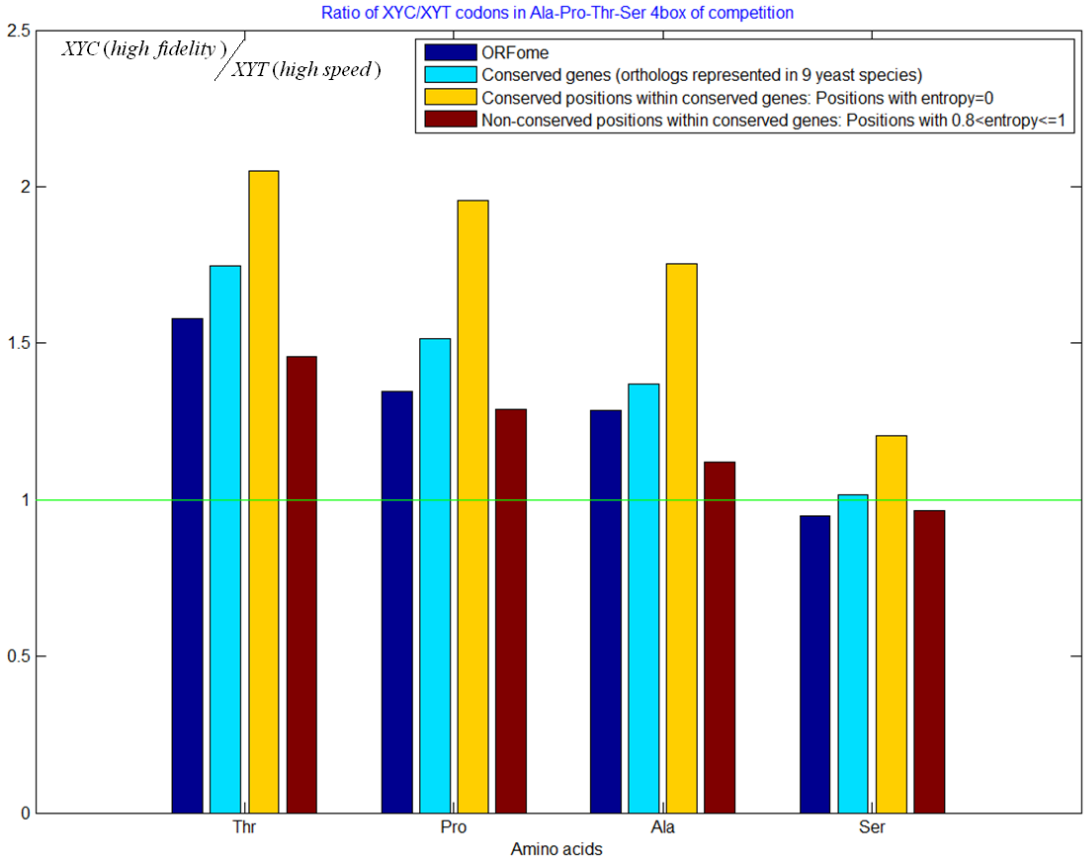


Figure 23: High fidelity codons are preferred in conserved genes, especially in their conserved amino acid positions. The y-axis denotes the ratio between the usage of the XYC and XYT synonymous codon pairs of four amino acids in *Y.lipolitca*. In each such codon pair, the XYT codon is translated via a perfectly matching codon-anticodon pairing type (thus refer to as "high speed" codon), whereas the XYC codon is expected to be translated with higher fidelity as it is subject to a lesser extent of competition from near-cognate tRNAs. The blue bars represent the full ORFome (~6000 genes); the light blue bars represent genes that are conserved among 9 yeast species (~1900 genes); the yellow bars represent the most conserved positions in the set of conserved genes, and the brown bars represent the non-conserved positions in these genes. Conservation level of a given position is shown in entropy terms, and range between 0 (in case that all species have the same amino acid in the very same position), to 1 (where for a given position, each species use different amino acid).

All together, our results imply for a trade-off between speed and fidelity of translation and in particular to a potential role of tRNA competition in shaping codon preferences of genes. These trends are predominantly enhanced at sites where mis-incorporations are most likely to disrupt protein functions.

8.3 Discussion

Classical studies typically define selection for both translation efficiency and accuracy by recognizing the over-representation of high-efficiency codons, i.e. – codons that are translated by abundant tRNAs. However, we realized that in the context of translation fidelity, codon-optimization should not only be thought of in terms of the availability of cognate tRNAs, but also in terms of the concentration of near-cognate tRNAs, i.e. – tRNAs with one base mismatch relative to the codon that may carry a different amino acid.

To investigate a potential role of codon choice in lowering the potential of amino acid mis-incorporation during translation, we inspected the choice between pairs of synonymous codons, in which both codons are solely translated by the same tRNA, but they differ from each other by their related probabilities to incorporate tRNAs with the wrong amino acids. Specifically, we focused on codon pairs in which the codon that is less likely to incorporate tRNAs with the wrong amino acid (hence favored in terms of translation fidelity) is translated via wobble interaction with the fully-matched tRNA of the second one, thus inferior compared to it in terms of the strength of the codon-anticodon pairing, and as a consequence - in terms of translation efficiency.

Reassuringly, we show that in highly conserved regions of yeast genes, the codons that are translated via wobble interaction but are exposed to low concentration of near-cognate tRNAs are strikingly preferred upon their counterpart codons.
We further noticed that, compared to the whole genome, the preference of high-fidelity codons over the high-efficiency codons in *Y. lipolitica* is especially pronounced in the most conserved positions of conserved genes, but is not seen in the non-conserved positions of such genes, implying for interesting trade-off between translation efficiency and the translation accuracy. Consistent with such balance between translation efficiency and translation accuracy, we observed that the extent, and in some cases even the nature, of choice between high-efficiency and high-fidelity codons varies between the most expressed genes and the most conserved genes in *S. cerevisiae*.

We observed preference of high-fidelity codons at the expense of high-efficiency codons through defining conservation of amino acid positions (through a

measure of entropy of amino acid composition in alignment columns), as well as by directly inspecting proteins active sites. Thus, the over-representation of the high-fidelity codons may reflect evolutionary pressure for translation accuracy in both structurally sensitive sites and sites where misincorporations are most likely to disrupt protein functions.

We further found that the preference of high-fidelity codons relative to high-efficiency codons in conserved regions of genes is obeyed by some but not all the amino acids, and is of species-specific nature. Yet, and intriguingly, we observed co-preference of high-fidelity codons among sets of four codon pairs of four different amino acids, in which each of the cognate tRNA of a given synonymous codon pair is a near cognate tRNA of one codon of each of the other three codon pairs. We observed such co-preference of high-fidelity codons in the 13 amino-acid-long active site of Serine/Threonine kinases in *S. cerevisiae*. Specifically, it involves Aspartic acid (the active residue), Asparagine and either Histidine or Tyrosine. Since the occurrences of each pair of these three amino acids in the motif signature is separated by small number of other amino acids (1-6 amino acids), our results may imply for deliberate depletion of near-cognate tRNAs from the putative local tRNA pool of the codons belonging to the active site (as suggested by the "local tRNA pool" hypothesis in this thesis and in (Cannarozzi et al. 2010)).

To date, there are few well-accepted measures of translation efficiency, but no common measure of translation accuracy. Consolidation of translation accuracy measure is complicated, as it requires assessing of both the availability of correct and incorrect tRNAs and the severity of incorporation of wrong tRNAs at each position, as well as information regarding potential position-dependent differences in the strength of the mismatch pairing. Whereas measures based on tRNA availability alone are poorly correlated with gene expression levels in higher organisms, an integrated model, which gauges both binding of cognate and near-cognate tRNAs, may deepen our understanding of translational selection as reflecting potential trade-off between speed and accuracy.

9. Summary of thesis

Selection for translation efficiency and accuracy is typically assessed by the adaptation between the codon usage of individual genes to the cellular tRNA pool. Traditional measures of translation efficiency rely on global and static attributes - the constant composition of codons in protein-coding genes and the constant number of tRNA genes in the genome. Hence, translation efficiency is considered as a fixed property of codons and genes along the physiological time scale.

In this study we show that translation efficiency of genes is a dynamic rather than a static trait. The reasoning behind this notion is our indications for both dynamics in the actual representation of codons in the translated transcriptome, as captured by observing expression of mRNAs upon different conditions, and dynamic availability of tRNAs, as captured by inspecting changes in tRNAs expression in cancer and in normal physiology. We further suggest that the dynamic nature of both the effective codon usage and the tRNA pool necessarily dictates dynamics in the adaptation between these two factors, leading to differences in the translation efficiency of the very same gene in different time points and upon environmental changes throughout organism life. Intriguingly, we even hypothesize dynamics in the translation efficiency of identical codons in different positions of the same gene, as may be governed by differential composition of cognate and near-cognate tRNAs in the putative local tRNA pool in the vicinity of the codons.

Intuitively, differences in translation efficiency of individual genes may be thought of as a means for regulation of the expression of the gene's products, namely – individual proteins. However, we reveal that a major dichotomy in mammalian cells between the codon usage of proliferation- and differentiation-related genes allows simultaneous changes in the translation efficiency of such entire gene sets, via coordinated changes in the tRNA pool. Indeed, our study suggests that cancer elevates the expression of tRNAs whose codons are enriched among the proliferation genes, while repressing the tRNAs that translate the differentiation genes. Such alternation in the tRNA pool might act as a feedback loop, which in turn promotes the cancerous process. All together, our results suggest that dynamics in translation efficiency affects the cell fate and that the tRNA expression profile in the cell at any moment may indicate for the state of the cell in normal physiology, and even further reflect oncogenic pathway signatures in abnormal physiology.

We also proposed new analytic tool to mine sequence and expression to decipher gene translation. Considering both the new conceptions introduced in this study and the new analytic tools introduced, this thesis may constitute the first stage towards consolidation of comprehensive dynamic model of translation efficiency, which may serve as a sensor for both the composition of the proteome and the status of the cell

10. References

Akashi H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* **136**(3): 927-935.

Barbarese E, Koppel DE, Deutscher MP, Smith CL, Ainger K, Morgan F, Carson JH. 1995. Protein translation components are colocalized in granules in oligodendrocytes. *J Cell Sci* **108 (Pt 8)**: 2781-2790.

Barski A, Chepelev I, Liko D, Cuddapah S, Fleming AB, Birch J, Cui K, White RJ, Zhao K. 2010. Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. *Nat Struct Mol Biol* **17**(5): 629-634.

Berg OG, Silva PJ. 1997. Codon bias in Escherichia coli: the influence of codon context on mutation and selection. *Nucleic Acids Res* **25**(7): 1397-1404.

Bucciantini M, Giannoni E, Chiti F, Baroni F, Formigli L, Zurdo J, Taddei N, Ramponi G, Dobson CM, Stefani M. 2002. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* **416**(6880): 507-511.

Buchan JR, Aucott LS, Stansfield I. 2006. tRNA properties help shape codon pair preferences in open reading frames. *Nucleic Acids Res* **34**(3): 1015-1027.

Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y. 2010. A role for codon order in translation dynamics. *Cell* **141**(2): 355-367.

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**(1): 65-73.

Comeron JM. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics* **167**(3): 1293-1304.

Cornut B, Willson RC. 1991. Measurement of translational accuracy in vivo: missense reporting using inactive enzyme mutants. *Biochimie* **73**(12): 1567-1572.

Crick FH. 1966. Codon--anticodon pairing: the wobble hypothesis. *J Mol Biol* **19**(2): 548-555.

Dairkee SH, Ji Y, Ben Y, Moore DH, Meng Z, Jeffrey SS. 2004. A molecular 'signature' of primary breast cancer cultures; patterns resembling tumor tissue. *BMC Genomics* **5**(1): 47.

Dittmar KA, Goodenbour JM, Pan T. 2006. Tissue-specific differences in human transfer RNA expression. *PLoS Genet* **2**(12): e221.

dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* **32**(17): 5036-5044.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**(2): 341-352.

Dunham I Kundaje A Aldred SF Collins PJ Davis CA Doyle F Epstein CB Frietze S Harrow J Kaul R et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.

Duret L. 2000. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet* **16**(7): 287-289.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc Natl Acad Sci U S A* **96**(8): 4482-4487.
Edelmann P, Gallant J. 1977. Mistranslation in E. coli. *Cell* **10**(1): 131-137.
Elf J, Nilsson D, Tenson T, Ehrenberg M. 2003. Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science* **300**(5626): 1718-1722.
Farabaugh PJ, Bjork GR. 1999. How translational accuracy influences reading frame maintenance. *Embo J* **18**(6): 1427-1434.
Flemming AJ, Shen ZZ, Cunha A, Emmons SW, Leroi AM. 2000. Somatic polyploidization and cellular proliferation drive body size evolution in nematodes. *Proc Natl Acad Sci U S A* **97**(10): 5285-5290.
Fluitt A, Pienaar E, Viljoen H. 2007. Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. *Comput Biol Chem* **31**(5-6): 335-346.
Gilchrist MA, Shah P, Zaretzki R. 2009. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. *Genetics* **183**(4): 1493-1505.
Gregersen N. 2006. Protein misfolding disorders: pathogenesis and intervention. *J Inherit Metab Dis* **29**(2-3): 456-470.
Gutman GA, Hatfield GW. 1989. Nonrandom utilization of codon pairs in Escherichia coli. *Proc Natl Acad Sci U S A* **86**(10): 3699-3703.
Hanahan D, Weinberg RA. 2000. The hallmarks of cancer. *Cell* **100**(1): 57-70.
Heger A, Ponting CP. 2007. Variable strength of translational selection among 12 Drosophila species. *Genetics* **177**(3): 1337-1348.
Hsieh AC, Liu Y, Edlind MP, Ingolia NT, Janes MR, Sher A, Shi EY, Stumpf CR, Christensen C, Bonham MJ et al. 2012. The translational landscape of mTOR signalling steers cancer initiation and metastasis. *Nature* **485**(7396): 55-61.
Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol* **151**(3): 389-409.
Ikemura T, Ozeki H. 1983. Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. *Cold Spring Harb Symp Quant Biol* **47 Pt 2**: 1087-1097.
Kaminska M, Havrylenko S, Decottignies P, Le Marechal P, Negrutskii B, Mirande M. 2009. Dynamic Organization of Aminoacyl-tRNA Synthetase Complexes in the Cytoplasm of Human Cells. *J Biol Chem* **284**(20): 13746-13754.
Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J Mol Evol* **53**(4-5): 290-298.
Karantza V. 2011. Keratins in health and cancer: more than mere epithelial cell markers. *Oncogene* **30**(2): 127-138.
Khazaie K, Buchanan JH, Rosenberger RF. 1984. The accuracy of Q beta RNA translation. 1. Errors during the synthesis of Q beta proteins by intact Escherichia coli cells. *Eur J Biochem* **144**(3): 485-489.
Klochendler A, Weinberg-Corem N, Moran M, Swisa A, Pochet N, Savova V, Vikesa J, Van de Peer Y, Brandeis M, Regev A et al. 2012. A transgenic mouse marking live replicating cells reveals in vivo transcriptional program of proliferation. *Dev Cell* **23**(4): 681-690.

Kramer EB, Farabaugh PJ. 2007. The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *Rna* **13**(1): 87-96.
Lavner Y, Kotlar D. 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene* **345**(1): 127-138.
Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD. 2003. A unification of mosaic structures in the human genome. *Hum Mol Genet* **12**(19): 2411-2415.
Lithwick G, Margalit H. 2003. Hierarchy of sequence-dependent features associated with prokaryotic translation. *Genome Res* **13**(12): 2665-2673.
Mamane Y, Petroulakis E, LeBacquer O, Sonenberg N. 2006. mTOR, translation initiation and cancer. *Oncogene* **25**(48): 6416-6422.
Man O, Pilpel Y. 2007. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat Genet* **39**(3): 415-421.
Meyerovich M, Mamou G, Ben-Yehuda S. 2010. Visualizing high error levels during gene expression in living bacterial cells. *Proc Natl Acad Sci U S A* **107**(25): 11543-11548.
Moura G, Pinheiro M, Arrais J, Gomes AC, Carreto L, Freitas A, Oliveira JL, Santos MA. 2007. Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure. *PLoS ONE* **2**(9): e847.
Oler AJ, Alla RK, Roberts DN, Wong A, Hollenhorst PC, Chandler KJ, Cassiday PA, Nelson CA, Hagedorn CH, Graves BJ et al. 2010. Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat Struct Mol Biol* **17**(5): 620-628.
Parker J. 1989. Errors and alternatives in reading the universal genetic code. *Microbiol Rev* **53**(3): 273-298.
Parker J, Holtz G. 1984. Control of basal-level codon misreading in Escherichia coli. *Biochem Biophys Res Commun* **121**(2): 487-492.
Pavon-Eternod M, Gomes S, Geslain R, Dai Q, Rosner MR, Pan T. 2009. tRNA over-expression in breast cancer and functional consequences. *Nucleic Acids Res* **37**(21): 7268-7280.
Pavon-Eternod M, Gomes S, Rosner MR, Pan T. 2013. Overexpression of initiator methionine tRNA leads to global reprogramming of tRNA expression and increased proliferation in human epithelial cells. *Rna*.
Precup J, Parker J. 1987. Missense misreading of asparagine codons as a function of codon identity and context. *J Biol Chem* **262**(23): 11351-11355.
Rodnina MV, Wintermeyer W. 2001. Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanisms. *Annu Rev Biochem* **70**: 415-435.
Sonenberg N, Hinnebusch AG. 2009. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136**(4): 731-745.
Sorensen MA. 2001. Charging levels of four tRNA species in Escherichia coli Rel(+) and Rel(-) strains during amino acid starvation: a simple model for the effect of ppGpp on translational accuracy. *J Mol Biol* **307**(3): 785-798.
Stansfield I, Jones KM, Herbert P, Lewendon A, Shaw WV, Tuite MF. 1998. Missense translation errors in Saccharomyces cerevisiae. *J Mol Biol* **282**(1): 13-24.
Stefani M, Dobson CM. 2003. Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J Mol Med* **81**(11): 678-699.
Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Mol Biol Evol* **24**(2): 374-381.

Tabach Y, Milyavsky M, Shats I, Brosh R, Zuk O, Yitzhaky A, Mantovani R, Domany E, Rotter V, Pilpel Y. 2005. The promoters of human cell cycle genes integrate signals from two tumor suppressive pathways during cellular transformation. *Mol Syst Biol* **1**: 2005 0022.
Toth MJ, Murgola EJ, Schimmel P. 1988. Evidence for a unique first position codon-anticodon mismatch in vivo. *J Mol Biol* **201**(2): 451-454.
Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**(2): 344-354.
Vander Heiden MG, Cantley LC, Thompson CB. 2009. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* **324**(5930): 1029-1033.
Warnecke T, Hurst LD. 2010. GroEL dependency affects codon usage--support for a critical role of misfolding in gene evolution. *Mol Syst Biol* **6**: 340.
Yang JH, Shao P, Zhou H, Chen YQ, Qu LH. 2010a. deepBase: a database for deeply annotating and mining deep sequencing data. *Nucleic Acids Res* **38**(Database issue): D123-130.
Yang JR, Zhuang SM, Zhang J. 2010b. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol Syst Biol* **6**: 421.
Zaborske JM, Narasimhan J, Jiang L, Wek SA, Dittmar KA, Freimoser F, Pan T, Wek RC. 2009. Genome-wide analysis of tRNA charging and activation of the eIF2 kinase Gcn2p. *J Biol Chem* **284**(37): 25254-25267.
Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol* **26**(7): 1571-1580.
Zouridis H, Hatzimanikatis V. 2008. Effects of codon distributions and tRNA competition on protein translation. *Biophys J* **95**(3): 1018-1033.

11. תקציר

במודל מקיף המעריך יעילות תרגום, התהליך צריך להיבחן כתהליך של ביקוש מול היצע, כאשר ההיצע ניתן על ידי זמינות מולקולות רנא מוביל, והביקוש הוא הייצוג הממשי של הקודונים במולקולות ה-רנא שליח המצויות בתא. מודלים רווחים של יעילות התרגום של גנים מניחים לרוב כי התהליך קבוע עבור כל גן לאורך חיי האורגניזם. ברם, מחקרים שהתפרסמו לאחרונה חושפים תמונה מורכבת יותר של דינמיות במאגר ה-רנא מוביל התאי. לקראת הדור הבא של מודלים של יעילות תרגום, אנו למדנו את הגורמים המניעים היצע דינמי של רנא מוביל, לצד הדינמיקה המשלימה האפשרית של הביקוש ל-רנא מוביל מצד הקודונים של הגנים. ההנחה העומדת בבסיסו של הצורך במודל מסוג היצע וביקוש היא שאם קודונים מסוימים מופיעים פעמים רבות במולקולות רנא שליח רבות בתנאי מחיה מסויימים, אזי יעילות התרגום שלהם תקטן, גם אם היצע ה-רנא מוביל עבור קודונים אלו גבוה. תזה זו כוללת ארבעה פרקים בהם חקרתי היבטים שונים של הדינמיות בהיצע וביקוש בתהליך התרגום.

בפרק הראשון, אנו חושפים נטייה כללית של מינים שונים להגדיל את הייצוג של קודונים המאופיינים ביעילות תרגום נמוכה בתא, כאשר הוא מצוי בתנאי עקה. מגמה זו מרמזת על תרגום לא יעיל של גנים המתבטאים במצבי עקה, שמקורו ככל הנראה בהיעדר לחץ אבולוציוני מספק לטובת שימוש בקודונים יעילים בגנים אלו.

תחת ההנחה שיעילות תרגום היא תכונה דינמית, בפרק השני אנו בוחנים את קיומם של שינויים אפשריים בהיצע וביקוש בתהליך התרגום בתאים סרטניים. לצורך כך, ניתחנו תוצאות מדידה של רמות ביטוי של מולקולות רנא שליח ורנא מוביל גילינו שרמות הביטוי של סוגים שונים של רנא מוביל משתנות בתאים סרטניים בצורה הדירה, ובפרט, חישבנו ומצאנו כי מאגר ה-רנא מוביל בתאים סרטניים צפוי להגביר באופן בררני את יעילות התרגום של גנים הקשורים בתהליכים של ריבוי ושגשוג תאים. יתר על כן, אנו מראים כי השונות בהשפעה של מאגר ה-רנא מוביל הסרטני על יעילות התרגום של קבוצות גנים שונות נובעת ביסודה מדיכוטומיה, דהיינו הפרדה חותכת בין הקודונים השכיחים בגנים הקשורים בתהליכי ריבוי ושגשוג של תאים, לבין הקודונים הנפוצים בגנים המעורבים בתהליכי התמיינות. אבחנה זו בין קבוצות הגנים הנזכרות לעיל, על בסיס העדפות מובחנות של קודונים שונים, לא זוהתה טרם מחקרנו זה. באופן ספציפי, נראה כי סרטן מגביר את רמות הביטוי של סוגי רנא מוביל אשר מתרגמים קודונים הנפוצים בגנים הקשורים בריבוי ושגשוג תאים. בד בבד נראה כי סרטן מפחית את רמות הביטוי של סוגי רנא מוביל אשר מתרגמים קודונים השכיחים בגנים הקשורים בתהליכי התמיינות. למעשה, אנו מראים כי מאגר ה-רנא מוביל בתאים סרטניים מגביר את יעילות התרגום של אותם הגנים אשר רמת הביטוי של ה-רנא שליח שלהם עולה בסרטן. ממצא זה מעלה את האפשרות ששינויים ביעילות התרגום מתווכים החלפה בין מצב של ריבוי ושגשוג תאים לבין מצב של התמיינות תאים בעת תפקוד נורמאלי כמו גם בסרטן.

בחלק השלישי של התזה אנו קוראים תיגר על הקונספציה המסורתית של יעילות תרגום, בהציענו כי מאגרים מקומיים של מולקולות רנא מוביל "ממוחזר" המצויות בקרבה לקודון המתורגם באתר ה"A" על גבי הריבוזום עשויים להעלות את יעילות התרגום שלו. בדקנו ומצאנו כי במופעים עוקבים של אותה חומצת אמינו, גנים בעלי רמות ביטוי גבוהות נוטים להשתמש באותו הקודון. תוצאה זו תומכת בהשערה שלנו בדבר קיומו של מאגר רנא מוביל מקומי. בנוסף, גילינו כי עמדות שמורות של חומצות אמינו מסוימות בגנים של שמרים נוטות להיות מקודדות ע"י קודונים אשר, באופן יחסי, מאופיינים בסבירות נמוכה לקשור רנא מוביל עם חומצת אמינו שגויה, אף אם קודונים אלו נחותים בהיבט של יעילות התרגום. ממצא זה עומד בסתירה להנחה הרווחת הקושרת בין תרגום מדוייק ובין שימוש מועדף בקודונים המאופיינים ביעילות תרגום גבוהה.

בסיכומו של דבר התיזה שלי מניחה יסודות למודל חדש ליעילות ודיוק תרגום של חלבונים. במודל זה דינאמיות של כל רכיבי התהליך מגולמת מפורשות, דבר המסייע להשגת תאור מהימן של הפיכת הטרנסקריפטום לפרוטאום.