Deciphering the codons’ code:
Selection beyond amino acid sequence.

November 2013
# Table of Contents

1 Abstract .......................................................................................................................... 3

2 תקציר ............................................................................................................................... 4

3 Introduction ......................................................................................................................... 5

3.1 Codon Usage and Bias .................................................................................................. 5

3.2 Translation Efficiency ................................................................................................. 6

3.3 The tRNA pool ............................................................................................................. 8

3.4 Transcript Degradation ............................................................................................... 9

4 Methods ............................................................................................................................ 11

4.1 The role of codon selection in regulation of translation efficiency deduced from synthetic libraries (Navon et al. Genome Biology 2011) ............................................................... 11

4.2 A Comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool .......................................................................................................................... 14

4.3 Ribosome density governs patterns of mRNA cleavage in *Escherichia coli* 16

5 Results ................................................................................................................................ 17

5.1 The role of codon selection in regulation of translation efficiency deduced from synthetic libraries (Navon et al. Genome Biology 2011) ............................................................... 17

5.2 A Comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool .......................................................................................................................... 47

5.3 Ribosome density governs patterns of mRNA cleavage in *Escherichia coli* 54

6 Discussion .......................................................................................................................... 71

7 Literature ............................................................................................................................ 76

7.1 References .................................................................................................................... 76

7.2 List of publications ....................................................................................................... 80

8 Declaration ......................................................................................................................... 81

9 Appendix ............................................................................................................................ 82

9.1 A Comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool (Bloom-Ackerman *et al.*) ......................................................................................... 82
1 Abstract

The base of protein production is the translation process in which the codon sequence is decoded into amino acids by tRNAs. The actual codons used for a certain protein can influence different aspects of its gene expression. Codon choices were shown to influence protein levels and folding, mRNA stability and even the host fitness. In my research I focused on how codon usage affects different aspects of gene expression. My research includes a combination of three projects meant to better understand the sequence effects on the gene’s translation on one hand, and its mRNA stability on the other.

In my first project I examine how the distribution of codons along the mRNA affects the protein levels. Here I reanalyzed data from two experiments in Escherichia coli, examining how regions with “slow codons”, i.e. codons which take longer to be translated, influence the translation rate and its efficiency. I found that localizing the slowest codons in the 5’-end results in higher protein levels and the slower this region is the higher the protein levels.

A major parameter which affects the codon’s translational speed is the availability of the tRNA which translates it. In my second project we studied the tRNA pool of S. cerevisiae through a comprehensive deletion library of tRNAs. We found extensive backup between the tRNA copies and differential contribution to the cell’s fitness of identical tRNA copies. In addition, we found up-regulation of RNA Polymerase III genes in some deletion strains all suggesting additional levels of regulation of the tRNA transcription.

In my third project I studied the mRNA degradation in Escherichia coli and its coupling to the translation process. By combining mapping of 5’ RNA fragments with ribosome profiling data, I studied the relationship between the gene’s translation and degradation properties. I suggest that the ribosomes play a dual role in mRNA degradation: on one hand they enhance cleavage of the transcript in its local vicinity, immediately up stream to its attenuation site, yet globally on the transcript they protect it from degradation.

Together my studies shed light on some of the forces behind codon usage. My studies show how codons can attenuate ribosome elongation and reveal that ribosome attenuation is also coupled with the mRNA degradation process.
The basic step in protein production is the translation, a process in which codon sequences are decoded into amino acids by ribosomes. The specific codons can affect various aspects of gene expression. It has been found that codons can influence protein levels and their folding, ribosome stability, and even cell health. My research focused on how codons influence different aspects of gene expression. My research was a combination of three projects, which aimed to shed light on the effects of codon sequences on translation of the gene from one direction, and ribosome stability from the other.

In my first project, I examined how codon distribution along the ribosome affects protein levels. In my second project, I analyzed data from two experiments regarding how areas with "slow codons", i.e., codons that take longer to be translated, affect translation rate and efficiency. I found that the location of the slowest codons at the 5' end is in line with higher protein levels, and the slower the area, the higher the protein levels.

A crucial parameter that affects translation speed is tRNA availability.

In my third project, I studied the deconstruction of ribosomes and their role in the translation process. By combining information about ribosome locations in the cell with translation profiles, I examined the relationship between translation and deconstruction of the ribosome. The results suggest that ribosomes play a dual role in deconstruction: on the one hand, they increase the chance of ribosome cleavage at a specific position, and on the other, they globally protect the ribosome from degradation.

The different parts of my research shed light on the factors affecting the use of different codons. My research shows that codons can affect ribosome movement and reveals a connection between translation and ribosome deconstruction.
3 Introduction

3.1 Codon Usage and Bias

During the translation process the nucleotide sequence of a gene is translated into the amino acid sequence of a protein. For translation, the nucleotides are grouped into triplets, resulting in 64 codons; these codons are commonly translated to 20 amino acids and stop codons. As a result, 18 out of the 20 amino-acids are translated by two to six different codons, called synonymous codons.

Although translated into the same amino acid, synonymous codons are not perceived the same by the translation system. Commonly across various genomes, there are around 40 different tRNA types per species. Some codons are translated by a fully matching tRNA while others are translated by non-fully matched ones, through the wobble interactions, originally described by Crick (Crick, 1966). According to the wobble hypothesis only the first two positions of a triplet codon need to be precisely paired with the tRNA anti-codon, while the pairing of the third nucleotide position of the codon may be ambiguous, and it varies according to the nucleotide present in this position. Crick’s wobble hypothesis was later extended to take into account different covalent modifications that occur on the tRNA molecule (Yarus, 1982, Agris, 1991, Agris, 2004). These modifications which are usually on the 34th nucleotide of the tRNA molecule (the first nucleotide of the anti-codon pair) either disable potential pairing between codon-anticodon pairing of different amino-acids or enable better pairing when the third position does not match (Yarian et al., 2002, reviewed in: Agris, 2004).

Two major observations suggest that there should be significant differences between codons that still code for the same amino acid. The first, the tRNA levels in the cytoplasm can be significantly different between different tRNAs for the same amino acid (Ikemura, 1981, Ikemura, 1982). Although relatively few measurements of such tRNA abundance levels were done (Ikemura, 1981, Ikemura, 1982, Dittmar et al., 2005, Zaborske et al., 2009) a quite reliable proxy for these levels was found in the form of the tRNA gene copy number (tGCN) (Percudani et al., 1997, Kanaya et al., 1999, Tuller et al., 2010a). From this
genome proxy, which can be examined at every fully sequenced genome, it is clear that the various codons are translated by a set of tRNAs which are very different in their relative abundance. The second observation is that there is a codon bias in genes, namely a nonrandom use of synonymous codons found in genomes of both unicellular and multi-cellular organisms (Ikemura, 1985). This bias was found to be particularly pronounced for highly express genes (Sharp & Li, 1986). It was also found that the genome codon bias correlates to the tGCN (Ikemura, 1981, Ikemura, 1982, Duret, 2000), suggesting connection between codon bias and tGCN to translation efficiency. Particularly, highly expressed genes were found to be preferentially encoded by codons that correspond to high abundance tRNAs (Sharp & Li, 1986).

3.2 Translation Efficiency

Based on the fact that highly expressed genes are biased toward “optimal” codons which correspond to the most abundant tRNA genes in the cell (Sharp & Li, 1986), two indices which calculate the adaptation of a sequence to the cellular translation capabilities were developed. The first index is the Codon Adaptation Index (CAI) (Sharp & Li, 1987). For each species a reference set of highly expressed genes is used to assess the relative translation efficiency score of each codon. Then a score for a gene can be calculated by the geometric mean of the translation efficiency values of its codons: \( CAI = \left( \prod_{k=1}^{L} W_k \right)^{1/L} \) where \( L \) is the number of codons in the gene and \( W_k \) is the relative translation efficiency score of the \( k \)th codon. This value reflects the usage frequency of the gene’s codons in the set of highly expressed genes. The second index is the tRNA Adaptation index (tAI) (dos Reis et al., 2004). This index assumes that the codon efficiency is derived from its tRNA concentration and its tRNA binding affinity. The affinities between each tRNA to each codon were calculated by optimizing the tAI index for highly expressed genes in yeast; these affinities are assumed to be the same for all organisms. As a proxy of the tRNA concentration the tRNA gene copy number (tGCN) of each tRNA type in every species is used. After the initial calculation of affinities which was done in the original paper (dos Reis et al., 2004), the species’ codons efficiency (\( W_i \)) can easily be
calculated by simply multiplying its tGCN by the previously calculate affinities \((1 - s_{ij})\), \(W_i = \sum_{j=1}^{n_l} (1 - s_{ij}) \cdot tGCN_{ij}\). As for the CAI, the tAI score of a gene is calculated as the genomic mean of all its codons: \(tAI = \left( \prod_{k=1}^{L} W_k \right)^{\frac{1}{L}}\) where \(L\) is the number of codons in the gene and \(w_k\) is the normalized codon’s efficiency of the \(k\)th codon \(w_k = W_k / \max(W_i)\).

Although these indices were shown to be good predictors for expression of natural genes (Sharp & Li, 1986, Man & Pilpel, 2007), in recent synthetic libraries experiments they failed to predict the abundance of a protein across different nucleotide sequence variants (Kudla et al., 2009, Welch et al., 2009). This failure suggests other sequence parameters affect the translation efficiency, shedding light on one of the potential limitations of both indices: they are indifferent to the order of the codons along genes, scoring equally alternative genes with same codon composition and yet different order of high and low efficiency codons.

In recent years other sequence parameters were shown to also affect translation efficiency. One of the most significant parameter is the mRNA secondary structure, in particular the folding energy of the ribosome binding site (RBS) region, which is also affected by codons close to the ATG (Lu et al., 2007, Kudla et al., 2009, Gu et al., 2010, Tuller et al., 2010b, Goodman et al., 2013). In addition codons near the ORF start were found to be under selection for low folding energy (Gu et al., 2010) but also for less efficient codons (Tuller et al., 2010a); suggesting that the codons regulate the flow of ribosomes on the transcript (Tuller et al., 2010a). To date it is quite clear that changes to the secondary structure significantly affect translation efficiency (Goodman et al., 2013, Kudla et al., 2009). Yet it is still unclear to what extent changes in each of these sequence parameters away from the ATG affects the translation efficiency of a given transcript and whether we have identified them all. However, it is clear that changes of synonymous codons, while not changing the amino acid sequence, do influence all the above sequence parameters thus changing translation efficiency.
### 3.3 The tRNA pool

The tRNA (transfer RNA) is one of the fundamental components of the translation machinery. The tRNA pool, numerically represented by the tRNA concentrations in the cell, was shown to play an important role in the dynamics of translation in previous papers (Ikemura, 1981, Ikemura, 1982, Dittmar et al., 2005, Man & Pilpel, 2007, Subramaniam et al., 2013). This is attributed to the fact that the rate limiting step in the elongation cycle of polypeptides is tRNA search and selection, (Varenne et al., 1984)

The tRNAs are transcribed by RNA Polymerase III (Pol III) which, along with RNA Polymerase I, dominates cellular transcription, together reaching more than 80% of total RNA synthesis in growing cells (Paule & White, 2000). The tRNA promoters are composed of two highly conserved sequence elements that reside within the transcribed region called the A-box and the B-box (reviewed in: Paule & White, 2000). For many years these two internal promoter elements along with the Pol III terminator were considered the main elements to regulate tRNA transcription. However, more recent papers have shown that the tRNA transcription levels also depend on upstream regulatory motifs (Giuliodori et al., 2003, Parthasarthy & Gopinathan, 2006).

tRNA genes may be present in multiple copies in a genome, e.g, *S. cerevisiae* has 274 tRNA genes encoding 42 different isoacceptors, human has 506 tRNAs encoding 48 different isoacceptors, and *E. coli* K12 has 86 tRNAs with 39 different isoacceptors (Chan & Lowe, 2009). It is assumed that the requirement for large quantities of tRNAs is met by multiple copies of the tRNA gene in addition to multiple rounds of transcription of each tRNA gene (Duret, 2000, Percudani et al., 1997, Kanaya et al., 1999). Indeed it was shown that the tRNA genome copy number is a reliable proxy for the tRNA levels (Percudani et al., 1997, Kanaya et al., 1999, Tuller et al., 2010a).

The tRNA pool is often assumed to be constant throughout the life of a cell (Tuller et al., 2010a, dos Reis et al., 2004). However recent papers reveal a more complex picture. Absolute tRNA levels were shown to change across growth conditions (Dittmar et al., 2005) and in different tissues (Dittmar et al., 2006). The binding of Pol III to the genome was also shown to vary between the
different tRNAs genes (Canella et al., 2010, Kutter et al., 2011). These new finding suggests the tRNA pool has the potential to be dynamic and to undergo regulated change in different growth conditions. A dynamic tRNA pool could grant the cell the ability to adapt the translation efficiencies to better fit the immediate demand. The questions of how dynamic is the tRNA pool and what are the mechanism that enable it to change still remains to be answered.

3.4 Transcript Degradation

The amount of protein synthesized in the cell is determined by two primary parameters: the mRNA levels of the genes and the rate of translation of this mRNA. mRNA levels in turn, are determined by both the rate of transcription and the rate of degradation. In bacteria the half-life of different transcripts varies significantly, from less than a minute to about an hour (Bernstein et al., 2002), thus allowing the stability of the mRNA to play a key role in regulating the transcriptome.

A mature bacterial transcript initially bears a triphosphate at its 5’ and a stem & loop structure at its 3’. In Escherichia coli, the degradation of most mRNAs is thought to begin with an internal endonucleolytic cleavage (Deana & Belasco, 2005), resulting in two RNA fragments. The upstream fragment is no longer protected by a 3’ stem loop and is thus promptly degraded by 3’ exonucleases. The downstream fragment is only monophosphorylated, leading to its rapid degradation by a series of internal cleavages and 3’ degradations (Deana & Belasco, 2005).

Since in bacteria the degradation is initiated from an internal cleavage, it is not surprising that translation parameters influence the mRNA degradation. Two ribosome-related mechanisms have a potential to stabilize the mRNA (reviewed in: Deana & Belasco, 2005). The first, closely spaced translating ribosomes can protect potential cleavage sites. The second, ribosome occupying the 5’ terminus of the transcript can impede access to cleavage site by disrupting the cleavage mechanism. Indeed it was found that a ribosome binding to the RBS helps to protect mRNAs, and as a consequence a less efficient RBS can give rise to destabilization of the mRNA (reviewed in: Deana & Belasco, 2005). On the other hand, stalled ribosomes could have either a stabilizing effect (Bechhofer &
or a destabilizing affect (Hayes & Sauer, 2003, Sunohara et al., 2004). In particularly it was shown that clusters of rare codon have a destabilizing effect (Sunohara et al., 2004). Furthermore stalled ribosomes were found to causes cleavage in variety of location without an identified nucleases, naming them “killer ribosomes” (reviewed in: Dreyfus, 2009).

It thus appears from biochemical analyses that the control of translation and mRNA degradation might be coupled. In particular features that affect translation efficiency may also influence the decay rate of the corresponding mRNAs. As described above, different codons have the potential to change both the secondary structure of the transcript and the ribosomes flow along it. As a result, degradation of mRNA might be sensitive to, and perhaps regulated by codon usage. To date the exact features which determine where and when a transcript will be cleaved are still unknown. Exact cleavages sites were mapped only for a handful of genes over the years (Cormack & Mackie, 1992, McDowall et al., 1994, Braun et al., 1996, Ehretsmann et al., 1992) and the nucleases cleavage preferences were not completely characterize. Thus, any hope to predict mRNA stability requires more work to understand both the cleavage mechanism and its potential coupling with translation.
4 Methods

4.1 The role of codon selection in regulation of translation efficiency deduced from synthetic libraries (Navon et al. Genome Biology 2011)

4.1.1 Defining the bottleneck

In this project we analyzed how a region of non-optimal codons can affect translation. We defined this region as the bottleneck, a region on the gene where the harmonic mean of the codons’ tAI (dos Reis et al., 2004) values is minimal. The codon tAI values are assumed to be proportional to the speed of its translation (Tuller et al., 2010a); higher tAI value, correspond to high tRNA abundance and affinity, thus faster translation. A harmonic mean of speeds is simply an arithmetic mean of the corresponding times. Hence looking for the region with the minimum harmonic mean of speed is equivalent to looking for the region which takes the longest time to translate. For each region the harmonic mean of speed is:

\[ \frac{1}{n} \sum_{c \in \text{Region}} \frac{1}{tAI_c} \]

where \( n \) is the region size, and \( c \) is the set of all the codons in the region (\( n \) codons). For more details see the Methods in the paper attached to the results chapter, 5.1

4.1.2 Choosing the bottleneck window size (\( n \))

Under a maximal density scenario (fast initiation rate), the distance between two consecutive ribosomes will be minimal. In this case, when two ribosomes are translating the same mRNA simultaneously, the minimum possible distance between the two translated codon (one by each of the ribosome) is one ribosome size (\( H \) codons). At any given moment during the translation process, two adjacent ribosomes would have translated exactly the same codons apart from the last \( H \) codons - the first of the two ribosomes has already translated them, and second is just about to start them. If the time it took the first ribosome to finish translating the \( n \)th codon \( T(n,1) \) is larger than the time it takes the second ribosome to translate the \( n-H \)th codon \( T(n-H,2) \) the second ribosome will “bump” into the first one. Therefore the region of \( H \) codons with maximum translation time determines whether and where a traffic jam will be
created (for detailed calculations see the Methods in the paper attached to the results chapter, 5.1). Consequently the minimal distance between two ribosomes should determine our window size. We adopted the average ribosome-to-ribosome distance measured by Brandt et al. (Brandt et al., 2009). They measured the mean distance between the center of mass of two ribosomes on actual bacterial polysomes to be 21.6\( \text{nm} \) which is about 21 codons (0.34\( \text{nm} \) per base). For more details see the Methods in the paper attached to the results chapter, 5.1.

4.1.3 The bottleneck parameters
A bottleneck is characterized by two parameters: its “location” and its “strength”:

The “location” of the bottleneck is defined as the location in the gene of the bottleneck’s first codon (\( k \) codons from the ATG).

The “strength” of the bottleneck is defined as the arithmetic average of \( 1/tAI \) values for the codons in the region, e.g. \( \frac{1}{n} \sum_{c \in \text{region}} \frac{1}{tAI_c} \) (the inverse of the harmonic mean).

The relative strength of the bottleneck is defined as the strength of the bottleneck divided by the average \( 1/tAI \) for the entire gene, e.g.

\[ \frac{1}{n} \sum_{c \in \text{bottleneck}} \frac{1}{tAI_c} \text{ \text{1 \text{sum} \text{over \text{bottleneck}}}} \] ; where \( l \) is the number of codon in the gene (excluding the stop codon).

The relative location of the bottleneck is defined as the location of the bottleneck divided by number of possible windows; e.g. \( \frac{k}{l-n+1} \) ; where \( k \) is the location of the bottleneck, and \( l \) is the length of the gene and \( n \) is the window size.
4.1.4 Finding the main anti-correlated codons

We used partial correlation to find the codons which contribute the most to the decrease in the cell’s fitness. The highest contributors were filtered according to the following steps:

1. Find codons which have a negative correlation to the OD.
2. For all codons left, we calculated the partial correlation matrix $M(i,j) =$ partial correlation (codon $i$, OD $|$ codon $j$).
3. Find the minimum absolute value of the partial correlation, for each codon.
   
   Rank the codons in a descending order accordingly.

For more details see the Methods in the paper attached to the results chapter, 5.1.

4.1.5 Calculating the codon usage in the genome and transcriptome

To calculate the codon usage in the genome we counted each codon appearance in all the ORFs and normalized by the total number of codons. For this analysis we used the genome of $E. coli$ B21 strain (which was used by Kudla et al.) was downloaded from NCBI; [Refseq: NC_012947 (Jan 11, 2010)].

To calculate the codon usage of the transcriptome of each gene was calculated by multiplying the mRNA levels measurements for the gene by the codon usage of the same gene. The contributions of all genes were summed for each codon and then divided by the total sum of all codons. mRNA levels were taken from Lu et al. (Lu et al., 2007). For details see the Methods in the paper attached to the results chapter, 5.1.
4.2 A Comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool

4.2.1 Characterizing the growth dynamics and the phenotype of a deletion strain

To assess the contribution of each tRNA gene to cellular growth, we characterized the growth dynamics of each deletion strain. We chose to characterize each deletion strain by two growth parameters: growth rate and the size of the population upon entering the stationary phase, denoted “growth yield”. For each strain the relative-growth-rate and relative-growth-yield were calculated by normalizing its parameters to the wild type parameters measured in the same experiment. These parameters are then projected on a distribution of the wild type growth parameters (which was created by measuring the wild type multiple times in multiple days) and the number of standard deviations is calculated (\( \sigma \)), see Figure 1. Any deviations larger than two standard deviations from the mean were considered as “phenotypes” and deviations above three standard deviations were considered a “strong phenotype”. A negative deviation denotes impairment (worse than the wild type (WT)) while a positive deviation denotes improvement (better than the WT).

\[
\sigma = \frac{\text{Relative Growth} - 1}{\text{WT std}}
\]

![Figure 1. Characterizing a deletion strain. Schematic representation of growth measurements, analysis, and scoring. For each strain relative-growth-rate and relative-growth-yield were calculated in relation to the wild type strain. These parameters are then projected on a distribution of the wild type growth parameters. Sigma (\( \sigma \)) is denoted the number of standard deviations from the mean of the wild type.](image-url)
4.2.2 Motif Analysis

A sequence motif analysis was performed using the MEME online software (Bailey & Elkan, 1994) to identify motifs which are can explain difference in tRNA deletion strain fitness. The motif search was done on the upstream sequence of tRNA genes which exhibited a yield impairment phenotype in rich medium upon deletion (42 genes) versus the upstream sequence of tRNA genes which exhibited a phenotype in no more than two out of the six measured conditions (99 genes). To apply location constrains on the motifs, the MEME analysis was done in windows of size 9bp, looking for motifs of 4-8bp in length.

4.2.3 Microarray analysis

For five tRNA deletion strains the transcriptome was measured using Affymetrix yeast 2.0 microarrays. The microarray background adjustment was done using the Robust Multi-array Average (RMA) procedure followed by quintile normalization. For each strain, the fold change in expression for all genes was calculated by comparing the wild-type measurement in the same batch and averaged over two biological repeats.

The five strains were clustered. The cluster tree was based on the correlation between the mRNA fold changes of the different strains. For the clustering we used the top 50% of the sorted genes based by the gene variance across the strains.

4.2.4 Gene Set Enrichment

To expose what are the responses and underling molecular pathways that differentiate these two groups of tRNA deletion strains, we used Gene Set Enrichment Analysis (GSEA) software. GSEA computationally determines whether pre-defined set of genes shows statistically significant difference in representation between two biological groups (Subramanian et al., 2005, Mootha et al., 2003). For the pre-defined set of genes we used S. cerevisiae pathways as defined by KEGG (Kanehisa & Goto, 2000, Kanehisa et al., 2012).
4.3 Ribosome density governs patterns of mRNA cleavage in *Escherichia coli*

4.3.1 Identifying cleavage sites

In order to identify cleavage sites for both *E. coli* and *P. aeruginosa* we used data derived from a modified RNA deep sequencing protocol (Wurtzel *et al.*, 2010) to deduce location of endonucleolytic cleavage sites. The Mapping of *E. coli* 5'-end monophosphorylated RNA fragments were taken from the Quax *et al.* paper (Quax *et al.*, 2013), which was downloaded from: http://www.weizmann.ac.il/molgen/Sorek/Navon_data.fasta.gz. and the *P. aeruginosa* mapping were taken from Wurtzel *et al.* paper (Wurtzel *et al.*, 2012).

The 5'-end mapping for both datasets were filtered, keeping only the most reliable cleavage sites thus filtering out potential transcription-start-sites (TSS) and less reliable reads. For details on the filtering see the Material and Methods in the paper attached to the results chapter, 5.3.

4.3.2 mRNA folding energy and unpairing score calculations

To calculate the mRNA secondary structure around the cleavage sites we used Vienna RNAfold package (Lorenz *et al.*, 2011). For more details see the Material and Methods in the paper attached to the results chapter, 5.3.

4.3.3 Calculating the genes’ ribosome density

For this paper the ribosome density were constructed from ribosome occupancy data deposited in GSE35641 (Li *et al.*, 2012). The ribosome occupancy for each gene was normalized to the sum of the ribosome occupancy over all the mRNA including its UTRs (50 nucleotides upstream and downstream). For more details see the Material and Methods in the paper attached to the results chapter 5.3.
5 Results

5.1 The role of codon selection in regulation of translation efficiency deduced from synthetic libraries (Navon et al. Genome Biology 2011)

Although translated into the same amino acid, synonymous codons are not perceived the same by the translation system. The translation efficiency of a codon is affected by the availability of the tRNAs translating it and their affinity to translate it. Thus, some codons will be perceived as more optimal than others by the translation system. The aim of this project was to study how regions of less than optimal codons affect the gene expression level and the host fitness.
The role of codon selection in regulation of translation efficiency deduced from synthetic libraries

Sivan Navon, Yitzhak Pilpel*

Abstract
Background: Translation efficiency is affected by a diversity of parameters, including secondary structure of the transcript and its codon usage. Here we examine the effects of codon usage on translation efficiency by re-analysis of previously constructed synthetic expression libraries in *Escherichia coli*.

Results: We define the region in a gene that takes the longest time to translate as the bottleneck. We found that localization of the bottleneck at the beginning of a transcript promoted a high level of expression, especially if the computed dwell time of the ribosome within this region was sufficiently long. The location and translation time of the bottleneck were not correlated with the cost of expression, approximated by the fitness of the host cell, yet utilization of specific codons was. Particularly, enhanced usage of the codons UCA and CAU was correlated with increased cost of production, potentially due to sequestration of their corresponding rare tRNAs.

Conclusions: The distribution of codons along the genes appears to affect translation efficiency, consistent with analysis of natural genes. This study demonstrates how synthetic biology complements bioinformatics by providing a set-up for well controlled experiments in biology.

Background
Understanding the mechanisms that control the efficiency of protein translation is a major challenge for proteomics, computational biology and biotechnology. Efficient translation of proteins, either in their natural biological context or in heterologous expression systems, amounts to maximizing production, while minimizing the costs of the process. Abundant genome sequence data now make it possible to decipher sequence design elements that govern the efficiency of translation. The codon adaptation index (CAI) [1] was the first measure to be introduced for gauging translation efficiency directly from nucleotide sequences of genes. This measure quantifies the extent to which the codon bias of a gene resembles that of highly expressed genes. The tRNA adaptation index (tAI) assesses the extent to which the codons of a gene are biased towards the more abundant tRNAs in the organism [2]. Despite several simplifying assumptions, both tAI and CAI are good measurements for predicting protein abundance from sequence [3,4]. Perhaps the most critical simplification of the two models is that they represent the translation efficiency of an entire gene by a single number – the average translation efficiency value over all its codons. As such, both CAI and tAI ignore the order in which codons of high and low translation efficiency appear in the sequence. Thus, two genes may share the same value of CAI or tAI and yet the order of high and low efficiency codons differs between them.

By analyzing dozens of genomes, we have recently shown that the order of high and low efficiency codons in biological sequences is under selection [5,6]. Specifically, examining such genomes revealed a clustering of low efficiency codons at the beginning of ORFs, mainly in the first approximately 50 codons. We termed this design the ‘translation ramp’, or ‘ramp’ for short, which might constitute a strategic early bottleneck in the flow of the ribosomes. Our model suggests that such ramps attenuate the ribosomes at the beginning of genes, thus allowing a jam-free flow of ribosomes beyond the ramp. We have shown that this design is predominantly

* Correspondence: pilpel@weizmann.ac.il
Department of Molecular Genetics, Weizmann Institute of Science, PO Box 26, Rehovot, 76100, Israel

© 2011 Navon and Pilpel; licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/2.0, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
obeyed by highly expressed genes [5,7], suggesting that it might support efficient production. Investigating natural genes has two obvious advantages: their availability in very high numbers, and the fact that they have been subject to selection and optimization by evolution. Similarly, using the totally asymmetrical simple exclusion process (TASEP), it was theoretically shown that slow codons can affect ribosome density and production rates depending on initiation rate, termination rate, and the rate of the slow codons and their distribution [8-12].

Yet, analysis of natural sequences also poses limitations. Natural genes represent a wide variety but their variability is uncontrolled and is influenced by confounding factors at many levels. For instance, even if two genes share the same translation efficiency profile, they may differ with respect to the strength of their promoter, the untranslated regions, the secondary structure and the amino acid sequence, all factors that may affect protein levels. Synthetic biology, which now offers the ability to synthesize and express designed genes, may complement the picture obtained from bioinformatics analysis of natural genes. Although the number of genes that can be synthesized is by orders of magnitude lower than the number of natural sequences, synthetic genes enable us to modify one variable at a time while keeping others constant. In several pioneering studies of this type, the nucleotide sequence of a single gene was randomized while amino acid sequence was kept constant. In particular, these studies generated libraries of artificial variants of genes' nucleotide coding sequences, while fixing other features, such as the untranslated regions and promoters. Analysis of one such library led to an important finding - that the stability of translation regions and promoters. Analysis of one such sequences, while fixing other features, such as the untranslated regions and promoters. Analysis of one such sequence led to an important finding - that the stability of translation regions and promoters.

Looking for the effects of codon usage on translation efficiency and whether the order of the codons is important, we set out to re-analyze data from the three synthetic libraries [13,14]. The original tAI value [2] is defined for an entire gene based on all its codons as:

$$tAI_g = \prod_{k=1}^{n} w_{i_k}^{1/l_{i_k}}$$

where \(l_{i_k}\) is the length of the gene in codons and \(w_{i_k}\) is the relative adaptiveness value of the codon defined by the \(k\)th triplet in the gene.

Here we refer to the \(w_i\) value of a single codon as the codon's tAI. This measure is an approximation of the codon's translation speed, since a codon is assigned with a high tAI if the various tRNAs that translate it are at high abundance and have high affinity towards it. Besides the tAI, there are other alternative approximations for the codon's translation speed [8,15,16] (see discussion in Additional file 1). Note that all current models have approximation as their basis, necessarily introducing inaccuracies in analyses that are based on them.

To investigate the effect of regions with less than optimal codons, for each gene we defined the 'bottleneck' as a region of a fixed number of codons, \(n\), where the (harmonic) mean of the codons' tAI value is minimal (the value of \(n\) is related to the distance between two consecutive ribosomes on the mRNA (see Materials and methods). Assuming the codon's tAI value is an approximation for the translation speed, then 1/tAI can be regarded as the codon's translation time and the bottleneck is the region with the longest average translation time.

The bottleneck of each gene is characterized by two parameters: the location of the bottleneck - that is, number of codons from the ATG in which it occurs - and the 'strength' of the bottleneck - the average time to translate all the codons within it. To allow comparisons between the different genes and libraries below, we refer to the relative, rather than absolute, form of these variables - the relative location of the bottleneck is its
location divided by the length of the gene, and the relative strength is the strength divided by the average strength (that is, the time it takes to translate the bottleneck regions divided by the total time of translation of the mRNA, or $1/\tau_{AI}$ of the entire gene).

We first analyzed 154 synthetic GFP genes in a library constructed by Kudla et al. [13]. All the synthetic GFP variants had the same amino acid sequence but different codon sequences. For these genes we calculated the bottleneck parameters using a window of length $n = 21$ codons. Note that there is uncertainty regarding the exact value of this parameter (see Materials and methods); however, experimentation with other window sizes in the range $14 < n < 30$ did not affect results qualitatively (not shown). Figure 1a shows the relative location of the bottleneck of all GFP genes versus the protein abundance of each translated gene (see Materials and methods). The relative location is anti-correlated to the protein abundance (Pearson correlation $-0.43$, $P$-value $3.4 \times 10^{-8}$; Spearman correlation $-0.46$, $P$-value $2.8 \times 10^{-9}$), indicating that genes that have the bottleneck closer to the ATG (designated here as the ‘proximal bottleneck’) tend to have higher protein abundance levels compared to genes whose bottleneck are located towards the 3’ end of the gene (designated the ‘distal bottleneck’).

As for the relative strength of the bottleneck, when examining the entire library of 154 genes we found a modest yet significant correlation with the protein abundance (Pearson correlation 0.38, $P$-value $1.9 \times 10^{-6}$; Spearman correlation 0.31, $P$-value $1.2 \times 10^{-4}$); that is, genes with long dwell times of the ribosome in the bottleneck regions tended to have higher expression levels. However, as seen in Figure 1b, this correlation is mainly contributed by genes that have a proximal bottleneck. Focusing on 86 of the genes with a proximal bottleneck (located between relative positions 0.16 to 0.28) a significant positive correlation emerged between the relative strength and the protein abundance (Pearson correlation 0.47, $P$-value $3.9 \times 10^{-5}$; Spearman correlation 0.44, $P$-value $2.1 \times 10^{-5}$). From Figure 1a it is seen that there are relatively few genes with a distal bottleneck that also have a similar relative strength; therefore, the influence of the relative strength on distal genes cannot be deduced.

Summarizing the analysis of the GFP library, the distribution of the codons along the transcript appears to affect the final GFP levels in the cell. A region of less efficient codons at the beginning of a transcript - for example, a proximal bottleneck - seems to enable higher protein levels. For genes with a proximal bottleneck it is also beneficial to have a relatively long dwell time of the ribosome, that is, a strong enough bottleneck. From this library we were not able to learn about the significance of the bottleneck strength in the case of genes with distal bottlenecks; however, other libraries with different distributions of bottlenecks can shed light on the question.

In another recent paper, by Welch et al. [14], two different proteins were synthesized: the DNA polymerase of Bacillus phage and an antibody fragment (scFv). For each protein there are approximately 40 different
sequences in which the amino acid was kept the same while changing the codon sequence. For both proteins, the location of the bottleneck is quite far from the ATG in most synthetic variants (relative distance of approximately 0.5 and higher; Figure S1 in Additional file 2), excluding the possibility of examining the effect of the proximal bottleneck on the expression of these two proteins. Nonetheless, we could still compute the correlation between the bottleneck’s parameters and protein abundance. Although less significant than in the case of the GFP library, both libraries showed an anti-correlation between protein abundance levels and the relative location of the bottleneck (Spearman correlation $-0.34$ ($P$-value 0.06) and $-0.40$ ($P$-value 0.03); Pearson correlation $-0.34$ ($P$-value 0.06) and $-0.16$ ($P$-value 0.40) for the scFv and the polymerase, respectively). Similar to the GFP library, such negative correlation indicates that proximal bottlenecks are often associated with higher expression levels. As was done for the GFP library, we looked at the correlation between protein abundance and the bottleneck relative strength (Figure 2) for specific locations, chosen based on Figure S1 in Additional file 2. Interestingly, while in the case of the GFP library a proximal bottleneck became more effective with increased relative strength, in the cases of scFv and the polymerase, which featured a distal bottleneck, the strength actually showed the opposite correlation; that is, genes with long dwell times in the bottleneck regions showed lower protein abundance (Spearman correlation $-0.43$ ($P$-value 0.02) and $-0.67$ ($P$-value $7.1 \times 10^{-5}$) for all genes of scFv and the polymerase, respectively). It is our understanding that a proximal bottleneck can have beneficial effects on protein production [5]. The bottleneck can delay the translating ribosome, causing a ribosome backlog (when in polysome), and can also reduce the density of the ribosome downstream. A proximal bottleneck minimizes the number of jammed ribosomes, thus reducing ribosome sequestering and collisions, two potential causes for a decrease in protein production. Assuming the bottleneck reduces the density of ribosomes downstream, a slower bottleneck (that is, a bottleneck with increased relative strength) will reduce even more downstream ribosome collisions, improving protein production, as seen with the GFP library. On the other hand, a distal bottleneck at the end of the ORF causes a long backlog, with no beneficial effects on expression levels. Since a bottleneck at the end of the ORF seems to have mainly negative effects on the protein translation rate, reducing its relative strength is beneficial, as seen in the case of the scFv and the polymerase.

To further verify our assumption that the bottleneck may have beneficial effects on protein abundance when they are located at the beginning of a gene, we looked at the distribution of locations of the bottleneck in natural *Escherichia coli* genes [Refseq: NC_012947] (Figure 3; Figure S2 in Additional file 2). Indeed, for most genes with a bottleneck of high relative strength (higher than 1.3), the bottleneck region is located in the first quadrant of the transcript (relative location smaller than 0.25). For 41% of genes with a bottleneck of high relative strength, the bottleneck is located in the first quadrant (hyper-geometric significant enrichment $P$-value...
abundance in the GFP library is due to the confounding we observe between bottleneck location and protein observations raised the possibility that the correlation mRNA, thus enabling a faster translation rate. These rate. It was suggested that a weak secondary structure mRNA near the initiation site influences translation conventions we found still hold even when examining gene sets with similar mRNA folding energy. We calculated the relative location correlation -0.27, (Pearson correlation -0.24, P-value 0.004; Spearman correlation -0.27, P-value 9.5 x 10^-4) and the relative strength at locations 0.16 to 0.28 (Figure 1) correlation of the codon in the sequence) and the fitness of the cell specific codon in the GFP sequence (number of copies that was expressing that GFP variant (Figure 5). Interestingly, the extent of usage of some codons is negatively correlated with fitness, is positively correlated for others, and for the rest is not correlated with fitness. The cases of negative correlation may indicate a burden on fitness due to using particular codons. In contrast, since fitness can only decrease due to GFP expression, cases of positive correlation between codon usage and fitness of the host cell, the fitness decline due to expression reflects only the pure cost of production.

Kudla et al. [13] showed that the measured optical density (OD), assumed to be proportional to the fitness of the host cell, is highly correlated with the CAI. Further analysis showed that the tAI is also correlated with OD (Pearson correlation 0.51, P-value 2.4 x 10^-11). These two similar measures describe the entire transcript and not a particular region within it. In contrast, we found that the bottleneck parameters that significantly correlate with protein abundance are not correlated with cell fitness. Thus, the factors that correlate with fitness and those correlating with protein abundance appear distinct in this library (Figure 4). It seems that while specific regions of the transcript affect protein abundance, the fitness is affected by the codon usage of the entire transcript.

Trying to understand the source for the correlation between the fitness and tAI or CAI, we examined the effect of individual codons on cell fitness. We analyzed the correlation between the usage frequency of each specific codon in the GFP sequence (number of copies of the codon in the sequence) and the fitness of the cell that was expressing that GFP variant (Figure 5). Interestingly, the extent of usage of some codons is negatively correlated with fitness, is positively correlated for others, and for the rest is not correlated with fitness. The cases of negative correlation may indicate a burden on fitness due to using particular codons. In contrast, since fitness can only decrease due to GFP expression, cases of positive correlation between codon usage in a gene and its host fitness likely reflect an artificial negative correlation of synonym codons; that is, the preference for not using its alternative codons rather than a preference for expressing the codon itself.
Thus, focusing on the codons that correlate negatively with fitness, we detected three codons whose usage correlates most significantly: CAU (Pearson correlation -0.69, P-value < 10^{-324}); AAU (Pearson correlation -0.68, P-value < 10^{-324}); and UCA (Pearson correlation -0.67, P-value < 10^{-324}) (Figure 5; Table S2 in Additional file 1). Further examination reveals inter-dependencies between the usage of some of these codons; in particular, the frequencies of CAU and AAU are highly correlated (r = 0.92, P-value 10^{-64}) among themselves (the reasons for internal correlation may have to do with GFP construction methods; see Kudla et al. [13]). Using partial correlation analysis between the usage of each codon, we identified UCA and CAU as the main codons contributing to the decrease in the fitness (see Materials and methods).

The number of occurrences of the UCA codon, encoding serine, in a single gene varies between zero to three appearances. This codon is the rarest out of the six serine codons in the E. coli genome [Refseq: NC_012947], though it is not extremely rare (12.2% of all serine codons, and 0.7% of all 61 codons in the ORFs of the genome; Table S2 in Additional file 1). However, in the transcriptome (that is, the genome, weighted by the mRNA expression level from each gene; see Materials and methods) UCA is one of the rarest codons (8.7% of all serine codons and 0.45% of all 61 codons). The UCA codon is exclusively translated by the tRNA_{UGA} [17]. The genome of E. coli has only one copy of this tRNA gene and, reassuringly, it was shown that a shortage of this tRNA decreases cell fitness [18]. The negative correlation between the copy number of the UCA codon and the fitness can thus imply that increased usage of the UCA codon causes a shortage of the corresponding tRNA, causing a decrease in fitness. Regarding codons CAU and AAU, they are negatively correlated with fitness (and with one another) yet we found no apparent reason for this.

Shortage of tRNAs explains some of the correlations between the usage of certain codons and fitness; however, it is not clear through which mechanism a shortage of tRNAs affects the fitness. The extensive usage of codons that correspond to rare tRNAs can affect the fitness in at least one of two alternative ways: by ‘consuming’ the tRNAs and sequestering them from participating in the translation of other transcripts; or through the unavailability of ribosomes that are delayed for longer times while searching for rare tRNAs. A simple means to distinguish between these two alternative options is to examine whether not only the number but also the location of such rare codons affects fitness. In particular, we expect that if the fitness-reducing effect of the rare codons is the jamming of ribosomes, then their utilization will be particularly harmful when located distally, closer to the 3’ end of the transcript. In contrast, if the fitness-reducing effect is predominantly due to the consumption of rare tRNAs, then it is not expected to show such location dependence. In reality, we observed no correlation with the location (Figure S3 in Additional file 2), suggesting that it is the consumption of the rare tRNAs, in this case, that compromises fitness.

![Figure 4 Correlation between the GFP experimental measurement and transcript calculated parameters](http://genomebiology.com/2011/12/2/R12)
Conclusions
As shown, a proximal and strong bottleneck is correlated with an increase in protein abundance. A proximal bottleneck can reduce the number of jammed ribosomes on a transcript. Therefore, it can reduce both the number of occupied ribosomes and the number of delayed ribosomes. Delaying ribosomes on the mRNA might increase their abortion rate, thus causing early termination of the translation [19], reducing protein levels. For ribosomes to jam, a fast initiation rate is required. This is usually the case in highly expressed genes, in cases of heterologous gene expression, and in synthetic libraries such as discussed here where high protein levels are desired. Due to amino acid sequence constraints for some genes, a naïve approach, using only optimal codons, might result in an unintentional distal bottleneck.

While the bottleneck parameters are correlated with protein abundance, they are not correlated with fitness. This suggests that while the occupation of more ribosomes sequesters them from the cell’s pool, for most genes in the GFP library it does not cause a shortage of ribosomes, enabling the cell to continue translating other transcripts. The decrease in fitness is correlated with the increased usage of codons UCA and CAU, suggesting a shortage of the complementary tRNAs.

Our results thus show that, along with mRNA stability, codon choice does affect translation efficiency, and that naïve averaged measures such as CAI and tAI do not capture this regulatory capacity. The results also
show that while codon choices do affect both translation efficiency and cell fitness, different aspects of codon selection affect differently the production capacity and costs. One direct conclusion from our results relates to the popular usage of ‘His-tags’, chains of histidine residues at carboxyl termini of genes in heterologous expression systems [20]. When using carboxyl-terminal His-tags in bacterial expression systems it would be advantageous to encode histidine with CAC rather than with CAU for two reasons: first, because CAU appears to correlate negatively with fitness; and second, in order to avoid a bottleneck towards the end of the gene.

When trying to understand the cell system, one realizes its processes are regulated on many different levels. As shown in this paper, synthetic gene libraries enabled us to control for a significant portion of gene variability and focus on the effects of regions with less than optimal codons (the bottleneck). Identification of bottleneck effects in synthetic genes thus completes Tuller et al.’s [5] bioinformatics work that identified clustering of low efficient codons at the beginning of ORFs of natural genes. The results further demonstrate how correlative conclusions made from observations of natural gene sequences can be complemented by synthetic genes, allowing decoding of the sequence features governing the efficiency of translation and it costs.

It is our belief that through carefully designed synthetic libraries many other regulation processes can be understood, thus completing the first step towards understanding the regulation process as a whole.

**Materials and methods**

**Defining the bottleneck**

The bottleneck is a region on a gene where the harmonic mean of its codons’ tAI values is minimal. For all codons except CGA, the tAI values were calculated using dos Reis et al.’s s-values [2]; for codon CGA the value 0.1333 was used. This codon is translated with tRNA_{ACG}; however, the s-value for this interaction is very high, resulting in a very low tAI value. This tAI value is smaller by at least an order of magnitude than the smallest tAI value, causing all other codons to have a relatively high tAI, disabling this analysis. Since CGA is actually translated by tRNA_{ACG}, we decided to change the s-value of this interaction to a more reasonable value, resulting in the above mentioned tAI value. Given the tRNA repertoire of *E. coli*, this change affects only the tAI value of codon CGA.

A codon tAI value is assumed to be proportional to the speed of the codon’s translation [5]; higher tAI values correspond to high tRNA abundance and affinity, thus faster translation. A harmonic mean of speeds is simply an arithmetic mean of the corresponding times. Hence, looking for the region with the minimum harmonic mean of speed is equivalent to looking for the region that takes the longest time to translate.

For each region the harmonic mean of speed is:

\[
\frac{n}{\sum_{i \in \text{Region}} \frac{1}{t_{AIc}}}
\]

where \( n \) is the region size, and \( c \) is the set of all the codons in the region (\( n \) codons).

To find the bottleneck, a sliding window of length \( n \) over the gene was used. The harmonic mean was calculated for each window and the window with the minimum value was identified. It should be noted that since we are averaging the translation time in a window, an incorrect window size might in some cases result in incorrect identification of the bottleneck. For example, if our estimated window size is too big, it might mask a cluster of a few slowly translated codons, of a more relevant size, that are surrounded by relatively rapidly translated codons. In most cases, however, the slow region is significant enough and its identification is not too sensitive to window size. Indeed, as mentioned in the Result and discussion section, our results did not change qualitatively for window sizes in the range 14 < \( n < 30 \).

**The bottleneck window size (n)**

Under a maximal density scenario (fast initiation rate), the distance between two consecutive ribosomes will be minimal. In this case, when two ribosomes are translating the same mRNA simultaneously, the minimal possible distance between the two translated codons (one by each of the ribosomes) is one ribosome size (\( H \) codons) (Figure S4 in Additional file 2). At any given moment during the translation process, two adjacent ribosomes would have translated exactly the same codons apart from the last \( H \) codons - the first of the two ribosomes has already translated them, and the second is just about to start them. If the time it took the first ribosome to finish translating the \( n \)-th codon, \( T(n,1) \), is longer than the time it takes the second ribosome to translate the \( n-H \)th codon, \( T(n-H,2) \), the second ribosome will ‘bump’ into the first one. That is, if \( T(n,1) > T(n-H,2) \), a traffic jam will be created. \( T(n,1) \) can be found by summing the time it takes the ribosome to assemble on the ATG (\( B \)) with the time it takes to translate the \( n \) codons:

\[
T(n,1) = B + \sum_{i=1}^{n} t(i)
\]

where \( t(i) \) is the time it takes to translate the \( i \)-th codon. The second ribosome gains access to the ATG only when enough codons (minimum \( H \)) are cleared after being translated by the first ribosome. As a result a
traffic jam will be created if \( Tw(k,H) > Tw(1,H) + B \), where \( Tw(k,H) \) is the time to translate \( H \) consecutive codons starting from codon \( k \):

\[
Tw(k,H) = \sum_{i=k}^{k+H-1} t(i)
\]

Therefore, the region of \( H \) codons with maximum translation time \( \left( \arg \max_{k=1:mRNA.length-H} \left( Tw(k,H) \right) \right) \) determines whether and where a traffic jam will be created (for a detailed calculation, see page 2 of Additional file 1). Choosing \( n \) in our bottleneck equation to be equal to \( H \), it is easy to see that our bottleneck is related to this maximum.

As can be seen from this analysis, the minimal distance between two ribosomes should determine our window size. The footprint of the ribosome, which is the actual protection of the ribosome from RNA degradation, was determined quite accurately to be ten codons [21]. Due to the structure of the ribosomes, we assume that there should be some space between two consecutive 30S subunits. As a result, although only ten codons are protected, the minimal distance between the two ribosomes should be larger. Therefore, we chose to adopt the average ribosome-to-ribosome distance measured by Brandt et al. [22]. They measured the mean distance between the center of mass of two ribosomes on actual bacterial polysomes to be 21.6 nm [22], which is about 21 codons (0.34 nm per base). In this paper, \( n \) was set to be equal to \( H \); that is \( n \) is set to 21 codons.

### The bottleneck parameters

A bottleneck is characterized by two parameters: its ‘location’ and its ‘strength’.

The ‘location’ of the bottleneck is defined as the location in the gene of the bottleneck’s first codon (\( k \) codons from the ATG). The relative location of the bottleneck is defined as the location of the bottleneck divided by the number of possible windows; for example, \( \frac{k}{l-n+1} \), where \( k \) is the location of the bottleneck, \( l \) is the length of the gene, and \( n \) is the window size.

The ‘strength’ of the bottleneck is defined as the arithmetic average of \( 1/tAI \) values for the codons in the region, for example, \( \frac{1}{n} \sum_{e \in \text{Region}} \frac{1}{tAI_e} \) (the inverse of the harmonic mean). The relative strength of the bottleneck is defined as the strength of the bottleneck divided by the average \( 1/tAI \) for the entire gene, for example, \( \frac{1}{l} \sum_{e \in \text{bottleneck}} \frac{1}{tAI_e} \); where \( l \) is the number of codons in the gene (excluding the stop codon).

#### Per-cell protein abundance

To get an estimate for protein expression per cell from the GFP library data [13], we normalized the measured protein abundance (measured by OD), which serves here as a proxy for the population size, the OD. The protein abundance levels for the data from Welch et al. [14] were measured while keeping the OD constant. Therefore, we can use this protein abundance as an already normalized protein level per cell.

#### Highly and lowly expressed genes of *E. coli*

The *E. coli* mRNA levels were taken from Lu et al. [23]. The highly expressed genes are the top 500 genes, and the lowly expressed genes are the bottom 500 genes (genes with no mRNA recorded were ignored). However, for both groups only genes that are longer than 100 codons were used.

### Finding the main anti-correlated codons

We used partial correlation to find the codons that contribute the most to the decrease in cell fitness. The highest contributors were filtered according to the following steps. First, find codons that have a negative correlation to the OD (29 codons). We were looking for codons that caused a decrease in the fitness; hence, only anti-correlated codons. Second, for all codons left, we calculated the partial correlation matrix \( M(ij) = \text{Partial correlation (codon } i, \text{ OD } | \text{ codon } j) \). Third, find the minimum absolute value of the partial correlation for each codon and rank the codons in a descending order accordingly. This gives us the codons with a correlation that cannot be explained by correlation to other codons (see Table S4 in Additional file 1 for a list of all codons with \( P \)-value < 0.1).

The codon at the top of the list is UCA, which is anti-correlated to the OD and its correlation cannot be explained by other codons. The second contributing codon is CAU, which has the highest partial correlation (-0.36, \( P \)-value \( 8.5 \times 10^{-6} \)) when controlling for the UCA codon. This codon is also the second codon in the ranked list. All other codons have a partial correlation < 0.2 with a \( P \)-value > 0.04 when controlling with one of the two codons (either UCA or CAU).

#### Calculating codon usage in the genome

The genome for *E. coli* strain B21 (which was used by Kudla et al. [13]) was downloaded from the NCBI ([Refseq:NC_012947], 11 January 2010). For each codon we counted its appearance in all the ORFs and normalized by the total number of codons.

#### Calculating codon usage in the transcriptome

mRNA levels were taken from Lu et al. [23]. If a gene did not have a measurement, it was assumed to have a
zero mRNA level. The measurements were done with E. coli strain K12 MG1655; thus, the sequence used for the calculation was different from that used for genome codon usage. The sequence was downloaded from NCBI ([Refseq: NC_000913], 1 April 2010). The contribution of each gene was calculated by multiplying the mRNA level measurements for the gene by the codon usage of the same gene. The contributions of all genes were summed for each codon and then divided by the total sum of all codons.

Additional material

Additional file 1: Supplementary methods. This file includes a discussion regarding codon translation speed, additional tables not included in the main text, and figure legends for the supplementary figures in Additional file 2.

Additional file 2: Supplementary figures. Additional figures not included in the main text.

Abbreviations


Acknowledgements

We thank the ‘Ideas’ program of the European Research Council (ERC), and the Ben May Charitable Trust for grant support.

Authors’ contributions

SN carried out all analyses. SN and YP conceived the work, analyzed the data and wrote the paper.

Received: 23 August 2010 Revised: 18 November 2010 Accepted: 1 February 2011 Published: 1 February 2011

References

18. Yamada Y, Matsusugi J, Ishihara H: tRNA1Ser(G34) with the anticodon GGA can recognize not only UCC and UCU codons but also UCA and UCG codons. Biochim Biophys Acta 2003, 1626:75-82.

Cite this article as: Navon and Pilpel: The role of codon selection in regulation of translation efficiency deduced from synthetic libraries. Genome Biology 2011 12:R12.
Supplement Text

**Codons translation speed**
In this paper we used the tRNA adaptation index (tAI) developed by dos Reis [2] as an approximation for the codon translation speed. The tAI index was developed mainly based on optimizing the translation efficiency of highly express genes. Two recent papers [15, 16] used evolutionary selection for translation efficiency as a means to learn the efficiency and coefficients of codon-anticodon interactions. They found that some non-standard combinations seem to be selected for and hence deduced to be translated at high rates. One of the major difference between their conclusion to the tAI is the value of the U:U interaction (codon with wobble-U translated by a tRNA with a wobble- U) for amino acid in a 4-box. In the tAI calculations this interaction was assumed to be forbidden (sU:U=1) while based on selection it was suggested that this interaction is quite potent, which should result in a low sU:U in the tAI model.

To investigate the U:U interaction effects we manually added the sU:U interaction to the tAI calculations at 4-box cases and scanned the effects over the full possible range of values of sU:U, form 0 to 1. First we looked how adding sU:U changed the bottleneck parameters. Figure S5 shows the correlation between our original bottleneck parameters and the newly calculated one. As expected the correlation decreases as sU:U decreases. The correlation decreases from 1 for sU:U=1 (which is as in the original dos Reis tAI values) to 0.61 for sU:U=0, for the bottleneck relative location and 0.51 for the bottleneck relative strength.

Next we looked how the modified tAI predicts expression levels of the *E. coli* mRNA levels [26] for 3 different groups: all genes, highly expressed (top 250) and lowly
express (bottom 250). As can be seen in Figure S6 for all genes and the highly expressed ones the correlations increase with an introduction of the U:U interaction (sU:U<1). However it seems that the main contribution to the increase occurs merely by introducing the interaction, i.e. reducing sU:U from 1 to 0.7 results in half of the increase in the correlation. For this graph we find it hard to deduce what should be the optimal sU:U value, but the result indeed suggests a desired correction for the tAI. Finally we look at the two main correlations we find in the GFP library: the correlation with between the bottleneck relative location to the protein abundance and the correlation between the bottleneck relative strength for genes with the same relative location (between 0.16-0.28) and protein abundance. For both correlations we see (figure S7) that the introduction of sU:U actually reduces the correlations.

The bottleneck window size (n) - detailed calculation

The bottleneck is the region that will have the most slowing down effect on the ribosome. This region will only have a bottleneck effect if it slows the ribosomes enough to affect consecutive ribosome. As explained in the following the size of this region has to be about the minimum ribosome-to-ribosome distance (denoted below as H).

We make first the following definitions:

1. Traffic jam will be caused if the time it takes the first ribosome to finish translation of the nth codon is longer than the time it takes the second ribosome to finish translation of codon n-H. In this case the second ribosome will not be able to proceed to the (n-H+1)-th codon since it will collide with the first ribosome and hence will be delayed.
2. A ribosome can start assembling on the ATG when the first H codons (the minimum distance required between 2 ribosomes) are cleared from the preceding ribosome. Let B be the time it takes the ribosome to bind and assemble on the ATG.

3. If \( t(i) \) is the time it takes to translate codon \( i \), and \( T(n,j) \) is the time it takes the \( j \)th ribosome to finish translation of the \( n \)th codon, where \( n \) is the codon at the leading edge of the ribosome.
   then:
   
   a. The time it takes the first ribosome to reach the \( n \)th codon:
   
   \[
   T(n,1) = B + \sum_{i=1}^{n} t(i)
   \]
   
   b. The time it takes the second ribosome to reach the \((n-H)\)th codon:
   
   \[
   T(n-H,2) = (\text{time until the assembly site is cleared}) + B + \sum_{i=1}^{n-H} t(i)
   \]
   
   \[
   \Rightarrow T(n-H,2) = B + \sum_{i=1}^{H} t(i) + B + \sum_{i=1}^{n-H} t(i)
   \]

4. For a traffic jam to be created, the time for the first ribosome to finish translation of the \( n \)th codon should be longer than the time it takes the second ribosome to finish translation of the \( n-H \) codon:

\[
T(n,1) > T(n-H,2)
\]

\[
\Rightarrow B + \sum_{i=1}^{n} t(i) > B + \sum_{i=1}^{H} t(i) + B + \sum_{i=1}^{n-H} t(i)
\]

\[
\Rightarrow \sum_{i=H+1}^{n} t(i) > B + \sum_{i=1}^{H} t(i)
\]
5. We define $Tw(k, H)$ is the time to translate $H$ consecutive codons starting from codon $k$ \( Tw(k, H) = \sum_{i=k}^{k+H-1} t(i) \). Therefore, a traffic jam will be created if:

\[
Tw(k, H) > B + Tw(1, H)
\]

\[
Tw(k, H) - Tw(1, H) > B
\]

6. For any traffic jam to be created along the gene:

\[
\arg \max_{k=1}^{mRNA length-H} (Tw(k, H)) - Tw(1, H) > B
\]

Inspecting the slowest region (of $H$ codons) for the specific gene (bottleneck strength), if the time to translate that region minus the time it takes to translate the first $H$ codons is longer than the time to finish “assembling” the ribosome ($B$) than traffic jam will be created.

As can be seen from the equation the region size which determine whether two consecutive ribosomes will collide has the size of the minimum distance between 2 ribosomes.
Supplement Figures

Figure S1: protein abundance vs. bottleneck relative location of data from Welch et al.’s libraries.

In the left figure (A) plotted all the scFv genes and in the right figure (B) plotted all the Polymerase genes. In both figures the x-axis is the gene’s relative location, the Y-axis is the per-cell protein abundance and the color is the gene’s relative strength.

Figure S2: The distribution (2D histogram) of the bottleneck of the E. coli genes.

X-axis is the relative location, Y-axis the relative strength and the color is the % of genes having a bottleneck matching the parameters. A all E. coli genes longer than 100 (well above region size and still maintain 90% of the E. coli genes) codons were plotted. B only the highest expressed genes were plotted. The 500 genes which have the highest transcript levels were chosen and from these the genes with 100 codons or longer were taken, making a total of 442 genes. C only the lowest expressed genes were plotted. The 500 genes which have the lowest transcript levels were chosen and from these the genes with 100 or longer were taken, making a total of 473 genes.

Figure S3: location of first copy of a codon vs. the fitness.

Each point represents a gene in the GFP library. A, plotted is the location of the first CAU codon for each GFP variant vs. the variants’ OD. In figure B, the location of the first UCA codon in the GFP variants is plotted vs. the variants’ OD.

Figure S4: distances between two consecutive ribosomes

The figure illustrates two consecutive ribosomes, on the same mRNA, with the second one (left) currently being assembled on the ATG. The size of a ribosome in the figure is $H$ codons. $H_L$ is the distance from the ribosome A-site to the left end of the
ribose, $H_E$ is the distance from the A-site to the right end of the ribosome. The illustration shows that the minimum distance between two ribosomes' A-sites is $H$ which is also “one ribosome size”. It can also be seen that in order for the second ribosome to start assembling on the ATG the first ribosome should have cleared the assembly area, e.g. translate $H$ codons.

**Figure S5: Correlation between the bottleneck parameters for different sU:U values**

The correlation between the bottleneck relative location (A), and bottleneck relative strength (B), calculated once with the original dos-Reis tAI values and with alternative values of sU:U.

**Figure S6: The gene expression prediction quality of the modified tAI values**

For each sU:U value we calculated the correlation between the gene’s tAI value to the expression level. In the figure the correlation is plotted for three different groups: all the *E. coli* genes (blue), only the highly express ones, top 250 (green) and the lowly express ones, bottom 250 (red).

**Figure S7: correlation between bottleneck parameters and the protein abundance for different sU:U values.**

**A** Correlation between the bottleneck relative location and the protein abundance for all the GFP variants, for different sU:U values. The blue line is the correlation between the relative location and the abundance while the green line is the p-value of each correlation. **B** Correlation between the bottleneck relative strength and the protein abundance for the 86 GFP variants for which the bottleneck is located in between 0.16 to 0.28 (as done in the main text) for different sU:U values. The blue
line is the correlation between the relative location and the abundance while the green
line is the p-value of each correlation.
Supplement Tables

Table S1. The correlations between the bottleneck relative strength to the protein abundance for the scFv and Polymerase libraries.

The relative location regions were chosen to incorporate many genes with the same relative location; the regions were chosen based on the plots in figure S1.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Relative location</th>
<th>Number of genes</th>
<th>Correlation with the bottleneck relative strength</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>scFv</td>
<td>0.9-1</td>
<td>25</td>
<td>-0.23; -0.32</td>
<td>0.27; 0.11</td>
</tr>
<tr>
<td>scFv</td>
<td>0-1 (All)</td>
<td>42</td>
<td>-0.43; -0.41</td>
<td>0.01; 0.02</td>
</tr>
<tr>
<td>Polymerase</td>
<td>0.48-0.52</td>
<td>13</td>
<td>-0.60; -0.67</td>
<td>0.03; 0.015</td>
</tr>
<tr>
<td>Polymerase</td>
<td>0.76-0.82</td>
<td>13</td>
<td>-0.38; -0.43</td>
<td>0.2; 0.14</td>
</tr>
<tr>
<td>Polymerase</td>
<td>0-1 (All)</td>
<td>39</td>
<td>-0.55; -0.67</td>
<td>0.0018; 7.1e-5</td>
</tr>
</tbody>
</table>

Table S2 – Codon parameters

For each codon the tables contains it amino acid, number of copies of complementary tRNA in the genome, its tAI value, the Pearson correlation with the OD measurements, the codon usage in the genome and the in the transcriptome. Except for the transcriptome all values are based on E. coli strain B, the transcriptome was calculated for E. coli K12 (see methods). When NaN (Not a Number) is listed it means that a correlation cannot be calculated due to a constant value of codons for all GFP variants.

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Codon</th>
<th># tRNA Copies</th>
<th>tAI</th>
<th>correlation</th>
<th>p-value</th>
<th>Codon usage in genome %</th>
<th>Codon usage in transcriptome %</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>AAU</td>
<td>0</td>
<td>0.39</td>
<td>-0.68</td>
<td>&lt;E-324</td>
<td>1.73</td>
<td>1.21</td>
</tr>
<tr>
<td>N</td>
<td>AAC</td>
<td>4</td>
<td>0.67</td>
<td>0.68</td>
<td>&lt;E-324</td>
<td>2.16</td>
<td>2.69</td>
</tr>
<tr>
<td>K</td>
<td>AAA</td>
<td>6</td>
<td>1.00</td>
<td>0.11</td>
<td>1.96E-01</td>
<td>3.37</td>
<td>4.25</td>
</tr>
<tr>
<td>K</td>
<td>AAG</td>
<td>0</td>
<td>0.32</td>
<td>-0.11</td>
<td>1.96E-01</td>
<td>1.02</td>
<td>1.29</td>
</tr>
<tr>
<td>T</td>
<td>ACU</td>
<td>0</td>
<td>0.20</td>
<td>0.16</td>
<td>5.71E-02</td>
<td>0.89</td>
<td>1.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>T</td>
<td>ACC</td>
<td>2</td>
<td>0.33</td>
<td>0.07</td>
<td>4.09E-01</td>
<td>2.33</td>
<td>2.60</td>
</tr>
<tr>
<td>T</td>
<td>ACA</td>
<td>1</td>
<td>0.17</td>
<td>-0.09</td>
<td>2.57E-01</td>
<td>0.68</td>
<td>0.48</td>
</tr>
<tr>
<td>T</td>
<td>ACG</td>
<td>1</td>
<td>0.22</td>
<td>-0.20</td>
<td>1.64E-02</td>
<td>1.45</td>
<td>1.02</td>
</tr>
<tr>
<td>S</td>
<td>AGU</td>
<td>0</td>
<td>0.10</td>
<td>-0.48</td>
<td>8.13E-10</td>
<td>0.86</td>
<td>0.54</td>
</tr>
<tr>
<td>S</td>
<td>AGC</td>
<td>1</td>
<td>0.17</td>
<td>0.48</td>
<td>8.13E-10</td>
<td>1.61</td>
<td>1.34</td>
</tr>
<tr>
<td>R</td>
<td>AGA</td>
<td>1</td>
<td>0.17</td>
<td>-0.16</td>
<td>5.48E-02</td>
<td>0.18</td>
<td>0.10</td>
</tr>
<tr>
<td>R</td>
<td>AGG</td>
<td>1</td>
<td>0.22</td>
<td>0.16</td>
<td>5.48E-02</td>
<td>0.11</td>
<td>0.05</td>
</tr>
<tr>
<td>I</td>
<td>AUU</td>
<td>0</td>
<td>0.30</td>
<td>-0.39</td>
<td>9.18E-07</td>
<td>3.05</td>
<td>2.57</td>
</tr>
<tr>
<td>I</td>
<td>AUC</td>
<td>3</td>
<td>0.50</td>
<td>0.42</td>
<td>1.21E-07</td>
<td>2.52</td>
<td>3.26</td>
</tr>
<tr>
<td>I</td>
<td>AUA</td>
<td>0</td>
<td>0.055</td>
<td>-0.22</td>
<td>8.02E-03</td>
<td>0.41</td>
<td>0.19</td>
</tr>
<tr>
<td>M</td>
<td>AUG</td>
<td>7 (3.5, 3.5)*</td>
<td>0.58</td>
<td>NaN</td>
<td>NaN</td>
<td>2.81</td>
<td>2.73</td>
</tr>
<tr>
<td>H</td>
<td>CAU</td>
<td>0</td>
<td>0.10</td>
<td>-0.69</td>
<td>&lt;E-324</td>
<td>1.28</td>
<td>0.95</td>
</tr>
<tr>
<td>H</td>
<td>CAC</td>
<td>1</td>
<td>0.17</td>
<td>0.69</td>
<td>&lt;E-324</td>
<td>0.98</td>
<td>1.13</td>
</tr>
<tr>
<td>Q</td>
<td>CAA</td>
<td>2</td>
<td>0.33</td>
<td>-0.43</td>
<td>5.75E-08</td>
<td>1.53</td>
<td>1.15</td>
</tr>
<tr>
<td>Q</td>
<td>CAG</td>
<td>2</td>
<td>0.44</td>
<td>0.43</td>
<td>5.75E-08</td>
<td>2.89</td>
<td>2.93</td>
</tr>
<tr>
<td>P</td>
<td>CCC</td>
<td>0</td>
<td>0.10</td>
<td>-0.42</td>
<td>8.35E-08</td>
<td>0.69</td>
<td>0.57</td>
</tr>
<tr>
<td>P</td>
<td>CCA</td>
<td>1</td>
<td>0.17</td>
<td>-0.35</td>
<td>1.08E-05</td>
<td>0.83</td>
<td>0.73</td>
</tr>
<tr>
<td>P</td>
<td>CCG</td>
<td>1</td>
<td>0.22</td>
<td>0.57</td>
<td>4.65E-14</td>
<td>2.35</td>
<td>2.54</td>
</tr>
<tr>
<td>R</td>
<td>CGU</td>
<td>4</td>
<td>0.67</td>
<td>0.03</td>
<td>7.23E-01</td>
<td>2.11</td>
<td>2.88</td>
</tr>
<tr>
<td>R</td>
<td>CGC</td>
<td>0</td>
<td>0.48</td>
<td>-0.30</td>
<td>1.92E-04</td>
<td>2.22</td>
<td>2.09</td>
</tr>
<tr>
<td>R</td>
<td>CGA</td>
<td>0</td>
<td>0.13</td>
<td>0.22</td>
<td>6.21E-03</td>
<td>0.35</td>
<td>0.17</td>
</tr>
<tr>
<td>R</td>
<td>CGG</td>
<td>1</td>
<td>0.17</td>
<td>0.06</td>
<td>4.53E-01</td>
<td>0.53</td>
<td>0.25</td>
</tr>
<tr>
<td>L</td>
<td>CUU</td>
<td>0</td>
<td>0.10</td>
<td>-0.11</td>
<td>1.95E-01</td>
<td>1.10</td>
<td>0.79</td>
</tr>
<tr>
<td>L</td>
<td>CUC</td>
<td>1</td>
<td>0.17</td>
<td>-0.08</td>
<td>3.08E-01</td>
<td>1.11</td>
<td>0.82</td>
</tr>
<tr>
<td>L</td>
<td>CUA</td>
<td>1</td>
<td>0.17</td>
<td>-0.42</td>
<td>7.56E-08</td>
<td>0.39</td>
<td>0.20</td>
</tr>
<tr>
<td>L</td>
<td>CUG</td>
<td>4</td>
<td>0.72</td>
<td>0.29</td>
<td>4.45E-04</td>
<td>5.33</td>
<td>5.80</td>
</tr>
<tr>
<td>D</td>
<td>GAU</td>
<td>0</td>
<td>0.30</td>
<td>-0.41</td>
<td>2.48E-07</td>
<td>3.20</td>
<td>2.97</td>
</tr>
<tr>
<td>D</td>
<td>GAC</td>
<td>3</td>
<td>0.50</td>
<td>0.41</td>
<td>2.48E-07</td>
<td>1.91</td>
<td>2.54</td>
</tr>
<tr>
<td>E</td>
<td>GAA</td>
<td>4</td>
<td>0.67</td>
<td>-0.09</td>
<td>2.79E-01</td>
<td>3.97</td>
<td>4.80</td>
</tr>
<tr>
<td>E</td>
<td>GAG</td>
<td>0</td>
<td>0.21</td>
<td>0.09</td>
<td>2.79E-01</td>
<td>1.77</td>
<td>1.80</td>
</tr>
<tr>
<td>A</td>
<td>GCU</td>
<td>0</td>
<td>0.20</td>
<td>-0.18</td>
<td>3.00E-02</td>
<td>1.54</td>
<td>2.31</td>
</tr>
<tr>
<td>A</td>
<td>GCC</td>
<td>2</td>
<td>0.33</td>
<td>-0.27</td>
<td>1.13E-03</td>
<td>2.57</td>
<td>2.09</td>
</tr>
<tr>
<td>A</td>
<td>GCA</td>
<td>3</td>
<td>0.50</td>
<td>0.46</td>
<td>6.25E-09</td>
<td>2.02</td>
<td>2.23</td>
</tr>
<tr>
<td>A</td>
<td>GCG</td>
<td>0</td>
<td>0.16</td>
<td>-0.42</td>
<td>9.49E-08</td>
<td>3.41</td>
<td>3.17</td>
</tr>
<tr>
<td>G</td>
<td>GGU</td>
<td>0</td>
<td>0.39</td>
<td>0.19</td>
<td>2.16E-02</td>
<td>2.48</td>
<td>3.39</td>
</tr>
<tr>
<td>G</td>
<td>GGC</td>
<td>4</td>
<td>0.67</td>
<td>0.25</td>
<td>2.44E-03</td>
<td>2.99</td>
<td>3.30</td>
</tr>
<tr>
<td>G</td>
<td>GGA</td>
<td>1</td>
<td>0.17</td>
<td>-0.58</td>
<td>8.66E-15</td>
<td>0.78</td>
<td>0.47</td>
</tr>
<tr>
<td>Amino acid</td>
<td>Copy number in GFP</td>
<td>Amino acid</td>
<td>Copy number in GFP</td>
<td>Amino acid</td>
<td>Copy number in GFP</td>
<td>Amino acid</td>
<td>Copy number in GFP</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------</td>
<td>------------</td>
<td>--------------------</td>
<td>------------</td>
<td>--------------------</td>
<td>------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>H</td>
<td>9</td>
<td>M</td>
<td>6</td>
<td>T</td>
<td>16</td>
</tr>
<tr>
<td>V</td>
<td>2</td>
<td>I</td>
<td>12</td>
<td>F</td>
<td>12</td>
<td>W</td>
<td>1</td>
</tr>
<tr>
<td>Y</td>
<td>3</td>
<td>L</td>
<td>21</td>
<td>P</td>
<td>10</td>
<td>Y</td>
<td>11</td>
</tr>
<tr>
<td>* UAA</td>
<td>0</td>
<td>NaN</td>
<td>0</td>
<td>* UAG</td>
<td>0</td>
<td>NaN</td>
<td>0</td>
</tr>
<tr>
<td>S</td>
<td>0</td>
<td>0.20</td>
<td>0</td>
<td>S</td>
<td>2</td>
<td>0.17</td>
<td>-0.67</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0.10</td>
<td>0.11</td>
<td>C</td>
<td>1</td>
<td>0.17</td>
<td>-0.11</td>
</tr>
<tr>
<td>* UGA</td>
<td>0</td>
<td>NaN</td>
<td>0</td>
<td>* UAG</td>
<td>0</td>
<td>NaN</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td>1</td>
<td>0.17</td>
<td>0.18</td>
<td>W</td>
<td>1</td>
<td>0.17</td>
<td>0.18</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0.20</td>
<td>-0.52</td>
<td>F</td>
<td>2</td>
<td>0.33</td>
<td>0.52</td>
</tr>
<tr>
<td>L</td>
<td>1</td>
<td>0.17</td>
<td>-0.54</td>
<td>L</td>
<td>1</td>
<td>0.17</td>
<td>-0.54</td>
</tr>
<tr>
<td>* Met is partly initiation tRNA and partly tRNA decoding regular Met codons. We assumed that about half of the Met tRNAs are used for initiation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table S3 Amino-acid usage in the GFP sequence**

For each amino acid the table lists the number of times it is used in the GFP protein.

![Table](attachment:table.png)
Table S4. List of all codons which even after partial correlation still had a significant (p-value<0.1) correlation.
The codons with p-value <0.1 are:

<table>
<thead>
<tr>
<th>Codon</th>
<th>Minimum negative partial correlation (p-value)</th>
<th>Controlling codon in the partial correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCA</td>
<td>-0.29 (0.0004)</td>
<td>CAU</td>
</tr>
<tr>
<td>CAU</td>
<td>-0.24 (0.003)</td>
<td>AAU</td>
</tr>
<tr>
<td>AGU</td>
<td>-0.17 (0.04)</td>
<td>UCA</td>
</tr>
<tr>
<td>AAU</td>
<td>-0.156 (0.06)</td>
<td>CAU</td>
</tr>
<tr>
<td>GGA</td>
<td>-0.156 (0.06)</td>
<td>CAU</td>
</tr>
<tr>
<td>GUC</td>
<td>-0.15 (0.07)</td>
<td>UCA</td>
</tr>
</tbody>
</table>
Figure S1
Figure S2
Figure S3

A

OD

location of first copy of codon CAU

B

OD

location of first copy of codon UCA
At least H codons translated before assembly
Figure S5

A. Bottleneck relative location

B. Bottleneck relative strength
Figure S6
Figure S7

A  Bottleneck relative location

B  Bottleneck relative strength in region 0.16-0.28
5.2 A Comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool

Researching the codon usage, we often refer to aspects such as the codons’ translation speed or the ability of the cell to translate a codon. This ability is affected by the availability of the relevant tRNAs, namely the composition of the tRNA pool. In this project our aim was to study the tRNA pool of *S. cerevisiae* and gain a global understanding on the tRNA availability in cells. This project is a joint project between Dr. Zohar Bloom-Ackerman and me. I wrote all the analysis scripts and took part in designing the experiments and analyzing the data. The manuscript was accepted to PLOS Genetics.

To this aim, we created a comprehensive tRNA deletion library in *S. cerevisiae*, where in each strain a single genomic tRNA gene was deleted. Our tRNA deletion library includes 204 deletions, out of the 275 genomic tRNA genes identified in *S. cerevisiae*, covering all 20 amino acids and 40 out of the 42 anti-codon families. In addition the library consists of a selection of double, triple tRNA deletion mutants.

To assess the contribution of each tRNA gene to cellular growth, we set to accurately characterize the growth dynamics of each deletion strain. To this end we implemented a robotic method to screen and score growth phenotypes of all tRNA deletion strains in multiple conditions (for more details see Method 4.2.1).

We screened the deletion library under a diverse set of growth conditions, including different metabolic challenges and stress-inducing reagents. In rich medium (YPD) only 13% of the library strains demonstrated a phenotype in growth rate and 27% showed a growth yield phenotype (Figure 2A-B). Most strains exhibited a phenotype only in one of the two parameters, which are extracted from distinct stages of the growth, resulting in no correlation between the two parameters ($r=-0.02$, $p$-value 0.8). Apart from the tRNAs who appear in only one copy in the cell (singleton) whose deletion strains are often dead or exhibit impaired growth, we could not explain the observed growth phenotypes, in either growth rate or yield, by either family size, or amino acid identity.
The percent of deletion strains exhibiting any growth phenotype varies between the different stresses (Figure 2C-D), indicating the demand for the different tRNA genes varies across conditions. In all stresses except proteotoxic stress most tRNA deletion strains do not exhibit any growth phenotype, indicating robustness to tRNA gene deletion. We showed that this robustness is enabled due to a wide spread backup between the different gene copies of the same tRNA family and between different families translating the same amino-acid.

![Figure 2](image)

**Figure 2. Deletion strain measurements in different conditions** (A-B) Distribution of phenotypes for the tRNA deletion library on rich medium (YPD), according to two growth parameters: growth yield (A) growth rate (B). Deletion strains were assigned to categories according to their σ values. Thus, highly impaired for σ<−3, impaired for −3<σ>−2, improved for 2<σ<3, and highly improved for σ>3 for σ calculations see Method 4.2.1. (C-D) Percent of strains exhibiting a growth yield (C) phenotype and growth rate (D) phenotype in various conditions. The color indicates the type of phenotype: impaired (blue) or improved (red).

Although it is often implicitly assumed that all tRNA copies contribute similarly to the cellular tRNA pool, comparison of the growth parameters of tRNA deletions from the same family revealed marked differences between seemingly identical family members. In particular, under rich medium, we found that deletions from 21 out of the 32 examined, multi-copy families, span a broad range of at least 10% difference in growth yield (Figure 3A). Such differences
were also exposed in the growth rate parameter but were milder and we thus focus on the growth yield parameter in all further analysis. To further investigate these differences we focused on the tR(UCU) family that contains 11 identical copies in the genome, of which five are represented in our library. In rich medium two copies (tR(UCU)E and tR(UCU)M2) showed appreciable reduction in growth yield (termed Major copies), while deletions of the other three copies (tR(UCU)M1, tR(UCU)G1 and tR(UCU)K) grow essentially as the wild-type (termed Minor copies). Further assessment of the Major and Minor copies across various stress conditions revealed that the hierarchy of Major and Minor is generally preserved (Figure 3B).

Since all family members have identical sequence, we hypothesized that differential contribution should be due to differences in their flanking regions. A complementation assay in which the different tRNAs from the UCU family were introduced to the tR(UCU)M2 deletion strain revealed different degrees of complementation. The different constructs differ only in the region flanking the tRNA gene (200bp); thus the variation in complementation capability can be attributed to the different sequences flanking the tRNA. Next we turned to look which sequence features could govern differential expression of the family members and hence their differential contribution to fitness. For this purpose we used the data set created by Giuliodori et al. (Giuliodori et al., 2003) in which identified conserved sequence elements upstream of S. cerevisiae tRNA genes. In their study they identified four conserved sequence elements located at positions -53 (T-rich), -42(TATA-like), -30(T-rich) and -13 (pol III TSS) with respect to the first nucleotide of the mature tRNA. Our analysis revealed that only the two Major copies contain the conserved TATA-like motif at nucleotide -42. Examining the entire tRNA deletion library in rich medium, we found that strains exhibiting impairment in growth yield were enriched for the same TATA-like motif (hypergeometric test, p-value 0.0089). In addition, the TSS motif at position -13 was enriched in deletion strains that exhibit impairment in either in growth rate or growth yield (Figure 3C). To reinforce these observations we ran the MEME motif search algorithm (Bailey & Elkan, 1994) screening the upstream sequences of tRNA deletion strains exhibiting impaired growth yield for enriched motifs (see Methods). As can be seen in (Figure 3D),
we found two significant motifs that resemble those found by Giuliodori et al.
both in sequence and location.

Figure 3. Different genes of the same tRNA family exhibit phenotype differences (A) Relative growth yield values of the tRNA deletion library strains in rich medium, sorted by anti-codon and amino-acid identity along the x-axis. Each dot along the vertical lines denotes the value (data are represented as mean of 3 biological repetitions +/- SEM) of a deletion strain of different tRNA gene of the respective family. The horizontal lines mark two standard deviations around the mean of the wild-type. Dots above or below these lines are considered non-normal phenotypes. (B) Relative growth yield of the five tR(UCU) members across different growth conditions, indicated on the x-axis. (C) Enrichment of conserved elements in tRNA genes divided according to phenotype observed in rich media for each growth parameter. Each column in the matrix denotes a conserved element as defined by (Giuliodori et al., 2003). Color bar indicates the -log10 of the hypergeometric p-value. (D) log10 E-value found by the MEME software for the most significant motif in a 9bp window starting from the position indicated by the x-axis. The LOGOs of the two significant motifs are displayed below, next to a number indicating its position. Position 0 is the first position of the mature tRNA.

Together these results indicate that the contribution to the tRNA pool and cellular fitness of different copies from the same tRNA family is far from equal. We provide one possible explanation, which can account for the observed
differences, implying that the sequences flanking tRNA genes play a role in determining their expression level.

To determine whether changes in the tRNA pool result in a distinct molecular signature, we examined five tRNA deletion strains using mRNA microarrays (see Methods). For each strain, we measured genome-wide changes in mRNA levels compared to the wild-type under rich growth conditions. The expression changes we exposed were modest and demonstrated a nice correspondence between the essentiality of the tRNA gene and the extent of changes in mRNA expression inflicted upon its loss. Clustering of the expression changes for all five deletion strains revealed that the strains could be divided into two groups: SC (the single copy family genes & initiator Methionine) and MC (the multi-copy family genes), Figure 4A. An example for this division can be found in the pronounced effect observed for the COS8 gene. This gene was extremely up-regulated (about 16 fold) in the SC group while unchanged in the MC group (Figure 4B). This group division recapitulated a division found between the strains when exposed to proteotoxic stress (see full paper at the appendix for details). Thus, these results, suggest different molecular signatures for the two groups, which are also related to the proteotoxic stress response.

To expose what are the responses and underlying molecular pathways that differentiate these two groups, we examined which KEGG pathways (Kanehisa & Goto, 2000, Kanehisa et al., 2012) differentiate between them. To this end we used Gene Set Enrichment Analysis (GSEA) software, which computationally determines whether predefined set of genes shows statistically significant difference in representation between two biological groups (Subramanian et al., 2005, Mootha et al., 2003). This analysis revealed a somewhat opposite signature between the two groups (Table 1). Pathways which are responsive to proteotoxic stress such as the “Proteasome” (FDR q-value <1E-5), and “Protein processing in endoplasmic reticulum” (FDR q-value 2E-3) are significantly induced in the SC group, relative to the MC group. While in the MC groups translation-related pathways such as “Ribosome biogenesis” (FDR q-value <1E-5) and “Ribosome” (FDR q-value 1E-4) are significantly induced compared to the SC group.
To further characterize these differences we focused on specific pathways. A more detailed examination of the expression changes observed for all the genes that constitute the proteasome complex revealed an up-regulation, to various extents in response to deletion of tRNAs from the SC group. The MC group demonstrated no change and even a slight down-regulation of these genes (Figure 4C), a trend which was further verified using RT-qPCR. These observations establish the notion that upon deletion of members of the SC group cells experience a proteotoxic stress. Another distinction between the groups was also observed in the RNA polymerase machinery pathway. Expression of genes that belong to this pathway were up-regulated only in the MC group (Table 1). Further examination revealed that this signal is due to the elevation of Pol III machinery (the polymerase responsible for tRNA gene transcription) and not Pol II machinery. The genes encoding for RNA Pol III machinery demonstrated up-regulation in the MC group and no change or even down regulation in the SC group (Figure 4D); further verified by RT-qPCR. This suggests a potential regulation of the tRNA pool by controlling the levels of the tRNA’s transcribing polymerase.

Figure 4. Molecular response to changes in the tRNA pool. (A) Dendrogram created by clustering changes in gene expression for five representative deletion strains. (B) Fold change of the COSS
(YHL048W) mRNA levels in each of the five deletion strains as measured by microarrays. (C-D) The fold change distribution of mRNA levels as measured by microarrays, of genes composing the Proteasome pathway (C) and the Pol III RNA Polymerase machinery module (D) as defined by the KEGG database. In both sub-figures values are plotted for the same five deletion strains: ΔtL(GAG)G (blue), tR(CCU)J (red), tM(CAU)C (green), tH(GUG)G1 (magenta) and tR(UCU)M2 (cyan).

To summarize, this work revealed additional levels of complexity within the tRNA pool including extensive backup interactions and differential contribution between tRNA copies. Combining these results with the fact that the essentiality of different tRNAs changes across conditions and the potential elevation of Pol III upon deletion all suggest the regulation of the pool is more complex than commonly accepted and perhaps even as complex as for mRNA genes.

The full results and conclusion are summarized in a paper, see appendix.

**Table 1 KEGG pathways differentiating between tRNA deletion sets.** KEGG pathways (Kanehisa & Goto, 2000, Kanehisa et al., 2012) for which changes in genes expression are significantly different between the two groups of tRNA deletion strains: MC (multi-copy) group (ΔtH(GUG)G1 and ΔtR(UCU)M2) vs. SC (single-copy) group (ΔtL(GAG)G, ΔtR(CCU)J, ΔtM(CAU)C) calculated with GSEA (Subramanian et al., 2005, Mootha et al., 2003). In the first column are pathways, which are higher in SC vs. MC and vice versa in the second column. The values are corrected for multiple hypothesis and the FDR q-values are indicated next to each pathway.

<table>
<thead>
<tr>
<th>Higher is SC than in MC</th>
<th>Higher is MC than in SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteasome (&lt;1E-5)</td>
<td>Ribosome biogenesis in eukaryotes (&lt;1E-5)</td>
</tr>
<tr>
<td>Oxidative phosphorylation (&lt;1E-5)</td>
<td>RNA polymerase (&lt;1E-5)</td>
</tr>
<tr>
<td>Endocytosis(2E-3)</td>
<td>Phenylalanine, tyrosine and tryptophan biosynthesis (&lt;1E-5)</td>
</tr>
<tr>
<td>SNARE interactions in vesicular transport (2E-3)</td>
<td>Pyrimidine metabolism (5E-5)</td>
</tr>
<tr>
<td>Protein processing in endoplasmic reticulum (2E-3)</td>
<td>Ribosome (1E-4)</td>
</tr>
<tr>
<td>Starch and sucrose metabolism (2E-3)</td>
<td>Lysine biosynthesis (1E-4)</td>
</tr>
<tr>
<td>Citrate cycle (TCA cycle) (0.01)</td>
<td>Histidine metabolism (4E-4)</td>
</tr>
<tr>
<td>Meiosis - yeast (0.01)</td>
<td>Cysteine and methionine metabolism (4E-4)</td>
</tr>
<tr>
<td>Homologous recombination (0.02)</td>
<td>Riboflavin metabolism (5E-3)</td>
</tr>
<tr>
<td>Mismatch Repair (0.02)</td>
<td>Arginine and proline metabolism (8E-3)</td>
</tr>
<tr>
<td>Cell cycle - yeast (0.02)</td>
<td>Valine, leucine and isoleucine biosynthesis (0.01)</td>
</tr>
<tr>
<td>MAPK signaling pathway - yeast (0.02)</td>
<td>Purine metabolism (0.03)</td>
</tr>
<tr>
<td>Fructose and mannose metabolism (0.02)</td>
<td>Sulfur metabolism (0.03)</td>
</tr>
<tr>
<td>Nitrogen Metabolism (0.02)</td>
<td>Tyrosine Metabolism (0.03)</td>
</tr>
<tr>
<td>Phagosome (0.03)</td>
<td>Folate biosynthesis (0.04)</td>
</tr>
</tbody>
</table>
5.3 Ribosome density governs patterns of mRNA cleavage in *Escherichia coli*

One of the main properties of an mRNA which can have a direct effect on the final gene expression is its stability. Changes in codon sequence were found to affect mRNA levels and stability (Kudla et al., 2009, Sunohara et al., 2004, Petersen, 1987, Kolmsee & Hengge, 2011). In this project our aim was to study the interplay between degradation and translation in *Escherichia coli*. The manuscript is under revision in *Nucleic Acids Research*. 
Ribosome density governs patterns of mRNA cleavage in *Escherichia coli*

Sivan Navon and Yitzhak Pilpel*

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

*To whom correspondence should be addressed. Tel:+972-8-9346058 ; Fax: +972-8-9344108 ; Email: Pilpel@weizmann.ac.il
ABSTRACT

Recent developments in microbial genomics allow the study of ribosome profiling and RNA fragments in the cell. Analysing ribosome profiles it was discovered that Shine-Dalgaro-like motifs embedded within coding region cause ribosome pausing in bacteria. Yet, what are the potential functions of such ribosome pausing and attenuation sites remains unknown. Using these developments we set to unravel a regulatory relation between translation and mRNA degradation. We used genome-wide mapping of RNA fragments to identify the Escherichia coli's mRNA cleavage sites and combined it with measurements of transcriptome-wide ribosome occupancy, exposing intricate relation between the two. We found a tendency to have a Shine-Dalgaro-like sequence and high ribosome occupancy downstream to cleavage sites, suggesting that the ribosome enhances degradation locally, immediately up-stream to its attenuation site. Conversely, genes with higher ribosome density around the start codon were often devoid of detectable cleavage points, suggesting that globally the ribosome protects the transcript. Our results expose for the first time on a genome-wide scale, a coupling between translation and mRNA degradation and ascribe the first functional role for ribosomal pausing in bacteria.

INTRODUCTION

The amount of protein synthesized in the cell is determined by two primary parameters: the mRNA levels of the genes and the rate of translation of this mRNA. mRNA levels in turn, are determined by both the rate of transcription and the rate of degradation. In bacteria the half-life of different transcripts varies significantly, from less than a minute to about an hour (reviewed in:1), thus allowing the stability of the mRNA to play a key role in regulating the transcriptome.

In Escherichia coli degradation of mRNA is initiated by endonucleolytic cleavage, the upstream fragments are degraded by 3'-to-5' exoribonucleases while the downstream fragments degrade through a series of successive endonucleolytic cleavages (2-4). Of the three site-specific endoribonucleases in Escherichia coli, RNase E is the main enzyme in mRNA decay (2,3). The common notion is that this enzyme cleaves at specific sites in single stranded regions of mRNAs that are typically AT-rich with no strong sequence specificity and with a stem&loop structure frequently found around the site (reviewed in: 2,3,5). For more on mRNA decay in bacteria, see reviews by (2,3,6).

The mRNA stability is tied to the translation process through two known stabilizing mechanisms: first, the ribosome can simply protect the mRNA by masking a potential cleavage site, namely the ribosome footprint protects the region from actually being cleaved (Reviewed in: 4). The second mechanism that coordinates translation and mRNA decay is through the blocking of transcript's 5'. RNase E degradation pathway requires access to the 5'-end of the transcript and when this access is blocked the mRNA is
stabilized. For both mechanisms an efficient ribosome binding site (RBS) can have a stabilizing effect on the mRNA (4,7). Stalled ribosomes on the 5’ end of the ermC were shown to stabilize the gene in B. subtilis (8). However not only the 5’ UTR features can affect the mRNA stability but also properties of the ORF which affect the ribosome elongation, such as codon usage and Shine-Dalgarno like motifs (9, see below). Rare codons in the RpoS gene were found to stabilize its transcript, potentially by concentrating the ribosome on the mRNA thus protecting it from RNase E attack (10). Earlier Petersen showed that altering codons immediately downstream to the ribosome binding site affect the transcript stability (11). It was suggested that these codons affect the transcript stability by changing the ribosome density on the gene (12). Though coupling between translation and the degradation was exposed for specific genes in the past, whether it is a genome wide phenomenon and what are the coupling mechanisms remain to be seen.

Until now, analysis of cleavage sites and their sequence properties was done by analyzing a handful of genes and different oligonucleotides. Here we set to analyze the cleavage sites of more than a thousand messenger RNAs in Escherichia coli. By analyzing 5’ RNA sequencing reads mapped onto the genome (13,14) we identified mRNA cleavage sites. Combining this data with the recent data on transcriptome-wide ribosome occupancy (9) we explored the relation between translation efficiency and the mRNA cleavage and stability. We discovered a peak of higher ribosome density 25 bps downstream to the cleavage sites potentially due to higher anti-Shine-Dalgarno (aSD) affinity of the region. This suggests that accumulation of ribosomes at an attenuation site enhance local mRNA cleavage immediately upstream to them. On the other hand, higher ribosome density at the 5’ termini of genes was found to be associated with more stable mRNAs, suggesting that globally the ribosome exerts a protective effect on the transcript. Thus, by analyzing cleavage sites of many genes this work sheds light on the potential modes of coordination between translation and degradation.

**MATERIAL AND METHODS**

**Identification of RNA cleavage sites from RNA deep sequencing data**

We used data derived from a modified RNA deep sequencing protocol (15) to deduce location of endonucleolytic cleavage sites. The Mapping of E. coli 5’-end monophosphorylated RNA fragments were taken from the Quax paper (13) which were downloaded from [http://www.weizmann.ac.il/molgen/Sorek/Navon_data.fasta.gz](http://www.weizmann.ac.il/molgen/Sorek/Navon_data.fasta.gz). Analyzing the data we noticed that the 5’-end of tens of thousands of fragments was identified to be inside coding regions. Fragments inside ORF are less likely to be transcription start site, suggesting that these fragments represent the downstream residue of an endonucleolytic cleavage site. Since E. coli does not have a 5’-to-3’ exoribonuclease and endonucleolytic cleavage is a main step in the RNA degradation pathway, such data can be used to deduce the genome wide features of cleavage sites that cause mRNA degradation.
Analyzing only coding regions, the data consists of 13,926 unique potential cleavage sites in the genome, which represent mapping of unique RNA fragments (we set a quality threshold such that each unique location had to appear at least in two independent reads in the data to be considered). These 13,926 cleavage sites distributed over 2218 mRNAs of the *Escherichia coli* reference genome NC_000913, implying that the rest of the 1923 annotated ORFs do not show even a single cleavage point in our final dataset. After filtering of sites that less not likely to represent cleavage sites (see Materials and Methods) we focused on 4157 high quality sites.

**Filtering cleavage sites**

For both the *E. coli* and the *P. aeruginosa* all identified sites were clustered into groups (clusters) based on the distance between consecutive sites, two sites which were separated by less than 5 nucleotides were clustered together. As a result different clusters had at least 5 nucleotides between their closest sites.

Filtering the *Escherichia coli* data: For a site to be considered as a cleavage site it should fulfill three criteria: (i) It needs to be the site with highest number of reads in its cluster, (ii) it needs to have at least three reads and no more than 50, (iii) it needs to be at least 10 nucleotides upstream or downstream from a known transcription start site (based on the EcoCyc database (16)).

Filtering the *Pseudomonas aeruginosa* data: sites that appeared in either the 37°C sample or the 28°C sample were considered, (as long as the site appeared in both the Tap+ and Tap-). For a site to be considered a cleavage site if it fulfilled three criteria: (i) It needs to be the site with highest number of reads in its cluster, (ii) it needs to have more than two reads and no more than 50 (averaging the Tap+ and Tap- reads), (iii) it needs to be at least 10 nucleotides upstream or downstream from the transcription start sites identified by Wurtzel *et al.* (14).

As Li *et al.* did in their kinematic analysis, for the analysis of the ribosome density around the cleavage sites we avoided regions with known changes in the ribosomes density. In particular we excluded sites which were close (up to 50 bases) to the start or the stop codon, since it is known that the ribosome dwells longer on these particular codons (17). By analyzing only sites in the middle of the gene, where the ribosome density should be relatively constant and there are no known causes for SD-like sequence to exist there, the observed changes can be associated to cleavage sites.

**mRNA folding energy and unpairing score calculations**

We used a sliding window of 20 nucleotides to create a folding energy and pairing score profiles for all mRNAs. Each window of 20 nucleotides was folded using Vienna RNAfold package (18). The free energy of the window was assigned to the 11th nucleotide in the window. The unpairing score for each nucleotide was calculated by counting the number of times it was unpaired over it total number of appearances (20
appearances for nucleotides far from the edges of the transcript). The secondary structure analysis was run with different window sizes (in the range of 10 to 100 nucleotides) with no significant differences in the results (not shown).

**Calculating the genes’ ribosome density**

Using data deposited in GSE35641 (9) we constructed the genes’ ribosome occupancy. The ribosome occupancy for each gene was normalized to the sum of the ribosome occupancy over all the mRNA including its UTRs, to include 50 nucleotides upstream and downstream to it. We chose 50 nucleotides since the average distance of the main TSS site identified in the E. coli 5’ data is 49 nucleotides. Then the ribosome density profile was average between the two repeats to get the final ribosome density of a gene.

**Defining expressed, cleaved and not-cleaved genes**

For the expression level of the genes we used the average mRNA in supplement database 2 in Lu et al. paper (19), which averages three different mRNA level papers. After ignoring all the genes which had no mRNA level reads and genes which are shorter than 200 bases, we ended up with 2025 genes, we took the top 50% of these genes which had the highest mRNA levels ending up with 1012 which we defined as expressed genes.

From these 1028 genes were constructed two set of genes. The first set included only genes which had no cleavage site at all, a total of 246 genes. The second set included genes which had at least one cleavage site after filtering a total of 541 genes. The rest of the expressed genes had cleavage site which did not pass the filter (see filtering cleavage sites).

**RESULTS**

**Sequence information is found around cleavage sites**

Examining the sequence content around the cleavage site some consensus emerges, mainly in nucleotides -1 to 1 (“0” is defined to be the first nucleotide position immediately downstream to the cleavage site, i.e. the first nucleotide in the RNA fragment detected). Reassuringly we noticed that indeed just around the cleavage site there is an AT-rich region, about 70% of nucleotides are either A or T for all three sites (-1 to 1), see Supplement Figure 1. Clustering the sites based on their sequence (using the Jukes-Cantor distance (20)) revealed two potential motif groups. The first motif, [G/A]N↑[A/T]TT (↑ denotes the cleavage site, Supplement Figure 1B) includes the core consensus sequence ATT suggested by Ehretsmann et al. (21) that was obtained from a handful of data points, while the first [C/A]T↑ (Supplement Figure 1C) was not detected before, to the best of our knowledge.

**High ribosome occupancy downstream of cleavage sites**
A relation between the presence and density of ribosomes and mRNA degradation was repeatedly suggested \((4,7,10,12)\), particularly due to the potential of the ribosome to protect the mRNA from degradation (reviewed in: 4). Recent profiling of ribosome density along mRNAs in *Escherichia coli* \((9)\) allows us to examine potential relation between the ribosome density along transcripts and cleavage sites.

First we examined the ribosome occupancy around the cleavage sites, Figure 1A. We worked throughout with ribosome occupancy profiles that are normalized for each gene individually, thus depicting relative changes in occupancy along a gene, and ignoring absolute differences between genes. As can be seen in the figure, around 12 bases downstream to the cleavage site there is a deep in the ribosome density profile (Wilcoxon rank sum test \(p<1e^{-27}\) compared to the density of all 100 bases around the cleavage site) and around 24 bases downstream there is a peak in the ribosome density profile (Wilcoxon rank sum test \(p<1e^{-31}\) compare to the density of all sites 100 bases around the cleavage site). The differences in the ribosome density could not be explained by codon bias (see supplement Figure 2).

Following Li et al.’s \((9)\) work we examined whether Shine-Dalgarno (SD) like sequences drive ribosomes pausing downstream to our candidate cleavage points. Using the same method as in Li et al. we constructed the anti-Shine-Dalgarno \((\text{aSD})\) affinity profile of genes (the affinity to bind to the complementary sequence of the Shine-Dalgarno, which is at the 3’ end of the 16S rRNA) profile around each cleavage site. We found a high aSD affinity about 13 bases downstream to the cleavage site (Wilcoxon rank sum test \(p<1e^{-12}\)), see Figure 1B; this peak in the aSD affinity is located 11 bases upstream to the peak of the ribosome occupancy. Since the ribosome occupancy data are aligned to the A-site, the peak in the aSD affinity is 8 nucleotides upstream to the P-site which is in range for an effective SD \((22)\).

To further examine the relation between cleavage sites and SD-like sequences we turned to another bacterial species, *Pseudomonas aeruginosa*. Using 5'-end mapping of RNA fragments published by Wurtzel et al. in this specie \((14)\) we calculated the aSD affinity around the *P. aeruginosa* cleavage sites (see Materials and Methods). As in *E. coli* we find a high aSD affinity about 13 bases downstream to the cleavage site (Wilcoxon rank sum test \(p<1e^{-4}\)), see Figure 1C.

High ribosome occupancy downstream potentially increases cleavage probability by keeping the region upstream to it free of secondary structure, as required by RNase E. To further investigate the connection between cleavage and the structure of the RNA we computed predicted mRNA secondary structure around the cleavage sites. We examined the averaged free energy profile around the cleavage sites and searched for significant differences in free energy and potential pairing of the nucleotides (see Material and Methods). For comparison we examined profiles of randomly permuted sequences around cleavage sites and sequence without cleavage sites. Reassuringly we found differences in the free energy around the cleavage site (Figure 2A). While around the cleavage site the energy is 60% higher
(looser structure), the structure becomes tight (low free energy values) further upstream and downstream from the cleavage site. In addition, from Figure 2C we notice there is a significant difference between the amounts of base pairing of some nucleotide position around the cleavage site, mainly base 0 and base -6 (Wilcoxon rank sum test compared to background: p=3e-21, p=2e-54 respectively). Similar figures were obtained for folding P. aeruginosa, see figures 2B and 2D

Ribosomes stabilize transcripts

While so far we revealed that ribosomes can affect locally degradation by enhancing cleavage immediately up-stream, here we turned to examine more global effect of ribosomes on stability of their harboring transcripts. As mentioned above, we identified one or more cleavage sites in only about a half of the E. coli genes; for the rest of the genes no such site was seen in the data. To further investigate why some genes show cleavage sites and others do not, we examined the ribosome density of cleaved and non-cleaved genes. For this analysis we avoided genes with low expression levels (see Material and methods) as these are unlikely to show cleavage events. The expressed genes were divided into two sets: without any observed cleavage site, and genes with at least one cleavage site (see Material and Methods). As can be seen in Figure 3A, expressed genes without cleavage sites have significantly higher ribosome density around their start codon. This observation might suggest that high ribosome density close to the 5’ UTR exerts a global protecting effect on RNAs from cleavage.

To further check the relationship between the ribosome density around the start codon and the gene cleavage we divided all of the E. coli genes into 3 equal groups based on their relative ribosome density in nucleotide positions -6 to +4 relative to the ATG. For each group we checked the fraction of expressed genes which had cleavage sites, Figure 3B. The set of genes with low relative density had significantly higher fraction of genes with cleavage sites than expected (hyper-geometric probability 0.0015) and the set of genes with high density had significantly lower fraction of genes with cleavage sites than expected (hyper-geometric probability 0.0028). To conclude, genes with low relative ribosome density around the start codon are more likely to have cleavage sites which suggest that stalled ribosomes around the ATG protect the transcript from cleavage.

DISCUSSION

The modified RNA deep sequencing protocol (15) enables, as far as we know for the first time, to analyze cleavage sites of bacteria on a single nucleotide resolution on a genome-wide scale. This genome-wide analysis reinforces the notion that the mRNAs are cleaved in AT-rich regions that have a relatively loose structure.

An important aspect of cleavage control that is analyzed here is the coupling between translation and degradation. Coupling between layers of gene expression regulation is a common theme (23) and here
we provide one such mode of coupling. We notice high relative ribosome occupancy about 25 bases downstream to the cleavage site suggesting the ribosome enhances degradation locally. Due to the ribosome size (foot-printing of 25-42 nucleotides (9)) a cleavage site which is 25 bases upstream to the ribosome A-site will most likely be free of the ribosome but very close to it, thus with limited ability to refold. The single stranded requirement of the RNase E is thought to be maintained by the RNA secondary structure; however here we suggest an additional mechanism. We suggest that the ribosomes take an active role in keeping the cleavage sites single stranded by dwelling a bit longer just downstream of the cleavage sites.

While the ribosome enhances degradation locally, we found that globally it exerts global protection to the transcript. On a genome-wide scale we notice that high ribosome occupancy around the start codon may stabilize the transcript. Such ribosomes have the potential to mask essential cleavage sites and jam other ribosomes on the 5’ UTR, thus blocking it from the degradation machinery which requires access to the 5’ end of the transcript.

In the future it will be interesting to repeat this analysis on data from different bacteria, on mutants which have different RNases knock down and in different stress conditions, which could activate additional degradation pathways (24) and modify the ribosome patterns genome wide (9). Such data can shed more light on the degradation process, its coupling to the translation process and its different enzymes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We acknowledge grant support from the European Research Council and the Ben-May Charitable Trust. YP is the incumbent of the Ben-May Professorial Chair. Special thanks to Omri Wurtzel a former PhD. Student at Rotem Sorek’s lab for extensive assistance throughout. We also thank Gene-Wei Li from Jonathan Weissman’s lab for his help with the ribosome profiling data and analysis.

FUNDING

This work was supported by the European Research Council (through grant number 205199-ERNBPTC); and the Ben-May Charitable Trust. Funding for open access charge: the European Research Council.
REFERENCES


TABLE AND FIGURES LEGENDS

**Figure 1. Ribosome Occupancy and SD-like sequences around the cleavage site**

(A) For each base aligned around the cleavage site the median of the ribosome density profile in *E. coli* is plotted. The confidence intervals plotted by thin lines were found using bootstrapping. The black line is the median of all the bases between -50 to 50 around the cleavage site regardless to their location. (B-C) mean aSD affinity profile around the cleavage site for the *E. coli* sites (B) and the *P. aeruginosa* (C). The profile was calculated by averaging the aSD binding affinity (as calculated by Li et al (9)) for each base over all cleavage sites. The confidence intervals (thin lines) were calculated using bootstrapping. The plot was smoothed with a moving average of 3 bases to reduce frame effects. The black line is the average of all point in the ±50 region. As can be seen there about 11 nucleotides upstream of the high ribosome occupancy there is a high aSD affinity which could explain the ribosome occupancy.

**Figure 2. Secondary structure features around the cleavage site**

In the figure are plotted the mRNA secondary structure properties around the *E. coli* (A&C) and *P. aeruginosa* (B&D) cleavages sites calculated as describe in the Material and Methods. (A, B) The free energy aligned around the cleavage site averaged over the cleavage site sequences; *E. coli* (A) and *P. aeruginosa* (B). (C, D) The unpairing score profile aligned around the cleavage site averaged over all the sites; *E. coli* (C) and *P. aeruginosa* (D). In each subplot in addition to the profile for all cleavage sites (solid blue line) are plotted two additional controls: (i) after randomly permutating the sequences of the cleavage site (line-dot black), and (ii) sequences without cleavage sites (dashed gray).

**Figure 3. Ribosome Density of genes with and without cleavage sites**

(A) From the express genes were two groups were constructed: (i) expressed genes with at least one cleavage site (red) (ii) expressed genes without any site, even ones that didn’t pass our filter (blue). For each group the median of the ribosome density is plotted when the genes are all aligned to the start codon (starting with nucleotide 0). It is important to notice that while the ribosome is assembled on the “P site”, the ribosome’s location is mapped to its “A” site in the ribosome data. Therefore high ribosome occupancy on the ATG is shifted toward the second codon. The error bars are the standard deviation for nucleotide three calculated using bootstrapping. (B) The genes in subplot (A) were divided into 3 groups.
depending on their level of ribosome density in nucleotides -6 to 4 (thresholds were set to have 3 equal size groups when analyzing all genes). For each group of genes the fraction of cleaved genes is plotted. The error bars (red) are the standard deviation calculated using bootstrapping. * indicates that the level of the cleaved genes is significantly different than expected. The 'low' occupancy group has more cleaved genes than expected (hyper-geometric probability 0.0015) while the 'high' occupancy group has less than expected (hyper-geometric probability 0.0028)
Figure 1

A

Ribosome Density

B

base location relative to cleavage site

C

affinity for aSD [kcal/mole]
Figure 2

A. *Escherichia coli*

B. *Pseudomonas aeruginosa*

C. 

D.
Figure 3

A

B

Fraction of genes with cleavage sites

Ribosome Density

low  medium  high
Supplementary Figures

Supplement Figure 1. Sequence information of the cleavage sites

(A) Sequence logo (Crooks et al., 2004) of 4157 cleavage sites. After clustering we get two groups of genes: (B) Sequence logo of the first group of genes (2321 genes), (C) Sequence logo of the second group of genes (1836). In all subplots base 0 is the first nucleotide just after the cleavage site (the cleavage is between base -1 to base 0). (D-F) is the probability logo of: all 4157 cleavage sites (D), first group of genes (E), second group of genes (F).
Supplement Figure 2. Codon’s translation time around the cleavage sites.

The figure shows the mean time it takes to translate the region around the cleavage site. One over the codon adaptiveness index from the tAI definition (dos Reis et al., 2004; Tuller et al., 2010) was used as a proxy for the time it takes to translate each codon. We constructed a translation time profile for each gene by dividing the time it takes to translate the codon evenly between its 3 nucleotides. Then we aligned these profiles around the reliable cleavage sites to get the mean translation time around the cleavage sites.
6 Discussion

The ability to predict final protein levels based on the nucleotide sequence is a major challenge for proteomics, computational biology and biotechnology. During the translation process the codon sequence is decoded into amino acids by tRNAs. Most amino acids are coded by more than one codon. Though resulting with the same polypeptide the levels of protein were found to change by up to 100 fold just by changing the codon sequence which in turns changes the mRNA secondary structure and the translation rates (Kudla et al., 2009, Welch et al., 2009, Subramaniam et al., 2013). To predict these changes we need to understand which additional information is coded by the codons in addition to the amino-acid sequence.

During the years a few indices were developed to try to predict protein levels based on the codon sequence. The two famous indices are the CAI (Sharp & Li, 1986) and the tAI (dos Reis et al., 2004). Both indices score a gene according to the adaptation of its codon sequence to the cellular translation capabilities. Though these indices were shown to correlate with protein levels in different organisms, they failed to predict protein levels in synthetic experiments in which the same protein was expressed with different codon combinations.

The basic assumption that underlines these indices and their weakness point is that they only look at the gene’s codon composition regardless to the codon’s location. In an earlier work we realized that in many organisms across the three domains of life the codons are not distributed evenly across the gene. We found that the first codons tend to be “slow codons”, i.e. codons which takes longer to be translated (Tuller et al., 2010a). In the work that followed this initial publication we focused on the single gene level and examined how regions of “slow codons” affect protein expression. We found that localizing the slowest codons in the 5’ end is correlated with higher protein levels, while localizing them in the 3’ end correlates with lower protein levels. In addition, we found that the time it takes to translate these codons affected the protein levels, if the slowest codons are indeed located at the 5’ end, then the slower region the higher will the protein levels be. In contrast if the slowest region is closer to the 3’ end of the transcript then the slower that region is the lower, on average, is the expression
level. By analyzing the codons of a given gene and locating the slowest region we believe that we can improve the prediction of protein levels and improve the design of synthetic genes.

In the past few years, much work was done to in the field trying to settle whether different codon usage for the same protein results in different expression level is due to secondary structure changes in the mRNA or due to translation efficiency of different codons. The main focus was on the N-terminal codons which are enriched with codons corresponding to rare tRNAs (Tuller et al., 2010a). As first showed on a small library by Kudla et al. (Kudla et al., 2009) and recently, on a much larger library, by Goodman et al. (Goodman et al., 2013) a reduced RNA structure at the 5’-end and not codon rarity itself is responsible for expression increases. This goes along with findings by Li et al. (Li et al., 2012) which showed by ribosome occupancy data that on rich medium the rate of translation of all codons in E. coli is the same. While the secondary structure seems to be the dominate factor for the N-terminal these results do not explain how codons far from the N-terminal also affect the expression level (Welch et al., 2009). In addition it was recently shown that in amino-acid starvation condition changes in protein expression were due to limited availability of tRNAs (Subramaniam et al., 2013). These limitations should also be taken into account when trying to express proteins in extremely high levels since in that case the synthetic protein codon usage might cause by itself a tRNA shortage in the cell, as suggested by our analysis of the Kudla library (Navon & Pilpel, 2011). Thus, although currently secondary structure is the leading answer I believe that as in many field in biology the answer is not secondary structure or codon translation efficiency but both. Only by understanding both mechanisms and their limitation we will be able to predict when each mechanism will dominate and improve our protein level predictions.

As found in our analysis of the Kudla’s library (Kudla et al., 2009) and also shown by Li et al. and Subramaniam et al. shortage is tRNAs can affect the cell’s fitness and protein levels. Thus, to better predict the protein levels we need to gain a better understanding of the tRNA pool. Since measuring the tRNA concentrations in the cell is far from trivial, we would like to learn about it regulation, thus enabling us to predict their levels just from the genomic sequence.
To study the tRNA pool we created a comprehensive tRNA deletion library in *S. cerevisiae*. When deleting the tRNA genes we found that there is extensive backup between the different tRNAs, resulting in robustness to deletion of most tRNA genes. This suggests that either the tRNAs usually transcribed in higher amounts than required for growth in rich medium or that the transcription rate of the remaining genes increases. The idea that decrease in tRNA levels can be sensed by the cell and result in increased transcription rate of related tRNAs is very intriguing and resembles the backup mechanism found between paralog mRNA genes (Kafri *et al*., 2006).

The tRNA genes are transcribed by RNA polymerase III which binds to internal promoters inside the tRNA sequence. Since many tRNA genes have identical sequences or at least identical internal promoter it is reasonable to expect that those tRNA will have the same transcription rate. Yet, in our study we discovered that even identical tRNA contribute differently to the tRNA pool. The results prove the tRNA surrounding may affect its transcription. We found that upstream motifs maybe the source for the identified differences between the tRNAs. Upstream motifs were shown to affect the tRNA transcription by facilitating the binding of RNA polymerase III (Giuliodori *et al*., 2003, Parthasarthy & Gopinathan, 2005). This again reminds regulation mechanisms of mRNA transcription by RNA polymerase II and suggests the tRNA transcription is more complex than previously assumed and might be as complex as mRNA transcription. This conclusion is reinforced by our results from microarray analysis of a few tRNA deletion strains where we found up-regulation of RNA polymerase III genes in some deletion strains indicating the ability to regulate of the tRNA transcription.

While mRNA transcription rates is studied and modeled thoroughly (Sharon *et al*., 2012) the tRNA transcription is neglected. Our work showed the complexity of the tRNA pool and took the first step toward understanding the underling mechanisms that regulate it. Nevertheless, since the tRNA pool may affect the cell’s fitness (Navon & Pilpel, 2011) and the protein levels (Subramaniam *et al*., 2013), it is clear that for better predictions of protein levels we need better predictions of tRNA levels in the cell.
Researching the information coded by the codons it is clear that the most direct question is how the codon usage influences the translation process through the regulation of the ribosome elongation rates. However, codon usage was shown to also affect the mRNA stability (Sunohara et al., 2004, Petersen, 1987, Kolmsee & Hengge, 2011), the protein folding (Pechmann & Frydman, 2013) and even the host fitness, upon expressing the protein (Kudla et al., 2009). The codons affect these properties of the transcript either through the coupling between the mRNA degradation/folding process with the translation process or through additional levels of information overlaid on the codon sequence. For example, different combination of codons result in different nucleotide sequences thus creating/eliminating overlaid information as sequence motifs.

During my research we encountered twice a coupling between the different processes. One of the results of the tRNA deletion library is that some deletion strains have more misfolded proteins and an elevated protein quality control pathway. This result suggests that reduced levels of a specific tRNA in the pool resulted in a reduced translation efficiency of the related codon(s), which in turn changed the ribosome elongation rate and hampered the correct folding of the protein. A coupling between codon usage and folding was also found by others (Pechmann & Frydman, 2013). However, the exact mechanisms are still unclear and our ability to predict which codons are essential for correct folding is still lacking. A unique feature of our approach is that we disrupted translation-folding coupling without manipulating the translated mRNA directly, but only by modifying the tRNA supply.

The second coupling we explored in my research is the coupling between codon usage and mRNA stability. The work by Kudla et al. who created a synthetic library of the same GFP protein coded with many different codon combinations revealed that codon usage can affect mRNA level even by 6 fold. Understanding this coupling is essential for synthetic biology in which the promoters are used to optimize transcript levels and codons are used to optimize translation efficiency. In my research I focused on mRNA degradation in E. coli and its coupling to the translation process. In bacteria, the mRNA degradation process consists of a series of endonucleolytic cleavages. By combining cleavage site information with ribosome profiling data we found that the ribosomes play a dual role: locally their
accumulation at strategic locations enhances mRNA cleavage immediately upstream, globally their high density on the transcript 5’ termini is associated with a tendency to be protected from degradation.

Surprisingly, the ribosome accumulation which enhances the degradation is not caused by slow codons but rather by the affinity of the downstream sequence to the anti-Shine-Dalgarno (aSD) sequence. This result goes along with the discovery by Li et al. (Li et al., 2012) who showed that in bacteria ribosome pausing is caused mainly by Shine-Dalgarno like sequences. In addition to the Shine-Dalgarno sequence, which was found to affect the ribosome elongation and mRNA stability, we also found that the cleavage sites themselves are not random and that they occur in locations with specific sequence information. However, not only the sequence is important for cleavage but also the transcript secondary structure, which is required to be single-stranded for cleavage. Thus, the secondary structure is another layer of information that causes indirect coupling between the codons and other cell process.

The work described in this thesis exposes four levels of dependencies between codons and gene expression that should be taken into consideration when designing proteins or trying to predict their expression: The first is the ability of the cell to translate the chosen codons based on the availability of the tRNAs. The second is the order of the codons and the load it exerts on the ribosomes. The third are the motifs and the secondary structure which are the results of specific codon choices. The four and perhaps the most complicated of all is the coupling between the different processes such as translation and degradation or folding. The done work here, from the point of view of the gene expression, was mainly done on E. coli while the work on the tRNA pool was done in S. cereviseae. Due to the many difference between the translation, degradation, maturation and coupling between the different processes in eukaryotes vs. prokaryotes additional work is required in all domain of life.

All together, these studies expand our understanding of the basic processes of gene expression, showing some of the less obvious effect of codon selection and putting forth the extensive coupling between the translation and degradation processes in prokaryotes.
7 Literature

7.1 References


7.2 List of publications


8 Declaration

I hereby declare that this thesis summarizes my independent efforts under the supervision of Yitzhak Pilpel.

The only project performed in collaboration was the tRNA deletion library project. This project was performed in collaboration with Dr. Zohar Bloom-Ackerman which was a fellow PhD student, from Yitzhak Pilpel’s laboratory at Weizmann Institute of Science who did all the experimental work. I wrote all the analysis scripts, did the bioinformatics analysis and took part in designing the experiments and analyzing the data.
9 Appendix

9.1 A Comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool (Bloom-Ackerman et al.)

Accepted to Plos Genetic
A Comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool

Zohar Bloom-Ackermann, Sivan Navon, Hila Gingold, Ruth Towers, Yitzhak Pilpel* and Orna Dahan

Department of Molecular Genetics

Weizmann Institute of science

Rehovot, 76100 Israel

*corresponding author, Pilpel@weizmann.ac.il

November 9, 2013
Abstract

Deciphering the architecture of the tRNA pool is a prime challenge in translation research, as tRNAs govern the efficiency and accuracy of the process. Towards this challenge, we created a systematic tRNA deletion library in *S. cerevisiae*, aimed at dissecting the specific contribution of each tRNA gene to the tRNA pool and to the cell’s fitness. By harnessing this resource, we observed that the majority of tRNA deletions show no appreciable phenotype in rich medium, yet under more challenging conditions, additional phenotypes were observed. Robustness to tRNA gene deletion was often facilitated through extensive backup compensation within and between tRNA families. Interestingly, we found that within tRNA families, genes carrying identical anti-codons can contribute differently to the cellular fitness, suggesting the importance of the genomic surrounding to tRNA expression. Characterization of the transcriptome response to deletions of tRNA genes exposed two disparate patterns: in single-copy families, deletions elicited a stress response; in deletions of genes from multi-copy families, expression of the translation machinery increased. Our results uncover the complex architecture of the tRNA pool and pave the way towards complete understanding of their role in cell physiology.
Transfer RNAs are an important component of the translation machinery. Despite extensive biochemical investigations, a systems-level investigation of tRNAs’ functional roles in physiology, and genetic interactions among them, are lacking. We created a comprehensive tRNA deletion library in yeast and assessed the essentiality of each tRNA gene in multiple conditions. The majority of tRNA deletions show no appreciable fitness defect when such strains were grown on rich medium. More challenging environmental conditions however revealed a richer set of specific-tRNA phenotypic defects. Co-deletion of tRNA combinations revealed that tRNAs with essential function can be compensated by members of the same or different anti-codon families. Often times, we saw that identical tRNA gene copies contribute deferentially to fitness, suggesting that the genomic context of each gene can affect function. Genome-wide expression changes in response to tRNA deletions revealed two distinct responses. When a deleted tRNA belongs to a family which contain multiple genes with the same anti-codon, the affected cells responded by up-regulating the translation machinery; but upon deletion of singleton tRNAs, the cellular response resembled that of proteotoxic stress. Our tRNA deletion library is a unique resource that offers a way towards fully characterizing the tRNA pool and their important role in cell physiology.
Introduction

Messenger RNA translation is a central molecular process in any living cell and is among the most complicated and highly regulated of cellular processes [1,2]. The tRNA pool is a fundamental component in that process, serving as the physical link between the nucleotide sequence of mRNAs and the amino acid sequence of proteins. In the cycle of translation elongation, tRNA selection is considered the rate-limiting step [3], therefore tRNA availability is one of the major factors that govern translation-efficiency and accuracy of genes [4,5].

Previous studies have established that efficient translation can increase protein levels and provide a global fitness benefit by elevating the cellular concentrations of free ribosomes [6,7], while accurate translation benefits the cell by reducing the metabolic cost of mis-incorporation events [8].

The tRNA pool is composed of various tRNA isoacceptor families, each family carries a different anti-codon sequence that decodes the relevant codon by Watson-Crick base pairing, or codons with non-perfect base pairing of the third nucleotide by the wobble interaction. tRNA families are further classified to isotypes if they carry the same amino acid. In all eukaryotic genomes, each tRNA family can be encoded by a single or multiple gene copies [9,10]. It was previously shown for several organisms that the concentrations of various tRNA isoacceptors positively correlates with the tRNA family's gene copy-number [11,12]. These observations along with detailed analysis of the relationship between gene copy-number of tRNA families and codon-usage, established the notion that the multiplicity of tRNA genes in yeast is not functionally redundant. Such multiplicity might establish the correct balance between tRNA
concentrations and the codon usage in protein-coding genes [13], thus justifying the use of the tRNA gene copy-number as a proxy for actual tRNA amounts [12,14,15].

The transcription of tRNA genes is catalyzed by RNA polymerase III (pol III), promoted by highly conserved sequence elements located within the transcribed region [16]. A genome wide analysis of pol III occupancy in yeast revealed that virtually all tRNA genes are occupied by the pol III machinery [17–19], and are thus considered to be genuinely transcribed. This observation, combined with the fact that tRNA genes within a family are highly similar, led to the notion that all copies within a family contribute equally to the total expression level and hence to the tRNA pool.

Although tRNAs have been extensively studied, until very recently many of the studies were performed on individual genes at the biochemical level. Only in recent years systematic genome-wide approaches started to complement the biochemical approach. These studies reveal a much more complex picture in which pol III occupancy, a proxy for tRNA transcription, varies within families and between tissues [17,20–22]. Expression however does not equal function, and so far no systematic study has been carried out to decipher the specific contribution of each tRNA gene to the tRNA pool and to the cell’s fitness.

To study the role of individual tRNAs and the architecture of the entire tRNA pool, we created a comprehensive tRNA deletion library in the yeast *S. cerevisiae*. The library includes 204 deletions of nuclear-encoded tRNA genes out of the total 275 present in the yeast genome. In addition, we created double deletions of selected tRNA gene combinations and of specific tRNAs with a tRNA modifying enzyme. We developed a robotic method to screen and score various fitness
parameters for these deletion strains, and applied it across various growth
conditions. This systematic deletion library revealed an architecture of genetic
interactions that feature extensive backup-compensations within and between
tRNA families. Such compensation capacity endows the organism with
robustness to environmental changes and to genetic mutations. We found that
different copies within a tRNA family contribute differently to the organism's
fitness, revealing a higher level of complexity in the tRNA pool's architecture,
possibly at the regulatory level. Finally, we observed two distinct molecular
signatures that underlie the cellular response to changes in the tRNA pool. First,
the deletion of non-essential single-copy tRNA genes invoked proteotoxic stress
responses, indicating a connection between aberrant tRNA availability and
protein misfolding. Second, the deletion of representative tRNAs from multi-copy
families triggered milder responses by up-regulating genes that are involved in
the translation process. Together our results uncover the complex architecture of
the tRNA pool reviling a profound effect on cellular fitness and physiology.
Results

**Generation of a tRNA deletion library in S. cerevisiae**

To gain a better understanding of the functional role of individual tRNA genes and their contribution to the tRNA pool, we created a comprehensive tRNA deletion library in *S. cerevisiae*, where in each strain a single nuclear-encoded tRNA gene was deleted. This methodology is based on recombining a selective marker into the genome at the expense of the deleted gene, as was done in the creation of the yeast ORF deletion library [23] (Figure 1A). A particular challenge in targeting specific tRNA genes for deletion by such a method stems from the high degree of sequence similarity within tRNA families, which can share 100% sequence identity. Consequently, in order to create specific gene deletions, we relied on unique sequences that overlap or flank the tRNA genes (see Supplemental text S1). Our tRNA deletion library contained 204 deletions out of the 275 nuclear-encoded tRNA genes identified in *S. cerevisiae* (see Materials and Methods). These deletions covered all 20 amino acids and 40 of the 42 anti-codon families. The remaining 71 tRNA genes were not deleted due to their complex genomic surrounding, since such deletions might affect neighboring potential features in their genomic vicinities. The library also consisted of 50 strains that represent various combinations of tRNA deletions, and co-deletions of selected tRNAs with the *TRM9* gene which codes for an enzyme that post-transcriptionally modifies tRNA molecules.

Although the majority of tRNA families contain multiple gene copies, there are six single-copy tRNA families in the *S. cerevisiae* genome. Out of these singleton families, four (*tS(CGA), tR(CCG), tQ(CAG) tT(CGU)* were found to be essential in
our analysis, which confirms previous reports [24–26](see Supplemental text S1). The remaining two singleton families (tR(CCU), tL(GAG)) were identified as non-essential upon deletion. All the tRNA genes that belong to multi-copy families were non-essential upon deletion.

**Cells were robust to tRNA gene deletions in rich medium but reveal sensitivity in challenging conditions**

To assess the contribution of each tRNA gene to cellular growth, we attempted to accurately characterize the growth dynamics of each deletion strain by implementing a robotic method to screen and score growth phenotypes of all tRNA deletion strains in a given growth condition. This fitness measurement approach allowed us to differentiate between physiological effects of the deletion under different growth phases, unlike the competition approaches for fitness measurement [27] that typically integrated all growth phases. We characterized each deletion strain by two growth parameters: growth rate and growth yield, the latter is defined as the size of the population upon entering stationary phase (Figure 1B and Supplemental figure S1A).

We began the characterization of the tRNA deletion library by growing the strains in rich medium. Under this condition, 13% of the deletion strains demonstrated a phenotype in growth rate and 27% showed a growth yield phenotype (Figure 1C-D and Supplemental figure S1B). Most strains exhibited a notable phenotype only in one of the two parameters. Strains that showed altered phenotypes in both growth rate and yield were rare (Supplemental figure S1B). Overall, most tRNA deletion strains did not exhibit any altered growth phenotype in rich medium, indicating robustness to tRNA gene deletion. Seven
percent of the tRNA deletion strains resulted in growth improvement, suggesting that for some genes the cost of retaining them in the genome and/or expressing them may exceed their benefit in this condition. Similar observations were also made on a selection of protein-coding genes in this species [28]. Apart from the singletons whose deletion strains were often dead or exhibit impaired growth, we could not explain the observed growth phenotypes in growth rate or yield by either tRNA family size or amino acid identity (Supplemental figure S2). To further examine the phenotypes of the tRNA deletion strains, we calculated the correlation to the mRNA expression level of adjacent genes and found none (see, Supplemental table S1 figure S3 and Supplemental text S1).

Given that yeast cells are constantly exposed to varying environmental conditions, their tRNA repertoire should differentially accommodate growth in various environments. We next examine whether stressful conditions would retain the robustness observed in rich medium or reveal another set of condition-dependent growth phenotypes. We screened the deletion library under a diverse set of growth conditions including different metabolic challenges and stress-inducing reagents reported in previous studies [29–31]. The fact that the production of tRNA molecules is considered energetically costly [32] prompted us to explore the effect of carbon limitation, alternative carbon sources and minimal medium on tRNA essentiality.

Growing the tRNA deletion library under stressful conditions revealed condition-specific phenotypes (Figure 2A-D, Supplemental table S2). In all but one of the examined conditions (Dithiothreitol-DTT, a reducing agent that also inflicts a general protein-unfolding stress), robustness to tRNA gene deletion was maintained. In the DTT condition, the phenotypes were surprising: while
multiple tRNA deletions exhibited impaired growth rates, many also demonstrated growth rate improvements (Figure 2B, 2D). As in the rich medium condition, we could not explain the observed growth phenotypes by either the family size, or the amino acid identity in all of the examined stress conditions.

Extensive redundancy underlies robustness to tRNA gene deletion

Our observations of robustness to tRNA gene deletions in rich medium, as well as several stressful growth conditions, prompted us to further explore the genetic architecture conferring this phenotype. Given that most tRNA families contain multiple gene copies, we hypothesized that at least part of the observed robustness might be the outcome of compensation provided by the remaining genes in the family. In addition, due to wobble-interactions, robustness may also be the outcome of compensation between families of the same isotype. Focusing on rich medium conditions, we generated selected combinations of multiple tRNA deletions. To examine the first possibility we created deletions of entire two-member and three-member tRNA families. As shown in figure 3A such family deletions resulted in either lethality (indicating a loss of the family’s function), or viability with growth impairment (indicating a partial compensation of the family's function by other families).

We then turned to examine in more detail the interactions within these essential three-gene families by examining the growth of various double deletion strains. Contrary to the common notion that suggests little or no functional redundancy between tRNA gene copies [13], we observed that in each of these families any one family member can sustain normal or near-normal fitness (Figure 3A, Supplemental figure S4A-B and Supplemental table S3). Similar observations
were made for essential two-gene families upon one member’s deletion (Figure 3A). Such results can either imply that yeast cells carry more tRNA copies than are actually needed to sustain growth under optimal growth conditions, or that a responsive backup mechanism might be at work, one that provides compensation by increasing the transcription of the remaining copies, as was previously observed in protein-coding genes [33–35]. We thus decided to investigate the expression levels of certain tRNA families, using RT-qPCR (Figure S5). For each deletion, we compared the expression level of the remaining copies belonging to the designated family to that of a wild-type strain. We observed in most strains an expected reduction in expression of the respective family. These findings suggest that in these families, tRNA supply exceeds the demand under rich medium conditions (Figure S5A). However in some cases there were no such decreases in expression, there were even observable increases, demonstrating that a responsive backup mechanism may have been at work, inducing the expression of the remaining family members following deletion of a certain member (Figure S5B).

Next, we turned to examine the surprising cases in which the deletion of an entire tRNA gene family resulted in a viable strain. We reasoned that in these cases a different type of compensation, which is based on wobble interactions across iso-acceptor families, came into play. To decipher this compensation mechanism we focused on the genetic interactions involving the two non-essential singleton families, tL(GAG) and tR(CCU) (Figure 3A).

In the absence of tL(GAG), the members of the tL(UAG) family represent the sole tRNA that can decode the CUN Leucine codons, and might be a candidate for providing compensation upon deletion of tL(GAG) even though such decoding
does not match the classic wobble rules [36]. Co-deletion of tL(GAG) with one of the tL(UAG) gene copies resulted in growth aggravation and negative epistasis. Deletion of the tL(GAG) together with two copies of the tL(UAG) family was lethal despite the fact that one copy of tL(UAG) still remained in the genome, indicating that a single tL(UAG) gene was insufficient to compensate for the loss of tL(GAG) (Figure 3B). The genetic interaction between tL(UAG) and tL(GAG) appeared specific, since co-deleting one copy of the tL(UAG) family together with two additional tRNA genes (tL(CAA)G3 and tW(CCA)G1) did not generate observable epistasis in either case (Figure 3B). We thus concluded that the tL(UAG) family is partially redundant to the tL(GAG) family, yet such redundancy was not sufficient to completely compensate for the loss of tL(GAG).

Similarly, the viability of the tR(CCU) deletion strain could be due to compensation provided by the 11 copies of the tR(UCU) family. Indeed the wobble rules are consistent with this assumption, but such interaction was never functionally demonstrated. Formally, demonstrating that the tR(UCU) family can compensate for the loss of the singleton tR(CCU) would amount to co-deleting all 12 tRNA genes. Looking for simpler means, we decided on a more economic, albeit indirect way. We co-deleted the singleton tR(CCU) with the Trm9 enzyme, which is responsible for methylating the third anticodon position of tR(UCU) and tE(UUC) [37]. It was previously shown that such methylation is needed for supporting the wobble interaction between tR(UCU) tRNAs and the AGG codon (the cognate codon of the CCU anti-codon) [37]. The tR(CCU)–trm9 double deletion strain was viable, but exhibited an appreciable aggravation of growth yield (Figure 3A and 3C). Thus our results confirm that the methylated tR(UCU) family can partially compensate for the loss of tR(CCU). We attempted to define a
more general role for the Trm9 modification enzyme in modulating the compensation mechanism between tRNA families. To this end we created 10 additional double deletions of the enzyme along with each of 10 tRNA genes from two glutamic acid families, one that is modified by the enzyme and one that is reportedly not modified by the enzyme [38] (see Supplemental figure S6). No epistasis was detected between the enzyme and any of these 10 tRNAs and hence, the data cannot support or exclude a putative similar role of the enzyme beyond the tR(UCU) family.

We thus conclude that there are two mechanisms that can account for the observed robustness for tRNA deletions under favorable growth conditions. The first is redundancy within a family, and its efficiency appears to be independent of the number of remaining tRNA gene copies. The second is compensation between families, which operates via wobble interactions.

**Identical tRNA genes contributed differentially to cellular fitness**

We then asked whether all copies within a family contribute equally to the tRNA pool. It is often implicitly assumed that all tRNA copies contribute similarly to the cellular tRNA pool. However, comparison of the growth parameters of tRNA deletions from the same family revealed marked differences between seemingly identical family members. In particular, under rich medium, 21 out of the 32 deletions examined from multi-copy families showed growth yield differences spanning a broad range of at least 10% (Figure 4A). Such differences were also detected in the growth rate parameter (Supplemental figure S7A) although they were less pronounced. We thus focus on the growth yield parameter in all further analysis. The phenomenon of differential contribution to fitness by
different family members was further enhanced when we grew the deletion strains on more challenging conditions such as low glucose (Figure 4B and Supplemental figure S7B). To further investigate the genetic interactions between differentially contributing tRNA copies within a given family, we focused on the tR(UCU) family. The tR(UCU) family contains 11 identical copies in the genome, 5 of which were represented in our library. In rich medium, two copies (tR(UCU)E and tR(UCU)M2) showed appreciable reduction in growth yield (termed Major copies), while deletions of the other three copies (tR(UCU)M1, tR(UCU)G1 and tR(UCU)K) grew essentially as the wild-type (termed Minor copies). Introducing a plasmid with the appropriate tRNA gene copy complemented the growth of all deleted copies (Supplemental figure S7C). To further assert the separation between the Major and Minor copies, we examined various pair-wise deletion combinations of these members. All pairs that included at least one Major member exhibited growth impairment upon deletion, while pairs that consisted of only Minor copies demonstrated either a slight growth defect or none at all (Figure 4C). Further analysis of genetic interactions of these family members with either the TRM9 gene, or with the above mentioned tR(CCU) gene that belongs to a different Arginine family, revealed a similar effect (Figure 4C). These results indicate that the loss of different tR(UCU) genes in the same genetic background does not affect the phenotype equally, Major copies are more essential than Minor copies and as such are also more essential in providing compensation within the family. We next turned to examine whether the hierarchy of Major and Minor copies is preserved across various stress conditions (Figure 4D). Examining essentiality in
several conditions, we observed the same phenomenon in which Major copies demonstrated a stronger effect on growth compared to Minor copies in most stress conditions. We also noted that the Minor copies showed a diverse response ranging from slight growth improvement, wild-type level growth to observable growth impairment. A potential scenario may be one in which the Major copies always actively contribute to the pool, while the Minor copies might be recruited at times of need to maintain efficient translation. Thus, the loss of a Major copy could only be partially compensated by the remaining copies.

Following these observations, we turned to examine possible genetic elements that might promote the phenomenon of differential contribution. Since all family members have identical sequence, we hypothesized that differential contribution should be due to differences in the vicinity of tRNA genes. To demonstrate this notion we performed a complementation assay, introducing different tRNAs from the UCU family, along with 200bp of their flanking sequences, to the tR(UCU)M2 deletion strain. We observed different degrees of complementation. Given that different constructs differ only in the region flanking the tRNA gene, the variation in complementation capability can be attributed to the different sequences flanking the tRNA (Supplemental figure S7D). The effect of sequences that flank tRNA genes on their transcription was reported in multiple studies[39–42]. In one such study Giuliodori et al. [42] performed an analysis of conserved sequence elements upstream of S. cerevisiae tRNA genes. They identified four conserved sequence elements located at positions -53 (T-rich), -42(TATA-like), -30(T-rich) and -13 (pol III TSS) with respect to the first nucleotide of the mature tRNA. We used these results to examine the entire tRNA deletion library and checked whether tRNA deletions that exhibited or that did
not exhibit altered phenotype in rich medium revealed enrichment for any particular motif (Figure 4E). We found that deletions exhibiting phenotypes of growth impairment were significantly enriched for the presence of specific motifs. In particular, deleted strains that exhibited impairment in growth yield had an enrichment for the TATA-like motif at position -42. In addition, the TSS motif at position -13 was enriched in deletion strains that exhibited impairment in both growth rate and yield. To reinforce these observations, we ran the MEME motif search algorithm [43] to screen the upstream sequences of tRNA deleted strains exhibiting impaired growth yield for enriched motifs (see Materials and Methods). Two significant motifs were found that resemble those reported by Giuliodori et al. in both sequence and location (Figure 4F).

Together these results indicate that the contribution to the tRNA pool and cellular fitness of different copies of the same tRNA family are far from equal. We provide one possible explanation, which can account for the differential essentiality, implying that the sequences flanking tRNA genes play a role in determining their expression level.

**Physiological effects of tRNA gene deletions on protein folding**

As mentioned above, screening the tRNA deletion library in the presence of the reducing agent Dithiothreitol (DTT), a drug that exerts a proteotoxic stress in the cell, showed severe phenotypic defect in many deletion mutants (Figure 2A, B). Yet, many of the strains that demonstrated growth reduction in other conditions were less sensitive than wild-type to this drug (Figure 2C, D). These findings point towards a connection between tRNA functionality and the protein folding state in the cells. To further explore this connection, we turned to thoroughly
characterize a selection of tRNA deletions in the presence of various proteotoxic agents. We chose two deletion mutants that exhibited either impaired or wild-type growth under DTT, namely \((tR(UCU)M2 \text{ and } tH(GUG)G1)\), both members of multi-copy families designated the MC group. In addition to the two viable single gene deletions \((tR(CCU)J \text{ and } tL(GAG)G)\), the initiator methionine \((tM(CAU)C)\) also demonstrated improved growth; we thus designated these three strains the SC group.

The various strains were treated with either DTT, Azetidine 2 carboxylic acid (AZC)- a toxic analog of proline [44], or Tunicamycin- a drug used to induce the unfolded protein response (UPR) in the endoplasmic reticulum (ER) [45]. The growth of each strain was characterized under each proteotoxic agent applied at several concentrations. The strains in the MC group demonstrated either growth impairment or wild-type growth under all examined conditions. However, the deletions of single-copy tRNAs and to some extent the imitator methionine demonstrated reduced sensitivity to all three proteotoxic agents (Figure 5A-C). The differences in relative growth for all the examined strains were apparent even at low concentrations and were consistent upon increase in the concentrations of these proteotoxic agents (Figure 5A-C).

The fact that the tRNA deletion strains from the SC group are resistant to proteotoxic agents led us to hypothesize that deleting these genes might inflict intrinsic and chronic misfolding stress, even at the absence of the drug. This stress results in the activation of relevant cellular response that protects cells from the aggravating effect of extrinsic proteotoxic stress. Such an effect is reminiscent of the cross protection effect observed between environmental
stressors [46], yet here it is manifested between a genetic perturbation and an environmental stress.

To directly examine whether changes in the tRNA pool induce proteotoxic stress in these strains, we examined the state of the protein quality control machinery using the naturally unstructured human protein VHL as a proteotoxic stress reporter [47]. In this system, the VHL protein can be destined to one of two cellular localizations. If the cell experiences protein-folding stress, the heterologous protein VHL will aggregate in inclusions (or puncta) due to saturation of the protein quality control machinery. In contrast, under normal conditions, the quality control machinery is available to properly deal with this naturally unfolded heterologous protein, thus it remains soluble in the cytoplasm and no inclusions are formed. For each of the five deletion strains, we quantified the number of VHL inclusions (puncta) in populations of yeast cells. This analysis revealed that indeed the tRNA deletions in the SC group exhibited a significant increase in the number puncta containing cells relative to the wild-type (Figure 5D and 5F), indicating saturation of the quality control machinery caused by intrinsic proteotoxic stress. The MC group did not exhibit inherent proteotoxic stress; their puncta containing cells count resembled that of the wild-type.

The inherent chronic proteotoxic stress observed for the SC deletions might provide them with the capacity to respond better to an additional external proteotoxic stress. To further explore this possibility we examined the state of the protein quality control machinery upon extrinsic proteotoxic stress induced by treatment with AZC. Treating the wild-type cells with AZC resulted in a rapid accumulation of the VHL protein in stress foci, indicated by increase in the occurrence of multiple inclusions [48]. As anticipated, the behavior of the SC
group demonstrated a significant increase in the presence of a single punctum upon AZC treatment, however the appearance of stress foci (multi-puncta) was significantly lower compared to the wild-type and to the MC group (Figure 5E and 5G). As in the previous experiment, the deletions of the MC group responded in a similar manner to that of the wild-type, displaying increased number of stress foci.

These results thus indicate that the deletion of some tRNA genes induced an inherent proteotoxic stress in the cell, demonstrating a physiological role of proper tRNA supply in protein folding by an undetermined mechanism. Such physiological response renders these cells relatively less sensitive, compared to other tRNA deletion strains and the wild-type, from the otherwise harmful effect of proteotoxic drugs.

**Different molecular responses to deletions of tRNAs from single and multiple copy families**

To determine whether changes in the tRNA pool result in a distinct molecular signature, we examined the same set of tRNA deletions (SC and MC groups) using mRNA microarrays. For each strain, we measured genome-wide changes in mRNA levels compared to the wild-type, under rich growth conditions. The expression changes we observed were modest and demonstrated a correlation between the essentiality of the tRNA gene and the extent of changes in mRNA expression upon its loss. Hierarchical clustering of the strains according to similarity in expression changes (Fig 6A and supplemental figure S8), revealed that the strains could be divided into two groups recapitulating the division to the SC and MC groups. An example for this division can be found in the
pronounced effect observed for the COS8 gene. This gene was extremely up-regulated (about 16 fold) in the SC group while unchanged in the MC group (Figure 6B). These results suggest different molecular signatures for the two groups, which are also related to the proteotoxic stress response.

To determine the responses and the underlying molecular pathways that differentiate these two groups, we examined which KEGG pathways [49,50] differentiate between them. We used Gene Set Enrichment Analysis (GSEA), a computational software which determines whether a defined set of genes shows statistically significant differences between two biological states [51,52]. This analysis revealed a somewhat opposite signature between the two groups (Table 1 and supplemental figure S8). Pathways which are responsive to proteotoxic stress such as the Proteasome (FDR q-value <1E-5) and Protein processing in endoplasmic reticulum (FDR q-value 2E-3) were significantly induced in the SC group relative to the MC group. While in the MC groups, translation-related pathways such as Ribosome biogenesis (FDR q-value <1E-5) and Ribosome (FDR q-value 1E-4) were significantly induced compared to the SC group.

To further characterize these differences we focused on specific pathways. A more detailed examination of the expression changes observed for all the genes that constitute the proteasome complex revealed an up-regulation to various extents in response to deletion of tRNAs from the SC group. The MC group demonstrated no change and even a slight down-regulation of these genes (Figure 6C), a trend which was further verified using RT-qPCR (Figure 6D). These observations establish the notion that cells experience proteotoxic stress upon deletion of members of the SC group. A further indication of proteotoxic stress in these deletion strains is the up regulation of COS8. The exact biological
function of this gene is still unclear, it was however found to interacts with *IRE1*, which is a hallmark regulator of the unfolding stress response [53].

An interesting distinction between the groups was also observed in the pathway consisting of the RNA polymerase machinery. Expression of genes that belong to this pathway were up-regulated only in the MC group (Table 1). Separating the RNA polymerase genes into modules corresponding to the different polymerases, revealed an interesting pattern. While the genes that encode RNA Pol II subunits did not change in any of the tRNA deletion strains (Supplemental figure S9), the genes encoding RNA Pol III machinery (the polymerase responsible for tRNA gene transcription) demonstrated up-regulation in the MC group and no change or even down regulation in the SC group (Figure 6E). These results were further verified by RT-qPCR (Figure 6F). Up-regulation of the pol III machinery for the MC group may suggest that in some MC deletion strains, the transcription of the remaining tRNA genes could increase, thus providing a possible molecular mechanism for backup compensation within families. Such response to deletions of tRNAs from the MC group could indicate the presence of a negative feedback loop, allowing the cell to respond to changes in the tRNA pool in the attempt to regain steady state levels.
**Discussion**

In this study, we investigated the genetic architecture of the tRNA pool and its effect on cellular fitness using a comprehensive tRNA deletion library. We found extensive dispensability of many tRNA genes, especially under optimal growth conditions. Such lack of essentiality has been studied in protein-coding genes, and is often interpreted to reveal a role for partially redundant genes and pathways providing backup compensation for the deleted gene [33,34,54–56]. Similar design principles are displayed in the architecture of tRNA genes, which exhibited significant gene redundancy and compensation (either partial or complete) among family members. An additional reason for apparent lack of essentiality of genes is the limited set of examined environmental challenges, and it was indeed shown for protein-coding genes that challenging gene deletion libraries to less favorable conditions exposes more essentiality [57,58]. We showed that a similar situation holds for tRNA genes. We found condition-specific functional roles for tRNAs, demonstrating increased demand for certain tRNA genes under certain defined conditions. This implies that the compensation within tRNA families changes across conditions. Such changes in the essentiality of tRNA genes can imply that the tRNA pool is dynamic and changes across conditions to accommodate cellular needs, as was recently suggested [59].

Further, we have discovered interesting architecture within families, which questions the prior notion that all tRNA gene copies contribute equally to the pool. Previous work has shown that Pol III transcription machinery displays different occupancy levels at various copies of the same tRNAs in the genome [21,22,60]. However, the potential phenotypic consequences of such
transcriptional differences have not been previously explored. We report that the flanking sequences around each tRNA gene contains motifs that are predictive of the deletion phenotypic consequences, potentially affecting pol III transcription machinery.

We further speculate that some tRNA genes, i.e. the Major copies, might be active across all conditions and with only partial functional redundancy, thus their loss cannot be fully compensated. Minor copies on the other hand are either not transcribed or have a modest contribution to the tRNA pool, with complete functional redundancy by other copies, thus their loss can be fully compensated. Such architecture could provide the cell with means to respond in a dynamic manner to changes in the environment, by transcribing varying portions of the members of each tRNA family depending on demand. As such, differential contribution within tRNA families exposed an additional novel mean to regulate the tRNA pool and as a consequence to regulate the translation process.

An interesting finding was that changes in the tRNA pool elicit molecular changes in the cells even when no severe phenotype is detected. Our results demonstrated two distinct molecular signatures which can be attributed to the family architecture and the severity of the changes in the pool. Upon deletion of the two viable single copy tRNAs, and also upon deletion of one of the initiator tRNA methionine copies, the cell exhibited a response reminiscent of a proteotoxic stress. We were able to identify such a stress in these mutant cells. Although the exact mechanisms by which changes in the tRNA pool induces proteotoxic stress remains to be determined, we hypothesize that the elimination or reduction in these tRNAs may lead to events of amino acid misincorporation, ribosome frame-shifting or stalled protein synthesis.
terminations. Such events would have a clear impact on the protein quality control machinery of the cell by titrating chaperons to deal with misfolded or misassembled proteins. Translation errors such as incorrect tRNA selection and incorrect tRNA aminoacylation have been shown to induce proteotoxic stress in yeast [61,62]. Given that cells exploit chaperon availability as a sensing mechanism to induce a stress response [63,64], translation errors may lead to the onset such a response. On the other hand, deletions of tRNAs from multi-copy families results in milder effects on the tRNA pool due to the extensive redundancy or backup-compensation, and they indeed elicit a different cellular response from the one invoked upon deletion of single-member families. In the response to deletion of members from multi-gene families, the pol III transcription machinery seems to be up regulated. Such up-regulation would bring about induced transcription of tRNAs, this would act as a feedback mechanism to bring the tRNA pool closer to its normal state [65]. At least in one case (Supplemental figure S5) our results suggest the existence of such responsive backup among tRNA genes from the same family. Yet, a clearer relationship between changes in the tRNA pool, pol III activation, and tRNA transcription is still lacking. Regardless of the actual mechanism that determines the exact cellular response to tRNA deletions, the fact that such a response wiring exists may be beneficial for maintaining cellular robustness upon environmental changes and mutations.

This work provides for the first time a systemic tool to study the functional role of individual tRNA genes. Using this deletion library, we discovered a much more complex picture than was previously known. We anticipate that a high throughput mapping of all genetic interactions between pairs of tRNA genes (as
done for protein-coding genes) [66,67] would reveal the full genetic network. In addition, it might reexamine and potentially refine the wobble interaction rules from a genetic, rather than the traditional biochemical/structural perspective. The design principles defined in this study, consisting of massive gene redundancy as well as differential contribution of gene copies may provide cellular plasticity and allow the tRNA pool to accommodate various growth conditions and developmental planes. Deciphering the effects of tRNA variations as is found in some diseases such as cancer [68] and Huntington [69] can provide possible routes for future treatment. We provide this novel set of minimalist genetic perturbations in the translation machinery as a resource to the yeast community towards further characterization of this highly complex process as well as additional cellular processes.
Materials and Methods

Creation of tRNA deletion library

The complete tRNA pool of *S. cerevisiae* was obtained from the tRNA genomic database [70], where 286 tRNA genes are annotated. 13 tRNA genes are encoded by the mitochondrial genome and the remaining are nuclear-encoded. Here we focused on the nuclear-encoded tRNAs. Two tRNA genes that are annotated in this database as not determined, belong to the *ts(GCU)* family. Thus, the *ts(GCU)* family contains two additional members, *ts(GCU)L* and *ts(GCU)D*, both verified by PCR, bringing the total number of nuclear encoded tRNA genes to 275.

Deletion strains were constructed using a PCR-based gene deletion [71,72], in the genetic background of the Y5565 strain (*MATα, can1Δ::MFA1pr-HIS3, mfa1Δ::MFA1pr-LEU2, lyp1Δ, ura3Δ0, leu2Δ0*). The *S. cerevisiae* strain S288C reference genome sequence R57-1-1 downloaded from the Saccharomyces Genome Database was used for primer design. Each deletion construct contained 45 bp flanking or overlapping a tRNA sequence for specific recombination event, a unique barcode and the HPH antibiotics ‘cassette’, conferring resistance to the antibiotic hygromycin B, [73]. PCR products were transformed into yeast cells and single colonies were verified by PCR. Three colonies from each strain were used to verify phenotypes in growth analysis. A wild-type strain in which the same antibiotic marker was integrated 200bp upstream of the *tL(CAA)L3* locus was created as a control and was used in all analyses as wild-type. A complete list of all plasmids, yeast strains and PCR fragments can be found in Supplemental text S1 and Supplemental table S5.
Measurements of growth using OD reads

Strains were grown for two days at 30°C in YPD (1% yeast extract, 2% peptone, 2% glucose), diluted (1:50) into the appropriate medium in U-bottom 96-well plates and grown at 30°C (using TECAN Freedom EVO robot). The OD of the population in each plate was monitored every 30 minutes using a spectrophotometer at 600 nm (INFINITE200-TECAN). Each plate contained a wild-type strain to which the growth parameters of the deletions strains were normalized. The OD reads served for growth analysis and extraction of growth parameters. At least 3 biological repeats and 36 technical repeats were performed for each strain in each condition. Complete description of analysis and normalization procedures are provided in the Supplemental text S1.

Yeast growth conditions

Library strains were screened in the following growth conditions: YPD, SCD (0.67% Bacto-yeast nitrogen base w/o amino acids 2% glucose supplemented with amino acids), YP supplemented with 0.025% glucose, YP supplemented with 1% galactose, YPD supplemented with 0.5M NaCl, SCD supplemented with 1.5mM DTT. Growth measurements were also performed on YPD supplemented with increasing concentrations of the proteotoxic agents DTT, AZC and Tunicamycin.

Motif Analysis

A sequence motif analysis was performed using the MEME online software [43]. The motif search was done on the upstream sequence of tRNA genes which exhibited a yield impairment phenotype in rich medium upon deletion (42 genes) versus the upstream sequence of tRNA genes which exhibited a
phenotype in no more than two out of the six conditions (99 genes). To apply location constrains on the motifs, the MEME analysis was done in windows of size 9bp, looking for motifs of 4-8bp in length.

Analysis of protein quality control using VHL-CHFP reporter

Wild-type and tRNA deletion strains harboring the pGAL-VHL-mCherry (CHFP) fusion were grown overnight on SCD+2% raffinose, diluted into SCD+2% galactose and grown at 30°C for 6 hours. Cells were visualized using an Olympus IX71 microscope controlled by Delta Vision SoftWoRx 3.5.1 software, with X60 oil lens. Images were captured by a Photometrics Coolsnap HQ camera with excitation at 555/28 nm and emission at 617/73 nm (mCherry). Images were scored using the ImageJ Image Processing and Analysis software. The percentage of cells harboring VHL-CHFP foci was determined by counting at least 500 cells for each strain in three biological repetitions. Protein un-folding stress was induced with AZC at a concentration of 2.5 mM AZC (Sigma) following induction with galactose.

Analysis of genome wide expression changes

Cultures were grown in YPD medium at 30°C to a cell concentration of 1.5*10^7 cells/ml. Cells were then harvested, frozen in liquid nitrogen, and RNA was extracted using MasterPure™ (EPICENTER Biotechnologies). The quality of the RNA was assessed using the BIOANALYZER 2100 platform (AGILENT); samples were then processed and hybridized to Affymetrix yeast 2.0 microarrays using the Affymetrix GeneChip system according to manufacturer's instructions. The background adjustment was done using the Robust Multi-array Average (RMA) procedure followed by quintile normalization.
For each strain, the fold change in expression for all genes was calculated by comparing the wild-type measurement in the same batch and averaged over two biological repeats.

**Microarray analysis**

The cluster tree is based on the correlation between the mRNA fold change of the different strains. For the clustering we used the top 50% of the sorted genes based on the gene variance across the strains.

**Microarray data access**

The data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO) under accession number GSE47050. A list of the measured fold changes for all genes in each strain can be found in Supplemental table S4.

**RT-qPCR measurements**

Cultures were grown in YPD medium at 30°C to a cell concentration of $1 \times 10^7$ cells/ml. RNA was extracted using MasterPure™ (EPICENTER Biotechnologies), and used as a template for quantitative RT-PCR using light cycler 480 SYBR I master (Roche)(LightCycler 480 system) according to the manufacture instructions. A list of the primers can be found in Supplemental table S6.
Acknowledgments:

We thank all the members of the Pilpel lab for many fruitful discussions. We thank Daniel Kaganovich for providing the VHL-CHFP yeast plasmids. We thank Sebastian Leidel and Refael Ackermann for critical reading of the manuscript. We thank Ilya Soifer for assistance with the Robotic system. We thank Ofer Moldovsky, Yifat Cohen and Tslil Ast for their assistance with the VHL-CHFP system and the microscope analysis. We thank Nir fluman for assistance with the protein measurements.
References


Figure 1. Creation and analysis of tRNA deletion library.

(A) Schematic representation of the deletion process. 204 different tRNA strains were created using homologous recombination. In each strain, a different tRNA gene was replaced by a hygromycin B resistance marker.

(B) Schematic representation of growth measurements, analysis, and scoring. For each strain, relative-growth-rate and relative-growth-yield are calculated in relation to the wild-type strain. These parameters are then projected on a distribution of the wild-type growth parameters. Sigma (σ) is calculated according to the formula and denotes the number of standard deviations from the mean of the wild-type (see also Supplemental figure S1A). The color in the histogram are areas were: σ<-3 (blue), -3<σ<-2 (cyan), 2<σ<3 (yellow) and 3<σ (red). The same color code is used to define phenotypes in the pie charts (C and D).

(C-D) Distribution of phenotypes for the tRNA deletion library in rich medium, according to two growth parameters: relative growth yield (C) relative growth rate (D). Deletion strains were assigned to categories according to their σ values. Any absolute σ value larger than 2 was considered as non-normal phenotype, where negative sigma denotes impairment (worse than the wild-type) and positive sigma denotes improvement (better than the wild-type). Any absolute σ value larger than 3 was considered as a strong phenotype. Thus, highly impaired for σ<-3, impaired for -2>σ>-3, improved for 2<σ<3, and highly improved for σ>3, see also Supplemental figure S1B.
**Figure 2. Screening the tRNA deletion library across various growth conditions.**

(A) Percent of strains exhibiting a growth yield phenotype in various conditions. The color indicates the type of phenotype: impaired (blue) or improved (red).

(B) Percent of strains exhibiting a growth rate phenotype in various conditions.

(C-D) The σ values measured for both the growth yield (C) and the growth rate (D) for all deletion strains across six conditions. The color bar indicates the σ values, red denoting improvement and blue impairment. Each row denotes a tRNA deletion strain and each column denotes different growth condition. Strains are ordered on the y-axis according to amino acids (denoted by letter) and further separated into families (denoted by lines within the amino-acid box). Black rows denote lethal strains. Gray rows indicate strains for which the respective value was not measured.

**Figure 3. Extensive redundancy underlies robustness to tRNA gene deletion.**

(A) Schematic representation of the genetic interactions within and between tRNA families. Families are denoted by dark grey circles and grouped (black dashed line) according to their tRNA copy number. Each family is denoted by its anti-codon and amino-acid. A protein-coding gene i.e. TRM9 is denoted by a grey box. Each filled circle indicates a tRNA deletion strain. The lines connecting the deletion strains denote a co-deletion of these genes (a multi-tRNAs deletion strain). The color of the filled circles and lines denote the severity of the growth phenotype for the respective strain: blue for normal growth, purple for impaired growth (worse than wild-type) and red for lethality. (B) Epistasis values for multi-tRNAs deletion strains which contain the deletion of tL(GAG) and either:
one \( tL(UAG) \) gene, two \( tL(UAG) \) genes, \( tL(CAA) \) (which is a tRNA of different Leucine family), and \( tW(CCA) \) (which is a non-Leucine tRNA) as controls.

(C) Epistasis values for multi-tRNAs deletion strains which contain the deletion of \( trm9 \) with: the singleton \( tR(CC) \), and \( tR(ACG) \) which is a tRNA of different Arginine family and \( tW(CCA) \) which is a non-Arginine tRNA as controls. In both (B) and (C) epistasis values of the relative growth yield and growth rate are indicated in grey and green respectively. Data is presented as mean of 3 biological repetitions +/- SEM.

Figure 4. Differential contribution of identical tRNA gene copies.

(A-B) Relative growth yield values of the tRNA deletion library strains in rich medium (A) and low glucose (B), sorted by anti-codon and amino-acid identity along the x-axis. Each dot along the vertical lines denotes the value (data are represented as mean of 3 biological repetitions +/- SEM) of a deletion strain of different tRNA gene of the respective family. The horizontal lines mark two standard deviations around the mean of the wild-type. Dots above or below these lines are considered non-normal phenotypes (see also Supplemental figure S7).

(C) Relative growth yield values (data are presented as mean of 3 biological repetitions +/- SEM) of various double deletion combinations consisting of: five \( tR(UCU) \) family members, \( tR(CCU) \) and \( trm9 \) deletion strains as indicated on the x-axis, along with the five members of the \( tR(UCU) \) family each denoted by a different shape and color in the legend. (D) Relative growth yield of the five \( tR(UCU) \) members across different growth conditions, indicated on the x-axis.
(E) Enrichment of conserved elements in tRNA genes divided according to phenotype observed in rich media for each growth parameter. Each column in the matrix denotes a conserved element as defined by [42]. Color bar indicates the \(-\log_{10}\) of the hypergeometric p-value. (F) \log_{10} E-value found by the MEME software for the most significant motif in a 9bp window starting from the position indicated by the x-axis. The LOGOs of the two significant motifs are displayed below, next to a number indicating its position. Position 0 is the first position of the mature tRNA.

**Figure 5. Changes in the tRNA pool affect protein folding**

(A-C) Relative growth rate (compare to wild-type) of the following five deletion strain: \(tL(GAG)G\) (blue), \(tR(CCU)J\) (red), \(tiM(CAU)C\) (green), \(tH(GUG)G\) (magenta) and \(tR(UCU)M2\) (cyan). Strains were grown in media supplemented with increasing concentrations of the following proteotoxic agent: AZC (A) Tunicamycin (B) DTT (C).

(D) Percentage of cells that contain puncta in the populations of the above strains.

(E) Percentage of cells that contain puncta in the populations of the above strains following treatment with 2.5mM AZC. Data are presented as mean of 3 biological repetitions +/- SEM, in each repetition 500 cells were counted. (*) \(P < 0.001\) by Students t-test.

(F-G) Images of representative fields for the wild-type and \(tR(CCU)J\) deletion strain, without treatment (F) and following treatment with 2.5mM AZC (G).
Figure 6. Molecular response to changes in the tRNA pool.

(A) Dendrogram created by clustering changes in gene expression for five representative deletion strains, for more information see Materials and Methods.

(B) Fold change of the COS8 (YHL048W) mRNA levels in each of the five deletion strains as measured by microarrays. (C) The fold change distribution of mRNA levels as measured by microarrays, of genes composing the Proteasome pathway by the KEGG database [49], for each of the listed tRNA deletion strains.

(D) mRNA Fold change of 6 representative genes from the proteasome pathway measured by RT-qPCR. Presented values are the mean of 3 biological repetitions +/- SEM. The strain colors are as in (C). If the mRNA fold change in a specific strain is significantly different from 0 (t-test) it is marked with:* (p<0.05) or ** (p<0.005).

(E) The fold change distribution of mRNA levels as measured by microarrays, of genes composing the Pol III RNA Polymerase machinery module by the KEGG database, for each tRNA deletion strain. (F) mRNA Fold change of 6 representative genes from the Pol III KEGG module measured by RT-qPCR. Presented values are the mean of 3 biological repetitions +/- SEM. The strain colors are as in figure (C). If the mRNA fold change in a specific strain is significantly different from 0 (t-test) it is marked with:* (p<0.05) or ** (p<0.005).

In all the sub-figures (C,D,E,F) values are plotted for the same five deletion strains: *tL(GAG)* (blue), *tR(CCU)* (red), *tiM(CAU)* (green), *tH(GUG)*1 (magenta) and *tR(UCU)*2 (cyan).
### Table 1. KEGG pathways differentiating between tRNA deletion sets

<table>
<thead>
<tr>
<th>Higher in SC than in MC</th>
<th>Higher in MC than in SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteasome (&lt;1E-5)</td>
<td>Ribosome biogenesis in eukaryotes (&lt;1E-5)</td>
</tr>
<tr>
<td>Oxidative phosphorylation (&lt;1E-5)</td>
<td>RNA polymerase (&lt;1E-5)</td>
</tr>
<tr>
<td>Endocytosis(2E-3)</td>
<td>Phenylalanine, tyrosine and tryptophan biosynthesis (&lt;1E-5)</td>
</tr>
<tr>
<td>SNARE interactions in vesicular transport (2E-3)</td>
<td>Pyrimidine metabolism (5E-5)</td>
</tr>
<tr>
<td>Protein processing in endoplasmic reticulum (2E-3)</td>
<td>Ribosome (1E-4)</td>
</tr>
<tr>
<td>Starch and sucrose metabolism (2E-3)</td>
<td>Lysine biosynthesis (1E-4)</td>
</tr>
<tr>
<td>Citrate cycle (TCA cycle) (0.01)</td>
<td>Histidine metabolism (4E-4)</td>
</tr>
<tr>
<td>Meiosis (0.01)</td>
<td>Cysteine and methionine metabolism (4E-4)</td>
</tr>
<tr>
<td>Homologous recombination (0.02)</td>
<td>Riboflavin metabolism (5E-3)</td>
</tr>
<tr>
<td>Mismatch Repair (0.02)</td>
<td>Arginine and proline metabolism (8E-3)</td>
</tr>
<tr>
<td>Cell cycle (0.02)</td>
<td>Valine, leucine and isoleucine biosynthesis (0.01)</td>
</tr>
<tr>
<td>MAPK signaling pathway - yeast (0.02)</td>
<td>Purine metabolism (0.03)</td>
</tr>
<tr>
<td>Fructose and mannose metabolism (0.02)</td>
<td>Sulfur metabolism (0.03)</td>
</tr>
<tr>
<td>Nitrogen Metabolism (0.02)</td>
<td>Tyrosine Metabolism (0.03)</td>
</tr>
<tr>
<td>Phagosome (0.03)</td>
<td>Folate biosynthesis (0.04)</td>
</tr>
</tbody>
</table>

KEGG pathways [49] for which changes in genes expression are significantly different between the two groups of tRNA deletion strains: MC (multi-copy) group ($\Delta tH(GUG)G1$ and $\Delta tR(UCU)M2$) vs. SC (single-copy) group ($\Delta tL(GAG)G$, $\Delta tR(CCU)$, $\Delta tiM(CAU)C$) calculated with GSEA [51,52]. In the first column are pathways, which are higher in SC vs. MC and vice versa in the second column. The values are corrected for multiple hypothesis and the FDR q-values are indicated next to each pathway.
Supplemental Figures

**Figure S1. Growth measurements parameters.**

(A) Schematic growth curve of Optical Density (OD) vs. time. The red dots indicate the time points from which the growth rate (1) and growth yield (2) parameters are extracted. (B) Dot plot for all strains in the library grown in YPD. Each strain is represented by a blue dot, showing its sigma growth rate vs. its sigma growth yield values. The Pearson correlation coefficient is -0.019 indicating there is no correlation between the two parameters p-val 0.794.

**Figure S2. Phenotypes cannot be explained by family size and amino-acid identity.**

Sigma growth parameters for the tRNA library grown in rich medium are plotted in boxes sorted by either family size or amino-acid identity. For each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles. Sigma growth yield by family size (A) sigma growth rate by family size (B) sigma growth yield by amino-acid (C) sigma growth rate by amino-acid (D). Apart from the singletons whose deletion strains are often lethal or impaired, we could not explain the observed growth phenotypes, in either growth rate or yield, by either the size of the family, or the amino acid identity.

**Figure S3. tRNA deletion phenotype are not correlated to the expression of nearby genes.**

(A-B) the average expression level of the genes located upstream and downstream to the tRNA gene that was deleted in each strain vs. the sigma
growth yield (A) or the sigma growth rate (B). Relative growth parameters of tR(CCU)J deletion (black), tR(CCU)J deletion containing a centromeric plasmid harboring the tR(CCU)J gene (gray) and a strain deleted for the YJR055W gene which is the protein-coding gene located downstream of tR(CCU)J (white). As can be seen only the tR(CCU)J deletion strain exhibits growth rate impairment while the two other strains do not.

Figure S4. Single tRNA genes can sustain wild-type growth upon deletion of multiple members in three gene families.

(A-B) Relative growth rate (red) and growth yield (blue) values of double deletion combinations containing members of the tG(UCC) family (A) and the tS(UGA) family (B). In each experiment the mean of 3 biological repetitions is presented +/- SEM. Two σ around the mean of the wild-type are indicated by red and blue lines around 1 (wild-type value).

Figure S5. Compensation within some tRNA families is due to plasticity of the pool and transcriptional changes of the remaining copies.

RT-qPCR measurement of the RNA levels of the tS(UGA) family (A) and tL(UAG) family (B) upon deletion of various members of the family. Results are reported in terms of log2 fold change of the expression level in each of the indicated deletion strain compared to the wild-type. In both (A) and (B) the * indicates cases in which the fold change was significantly different from zero (t-test, p-value <0.05).
Examine a more general role for Trm9 in modulating the compensation between tRNA families we chose the second tRNA family that is modified by Trm9, tE(UUC), and in addition we examined the tE(UCU) family. Together these two families decode in a split codon box, in a similar manner to the Arginine UCU and CCU families. We created 10 double deletions, each consisting of the enzyme along with one of the tRNA genes of the two glutamic acid families and analyzed their interactions by epistasis.

Epistasis values for co-deletion strains which contain the deletion of trm9 with: the deletion of the two members of tE(CUC) family, and eight members of the tE(UUC) family. Epistasis values of the relative growth yield and growth rate are indicated in grey and green respectively. Data is presented as mean +/- SEM of 3 independent experiments.

Identical tRNA genes contribute differentially to the tRNA pool.

(A-B) Growth rate values of the tRNA deletion library in rich medium (A) and low glucose (B) sorted by families and amino-acid identity. The horizontal lines denote two standard deviations around the mean of the wild-type in that condition. Dots above or below these lines are considered phenotypes. (C) Relative growth yield values (data of 3 biological repetitions +/- SEM is presented) of five tR(UCU) deletion strains (Grey) and the corresponding complementation strains (White). Each complementation strain carries the deleted tRNA gene on a centromeric plasmid. The values are relative to the wild-type. In the complementation experiment, the wild-type harbors an empty plasmid. (D) Relative growth yield values of strain deleted for tR(UCU)M2 gene
(a major copy of the \textit{tR(UCU)} family- marked as $\Delta M2$ ), and $\Delta M2$ strains containing different centromeric plasmids. Each centromeric plasmid carries the \textit{tR(UCU)} tRNA flanked from each side by 200bp sequence identical to a the different members of the \textit{tR(UCU)} family.

\textbf{Figure S8. Expression changes of tRNA deletions.}

Expression changes for the five deletion strains. Each row indicates a gene and each column is a tRNA deletion strain. The genes and strains are sorted according to the clustering results (see Materials and Methods). The Color bar indicates the log2 fold change. The groups of genes enriched for relative pathways are indicated on the right (locations were found by looking at the highest hypergeometric enrichments for varying window sizes).

\textbf{Figure S9. Fold change of the Pol II pathway.}

(A) The fold change distribution of mRNA levels as measured by microarrays, of genes composing the Pol II RNA Polymerase machinery by the KEGG database for each of the listed tRNA deletion strains. (B) mRNA Fold change of 3 representative genes from the Pol II pathway measured by RT-qPCR. Presented values are the mean of 3 biological repetitions +/- SEM. The strain colors are as in figure (A). If the mRNA fold change in a specific strain is significantly different from 0 (t-test) it is marked with:* (p<0.05) or ** (p<0.005).

In both sub-figures (A, B) values are plotted for the same five deletion strains: \textit{tL(GAG)G} (blue), \textit{tR(CCU)J} (red), \textit{tiM(CAU)C} (green), \textit{tH(GUG)G1} (magenta) and \textit{tR(UCU)M2} (cyan).
Table S1. Correlation between tRNA phenotype and expression of nearby genes.

Table S2. List of all tRNA deletion strains in the library and their respective phenotypes across conditions.

Table S3. List of double deletion strains and their phenotypes.

Table S4. Microarray Fold change measurements for selected tRNA deletion strains.

Table S5. List of primers used to create the tRNA deletion strains.

Table S6. List of primers used for RT-qPCR experiments.

Text S1. Supplementary methods and note.
Figure 1

Click here to download Figure: F1.eps
Figure 2

Click here to download Figure: F2.eps
Figure 5
Click here to download Figure: F5.eps
Supplemental Text

Materials and Methods

Yeast strains and plasmids

- *S. cerevisiae* strain Y5565 (MATα can1Δ::MFA1pr-HIS3 mfa1Δ::MFα1pr-LEU2 lyp1Δ ura3Δ0 leu2Δ0) was used as the genetic background for the tRNA deletion library strains and the double deletion strains.
- *S. cerevisiae* strain BY4743 (MATa/α his3Δ1/his3Δ1 leu2Δ0/leu2Δ0 LYS2/lys2Δ0 met15Δ0/MET15 ura3Δ0/ura3Δ0) was used for analysis of tRNA essentiality in tetrad analysis.
- pAG32 DEL-MARKER-SET (EOUROSCARF) [1] was used to amplify the Hygromycin B resistance marker.
- pAG25 DEL-MARKER-SET (EOUROSCARF) [1] was used to amplify the Nourseothricin resistance marker.
- pFA6a-kanMX6 [2] was used to amplify the Kanamycin resistance marker.
- pRS316- Centromeric plasmid [3] was used for tRNA complementation assays.

Creation of a tRNA deletion library

In similarity to the yeast gene deletion library [8], the tRNA deletion library was constructed using homologous recombination, replacing each tRNA gene with the HPH antibiotics cassette. The DNA template for the recombination event was created using two sequential PCR reactions. The first PCR reaction created the basic template for the recombination containing the antibiotic marker and a 23bp sequence homologous to the desired tRNA. The second PCR reaction served to lengthen this template so it contains 45bp homologues to the desired tRNA deletion. The first PCR reaction used primers containing a genomic sequence that flanks either the 5' or 3' end of the tRNA (directly proximal and distal to the start and end of the gene respectively), 18 and 17bp of sequence common to all gene disruptions as was used in the yeast deletion collection, a 20 base pairs unique sequence (the 'molecular bar-code' TAG) and 22 base pairs of sequence, homologous to the HPH gene. The template for this PCR reaction was the pAG32 plasmid, encoding the hygromycin B phosphotransferase conferring resistance to the antibiotic hygromycin B. The second PCR
reaction used two tRNA gene-specific 45 bp primers, to extend the tRNA specific homology of the first PCR product to a total of 45 bp. The product of the second PCR served for yeast transformation. For a full list of primers see Supplemental table S5.

 Verification of tRNA deletion strains
For each tRNA deletion strain, five different colonies were selected, and verified by two PCR reactions. The junctions of the antibiotics cassette were amplified using two primer combinations. Each primer combination contained a primer homologous to sequence within antibiotics cassette of either the promoter or terminator of the marker gene, and a tRNA specific primer, homologous to sequence upstream or downstream of the tRNA deletion. Only colonies from which the correct products were amplified served for further analysis.

 Verification of lethality of tRNA deletions by tetrad analysis
To confirm the essentiality of double and triple tRNA deletions, diploid strains were created by mating two deletion strains, either single or double deletions. These heterozygous diploids were sporulated in SC medium (1% potassium acetate) for three days at 25°C. The resulting tetrads were then dissected using a Micromanipulator (Singer). Spores were then allowed to germinate at 30°C on YPD plates and scored on plates containing the appropriate antibiotics. The essentiality of single tRNA genes was confirmed by deletion of the gene from a diploid strain, BY4347, in a similar manner to the construction of the tRNA deletion library. The resulting colonies were verified, grown on rich medium, sporulated and scored. At least 100 tetrads were scored for each such deletion.

 Construction of complementation plasmids
In order to reintroduce into the cell a tRNA deleted gene we created a set of “complementation plasmids”. Complementation plasmids were constructed for multiple tRNA genes using gene specific primers homologous to 200 bp up and down stream of the tRNA gene which was deleted in the library. Genomic DNA of Y5565 was used to amplify the desired tRNA gene, followed by cloning into the pRS316 centromeric plasmid.

 Creation of multi-deletion tRNA strains
Double deletion of tRNA genes was constructed in a similar way to the creation of the tRNA deletion library strains. A strain from the tRNA deletion library served as the genetic
background for the deletion of an additional gene by means of homologues recombination. The sequential PCR was repeated for the second tRNA gene deletion using the same primers used for the creation of the tRNA deletion library. The templates for the first PCR round were either the pAG25 plasmid containing the nourseothricin N-acetyl-transferase that confers resistance to the antibiotics nourseothricin, or the pFA6a-kanMX6, encoding the kanamycin gene that confers resistance to the antibiotics G418. Triple deletions were created by mating double deletions with a single deletion.

**Growth measurements and normalization**
The measures obtained for each strain in the library were compared to wild-type values that were retrieved separately for each condition. The wild-type measures were done in three repeats and were then summarized by the mean and standard deviation values to generate normalized growth rate and normalized growth yield for each condition. This procedure allowed normalizing for potential variation between days, incubator slots, and locations of wells within the 96-well plate. These measures were obtained from three biological repeats containing wild-type cells in all plate positions, for all slots in the incubator. Each mutant was characterized by growth rate and growth yield. We consider a mutant as showing a phenotype if it deviated by more than two standard deviations from the wild-type mean in that condition in either growth rate or growth yield.

**Double deletion epistasis calculations**
The epistasis between any two deletion strains was calculated according to the non-scaled measure of epistatic interactions [9] using the following equation:

\[
\varepsilon = W_{XY} - W_X W_Y
\]

In which \(W_X\) and \(W_Y\) represent the growth values (separating growth rate and growth yield) of the single deletions and \(W_{XY}\) represents the growth values of the corresponding double deletion. The results of the calculation indicate the nature of the epistasis, \(\varepsilon = 0\) indicates no epistasis, \(\varepsilon < 0\) aggravation and \(\varepsilon > 0\) buffering.
No appreciable correlation between tRNA expression and the expression levels of nearby protein-coding genes

It has been demonstrated in the past (mainly for specific cases) that the expression of tRNA genes may affect the expression levels of nearby coding genes [6–8]. In order to explore a possible relation between the essentially of tRNA genes, as revealed here upon their deletion, and the expression of nearby genes we measured the correlation between the tRNA deletion phenotype and the expression level (as measured for the wild-type strain in our microarray data) of their upstream and downstream nearby genes. We found no correlation between the severity of growth yield or growth rate phenotypes and nearby gene expression level, indicating that high (or low) expression of a nearby gene is not a predictor of fitness change upon tRNA deletion (see Supplemental figure S3A,B and Supplemental table S1). In addition, we did not find any correlation between the distance to the closest genomic feature and the phenotype of the deletion strain.

Furthermore focusing on five tRNA deletions for which we measured genome wide expression changes, we examined the effect of the deletion on the expression levels of the adjacent genes. As can be seen in the following table:

<table>
<thead>
<tr>
<th>Deletion Strain</th>
<th>Upstream</th>
<th>Downstream</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gene Name</td>
<td>Distance bp</td>
</tr>
<tr>
<td>∆iM(CAU)C</td>
<td>YCR018C</td>
<td>-1017</td>
</tr>
<tr>
<td>∆R(CCU)J</td>
<td>YJR054W</td>
<td>-1006</td>
</tr>
<tr>
<td>∆tR(UCU)M2</td>
<td>YML071C</td>
<td>-253</td>
</tr>
<tr>
<td>∆tH(GUG)G1</td>
<td>YGL205W</td>
<td>-221</td>
</tr>
<tr>
<td>∆L(GAG)G</td>
<td>YGR106C</td>
<td>-889</td>
</tr>
</tbody>
</table>

Out of 10 protein-coding genes examined, only in two cases (YJR055W and YGL205W) we detect an appreciable change in the expression level (although in both cases it was less than two fold). This indicates that in most cases tRNAs deletion do not affect their surroundings.

The general lack of coordination between tRNA genes and nearby protein-coding genes is in agreement with the work of Conesa et al., [9], that have measured the expression level of all protein-coding genes in several Pol III mutants as well as few tRNA deletion strains. This analysis detected only a modest correlation between altered expression of Pol II-transcribed
genes and their proximity to class III genes.

For the deletion that demonstrated the highest fold change in the expression of a nearby gene \( tR(CCU)J \) and a severe phenotype, we verified that the fitness reduction observed in this strain is indeed due to the deletion of the tRNA gene using several different methods. First, using complementation assay we re-introduced the \( tR(CCU)J \) gene on a centromeric plasmid into the background of the deleted strain. This resulted in a complete abolishment of the fitness defect (Supplemental figure S3C). In addition, we measured the growth of a strain deleted for the \( YJR055W \) gene (the protein-coding gene located downstream \( tR(CCU)J \)) and found that unlike \( tR(CCU)J \) this strain does not exhibit a growth rate phenotype (Supplemental figure S3C). Taken together these two analysis demonstrate that the growth defect observed in cells deleted for \( tR(CCU)J \) is due to the lack of function of the CCU tRNA.
Supplemental references


Figure S1

A

B

Growth Rate
Growth Yield
corr: -0.018
p-value: 0.798
tV(AAC)L
tR(CCU)J
tW(CCA)M
tE(UUC)K
tG(GCC)J1
tL(UAG)J
Figure S2
Figure S3

A

B

C

Mean gene expression (log2)

Mean gene expression (log2)

σGrowth rate

σGrowth rate

σGrowth yield

σGrowth yield

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Relative Growth

Relative Growth

Growth rate

Growth rate

Growth yield

Growth yield

ΔtR(CCU)J

ΔtR(CCU)J

p_tR(CCU)J

p_tR(CCU)J

ΔYJR055W

ΔYJR055W

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Correlation: corr = -0.037

p-value: 0.63

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Correlation: corr = -0.037

p-value: 0.63

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Correlation: corr = -0.037

p-value: 0.63

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Correlation: corr = -0.037

p-value: 0.63

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Correlation: corr = -0.037

p-value: 0.63

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Correlation: corr = -0.037

p-value: 0.63

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Correlation: corr = -0.037

p-value: 0.63

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Correlation: corr = -0.037

p-value: 0.63

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Correlation: corr = -0.037

p-value: 0.63

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Correlation: corr = -0.037

p-value: 0.63

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Correlation: corr = -0.037

p-value: 0.63

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Correlation: corr = -0.037

p-value: 0.63

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Correlation: corr = -0.037

p-value: 0.63

Correlation: corr = 0.032

p-value: 0.68

Correlation: corr = 0.037

p-value: 0.63

Correlation: corr = -0.037

p-value: 0.63

Correlation: corr = 0.032
Figure S4

Panel A

- tG(UCC)D
- tG(UCC)D/O
- tG(UCC)N
- tG(UCC)N/D
- tG(UCC)N/O
- tG(UCC)O

Relative Growth

Growth Yield

Growth Rate

Panel B

- tS(UGA)E
- tS(UGA)E/P
- tS(UGA)I
- tS(UGA)I/E
- tS(UGA)I/P
- tS(UGA)P
Figure S5

A

B
Figure S7

A

B

Relative Yield

\[ \Delta M2 \]
\[ \Delta M2p_tR(UCU)E \]
\[ \Delta M2p_tR(UCU)K \]
\[ \Delta M2p_tR(UCU)M1 \]
\[ \Delta M2p_tR(UCU)M2 \]

Relative Yield

\[ \text{YPD} \]
\[ \text{0.5YPD} \]

Relative Growth

tRNA families

UGC
AGC
GCA
GUC
UUC
CUC
GAA
UCC
GCC
CCC
GUG
AAU
UUU
CUU
UAG
UAA
CAA
CAU
GUU
UGG
AGG
UUG
UCU
ACG
GCU
UGA
AGA
AGU
UAC
AAC
CCA
GUA

YPD

C

D

Relative Yield

\[ \text{deletion} \]
\[ \text{complementation} \]

Relative Yield

\[ \Delta M2 \]
\[ \Delta M2p_tR(UCU)E \]
\[ \Delta M2p_tR(UCU)K \]
\[ \Delta M2p_tR(UCU)M1 \]
\[ \Delta M2p_tR(UCU)M2 \]
Table S1. Correlation between tRNA phenotype and expression of nearby genes.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Upstream Gene</th>
<th>Downstream Gene</th>
<th>Average (up &amp; down genes)</th>
<th>Minimal distance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>yield phenotype</strong></td>
<td>r=0.06</td>
<td>r=-0.02</td>
<td>r=0.03</td>
<td>r=0.06</td>
</tr>
<tr>
<td></td>
<td>p-value:0.41</td>
<td>p-value:0.78</td>
<td>p-value:0.68</td>
<td>p-value:0.36</td>
</tr>
<tr>
<td><strong>absolute yield phenotype</strong></td>
<td>r=-0.03</td>
<td>r=0.02</td>
<td>r=0.01</td>
<td>r=0.07</td>
</tr>
<tr>
<td>(no matter whether it is</td>
<td>p-value:0.71</td>
<td>p-value:0.80</td>
<td>p-value:0.95</td>
<td>p-value:0.37</td>
</tr>
<tr>
<td>impairment or improvement)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>growth phenotype</strong></td>
<td>r=-0.01</td>
<td>r=-0.05</td>
<td>r=-0.04</td>
<td>r=0.04</td>
</tr>
<tr>
<td></td>
<td>p-value:0.92</td>
<td>p-value:0.50</td>
<td>p-value:0.63</td>
<td>p-value:0.61</td>
</tr>
<tr>
<td><strong>absolute growth phenotype</strong></td>
<td>r=-0.02</td>
<td>r=-0.04</td>
<td>r=-0.04</td>
<td>r=0.12</td>
</tr>
<tr>
<td>(no matter whether it is</td>
<td>p-value:0.78</td>
<td>p-value:0.60</td>
<td>p-value:0.61</td>
<td>p-value:0.08</td>
</tr>
<tr>
<td>impairment or improvement)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>any absolute phenotype</strong></td>
<td>r=-0.02</td>
<td>r=-0.01</td>
<td>r=0.00</td>
<td>r=0.08</td>
</tr>
<tr>
<td>(growth rate/growth yield</td>
<td>p-value:0.77</td>
<td>p-value:0.86</td>
<td>p-value:0.99</td>
<td>p-value:0.24</td>
</tr>
<tr>
<td>impairment/improvement)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The table shows the correlation (r) and the p-value between the deletion strain phenotypes and the expression of nearby genes. Each row shows the correlation between the sigma value of the growth yield/growth rate in rich medium (YPD) and the expression of the nearby gene/s taken from the microarray wild-type measurements (columns 1-3) and the distance to the closest genomic feature (column 4).