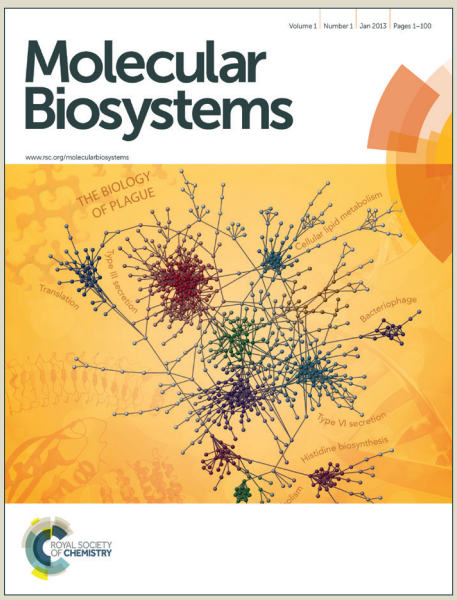
# Molecular BioSystems

Accepted Manuscript





ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

# Most Associations between Transcript Features and Gene Expression are Monotonic

Gilad Shaham[1] and Tamir Tuller[1,2]

[1]Department of Biomedical Engineering, the Engineering Faculty, Tel Aviv University, Israel. [2]The Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv.

**TT:** tamirtul@post.tau.ac.il

## Abstract

**Dozens of previous studies in the field have dealt with the relations between transcript features and their expression. Indeed, understanding the way gene expression is encoded in transcripts should not only contribute to disciplines, such as functional genomics and molecular evolution, but also to biotechnology and human health. Previous studies in the field mainly aimed at predicting protein levels of genes based on their transcript features. Most of the models employed in this context assume that the effect of each transcript feature on gene expression is monotonic.**

**In the current study we aim to understand, for the first time, if indeed the relations between transcript features (*i.e.*, the UTRs and ORF) and measurements related to the different stages of gene expression are monotonic. To this end, we analyze 5,432 transcript features and gene expression measurements (mRNA levels, ribosomal densities, protein levels, etc.) of 4,367 *S. cerevisiae* genes. We use the Maximal Information Coefficient (MIC) in order to identify potential relations that are not necessarily linear or monotonic.**

**Our analyses demonstrate that the relation between most transcript features and the examined gene expression measurements is monotonic (only up to 1%-5% of the variables, with significance levels of 0.001, are non-monotonic); in addition, in the cases of deviation from monotonicity the relation/deviation is very weak.**

**These results should help in guiding the development of computational gene expression modeling and engineering, and improve the understating of this process. Furthermore, the relatively simple relations between a transcript's nucleotide composition and its expression should contribute towards better understanding of transcript evolution at the molecular level.**

## Introduction

The association between various features of the transcript and its expression levels has been the topic of dozens of studies in recent years[1–7]. Earlier studies were mainly based on the codon composition of the coding sequence (ORF), and demonstrated that simple features based on the codon usage of the ORF highly correlate with the expression levels of the corresponding transcript [8–10]. Later studies exploited a much larger number of transcript features for predicting gene expression (mainly at the protein level)[1–7,11].

Among others, these features include GC content and folding in different parts of the transcript, length of the UTRs and ORFs, and more. Some examples of previous studies in the field include Lithwick and Margalit who measured several transcript features association to protein abundance in prokaryotes[12], and Kawaguchi and Bailey-Serres which analyzed the relation between transcript features and ribosomal loading in *A. thaliana*[13]. Ghaemmaghami *et al.* performed large scale measurements of protein levels in *S. cerevisiae* and showed that mRNA levels or codon usage measurements exhibit relatively limited correlation with protein levels[14]. Nie *et al*. examined several initiation, elongation and termination related features and found that the mRNA–protein correlation was affected the most by the features at elongation stages[15]. Transcript features have also been used for predicting protein levels and measurements of gene expression; for example, Tuller *et al.* used large scale data to obtain a predictor of translation efficiency in *S. cerevisiae*[3]; Vogel *et al*. used around 200 transcript features to explain 67% of protein abundance in human cell lines[1]; more recently, Zur and Tuller used 5,432 features to predict several expression levels variables[11].

All the previous studies in the field employ models that are based on a monotonic association assumption (*e.g.*, correlations and regressors) between transcript features and gene expression. However, the association of two variables may be very significant and yet non-monotonic. Some examples of non-monotonic associations include convex, concave, having multiple maxima, mutually exclusive, a combination of several associations, or even too complex to describe. If such associations between transcript features and gene expression exist, previous (and future approaches) focusing on monotonic relations may miss important transcript features. Thus, a fundamental open question in the field is related to the nature (monotonic or non-monotonic) of the relation between transcript features and their expression levels.

In the current study, we aimed to provide an initial answer to this question. To this end, we employ a recent novel statistical approach named the Maximal Information coefficient (MIC)[16]. MIC is a rank order statistic that discovers associations between variable pairs *without any prior knowledge regarding the type of association*. Previous studies already used MIC for various purposes, including biological research[17–20]. Unlike regular correlation, which is limited to specific types of associations (such as linear or monotonic), MIC is not built for specific association types and can detect data associations much more complex than simple correlation. Therefore, MIC fits well with our main goal. This approach has been employed on a dataset of 5,432 transcriptional features of *S. cerevisiae* endogenous genes[11] .We analyzed their association with various gene expression measurements in this organism: Protein Abundance (PA)[14,21,22], Ribosomal Density (RD)[23,24], mRNA levels[25], and Proteins per mRNA molecule (PPR).

## Results

### Analyzing large scale expression data using MIC

Briefly, MIC is based on the mutual information of binned values of an analyzed pair of variables[16]. The mutual information of a pair of variables ($x$ and $y$) measures the statistical dependency of the variables by computing the average over $\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$[26]; where $p(x,y)$ corresponds to the common distribution of the two variables, and $p(x)$ and $p(y)$ correspond to the marginal/independent distributions of the variables $x$ and $y$ respectively. If the two variables are independent then $p(x,y) = p(x)p(y)$; and therefore

$\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$ would be zero. Conversely, if there is high dependence between the variables we would expect this value to be high.

Practically, to generate distributions ($p(x), p(y)$, and $p(x,y)$) for the two analyzed variables the scatter plot is divided to a grid, and a value $p(x,y)$ is computed for each cell in the grid (see illustration in **Figure 1** A.)

This is essentially part of the calculation of mutual information and therefore the strategy in MIC is to calculate the mutual information of all the possible grids, normalize the values to be between 0 and 1, and return the maximum normalized value that was found. Low MIC values suggest low dependence between the variables and high values suggest high dependence. MIC should be able to detect any type of relation between pairs of variables[16].

A central question in this study is related to the monotonicity of the relation between transcript features and gene expression. To this end, we study the deviation between **MIC** and Spearman correlation **($\rho^2$)**, by subtracting the result of Spearman correlation from the MIC result ( $MIC - \rho^2$**).** This is expected to be close to zero for monotonic relations (illustration in Figure 1B; more explanations in the Methods section and in Reshef *et al.*[16])

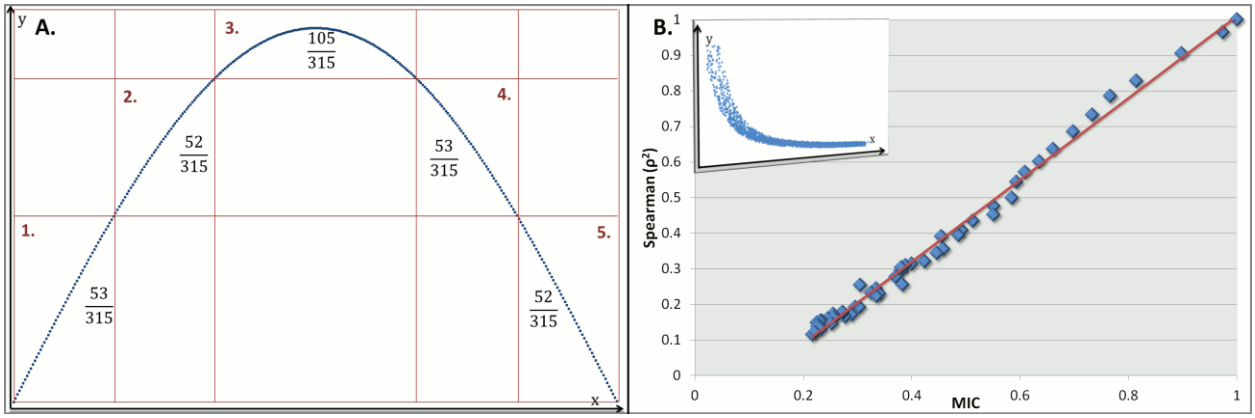



**Figure 1** – **A.** An example of a simple association between 2 variables with 315 data points and calculated probability using a grid. In this case the y-axis is divided into 3 (equal) groups of data points while the x-axis is divided into 5 (non-equal) parts giving us a total of 15 bins. In each bin, we count the number of points that lie within it. Thus, for example, in bin 1, the probability to see a data point is $\mathbf{53/315 = 0.1683}$. Using this method one can calculate the resulting mutual information for every grid. MIC searches for the grid that maximizes the mutual information for a given variable pair. **B.** An example of a **monotonic** association with its MIC and Spearman ($\boldsymbol{\rho^2}$) for 50 different levels of (uniform) random noise. Both MIC and Spearman ($\boldsymbol{\rho^2}$) can be used to detect this association type and output similar scores. The red line is a linear regression between the MIC scores and Spearman ($\boldsymbol{\rho^2}$) scores with $\boldsymbol{R^2 = 0.9921}$ and a slope close to one (1.15 in this case). Note that for totally random data Spearman would be zero, but MIC will not; thus, in this study we verify whether MIC is significantly higher than random by performing a permutation test (see Methods for more details). The inset includes the monotonic association we used to perform the analysis for this graph.

In practice, in order to detect relations between transcript-features and gene expression measurements that are not monotonic we performed the following steps (more details in the Methods section): First, for each expression data and feature pair we calculated its MIC (detects any statistical relation) and Spearman correlation (detects monotonic relations). We performed permutation tests (details in the Methods section) to estimate: 1) which MIC results are significant; 2) if the difference between MIC and Spearman is

significant (*i.e.*, significant deviation from monotonicity). Based on these tests we discovered two possible relations (monotonic and non-monotonic):

1. If the difference between MIC and Spearman $\rho^2$ is significantly low (for example, in at least 95% of the permutations the difference is equal or higher), and the Spearman p-value related to the association is significant, we say that the relationship of the transcript feature and expression is *significant monotonic*.
2. If the MIC score is significant and the difference between MIC and Spearman is significant (for example, in less than 5% of the permutations the difference is equal or higher), we say that the relationship of the transcript feature and expression is *significant non-monotonic*. Spearman correlation, in this case, is not enough to fully describe the association of the two variables.

We also considered stricter thresholds than the example above (*e.g.*, 99% and 1%; or 99.9% and 0.1%).

It is important to emphasize that according to our definition the meaning of a 'non-monotonic' relation is that the association *cannot be fully explained* via a monotonic (Spearman) relation. However, it does not mean that it cannot have a *monotonic component* and thus a significant Spearman correlation. In addition, we would like to mention that a relation can have low spearman correlation but still be significant if the number of points is high (the p-value is a function of the correlation but also the number of points).

As mentioned in the introduction, we analyzed 5,432 transcriptional features of *S. cerevisiae* endogenous genes and various gene expression measurements (PA, RD, mRNA, PPR); these data were downloaded from[11].

See **Figure 2** for an illustration of the process described above and a few examples of the results MIC and Spearman return for monotonic and non-monotonic functions.

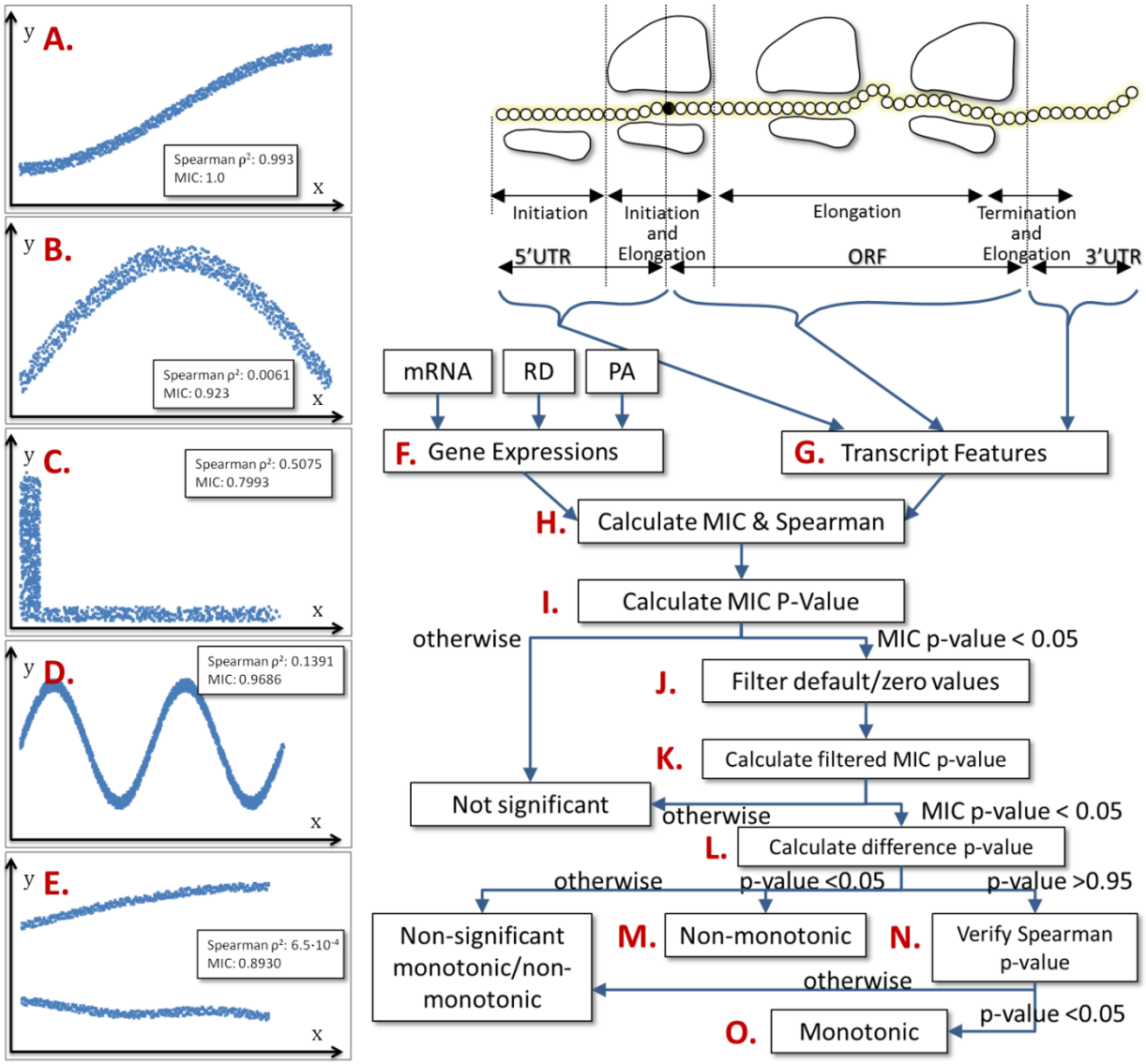



**Figure 2** – A-E: **Examples of associations with some random noise and their corresponding Spearman and MIC scores**. **A.** Monotonic association, Spearman is 0.993 and MIC is 1.0. **B.** Non-monotonic association, Spearman is 0.0061 while MIC is 0.923. MIC can detect many different types of non-monotonic associations whereas Spearman is more limited to monotonic associations. **C.** Mutually exclusive relationship, MIC is 0.7993, Spearman is 0.5075. **D.** Sinus (2 cycles), MIC is 0.9686, Spearman is 0.1391. E. Combination of two monotonic relations. MIC is 0.8930, Spearman is $\mathbf{6.5 \cdot 10^{-4}}$ **F.-O.: General framework of the analysis: F.** Large scale gene expression measurements are collected and normalized; beyond mRNA, RD and PA, we also calculate PPR by dividing PA by mRNA. **G.** Transcription (UTRs and ORFs) features are obtained for each gene. **H.** We calculate Spearman and MIC for each expression and feature pair. **I.** Empirical p-values based on permutation tests have been performed to estimate MIC significance. **J.** If the p-value is significant, then we filter default or zero values where applicable in order to be able to compare MIC and Spearman for monotonicity. **K.** We calculate the new MIC empirical p-value to verify result is still significant **L.** If the MIC is significant, then we calculate the p-value of the difference of $\boldsymbol{MIC - \rho^2}$ and classify each feature and expression pair to monotonic or non-monotonic (*i.e.*, whether the difference between MIC and Spearman is significant; details in the main text). **M.** If the p-value is below 0.05, then the relation is classified as *non-monotonic* since the significant relationship can be explained by MIC but not by Spearman. **N.** If the p-value is above 0.95, we verify the Spearman p-value is significant. **O.** If the p-value is significant then we say the feature is *monotonic* since the relationship can be equally explained by both MIC and Spearman. In all other cases we cannot say with high significance whether the relation is monotonic or non-monotonic.

### The features with strongest associations to expression levels measurements

At the first stage, we compared each feature to the four different measurements of expression levels: protein abundance (PA), mRNA levels, ribosomal density (RD) and proteins per mRNA molecule (PPR), using both MIC and Spearman correlation. The list of features and explanations about them appear in the Methods section; some of the features can be divided to large subsets (*e.g.*, the frequency of all codon pairs; see also Tables S1-S5).

**Figure 3** includes the features per gene expression measurement with the corresponding MIC and Spearman values. Features of interest (with emphasis on those with highest significance) are labeled in the graph. The definitions of all the features appear in the Methods section. This figure shows that the strongest associations have relatively high $\rho^2$ Spearman correlation scores. Hence, in case one would like to predict expression levels using transcript features, most of the attention could be focused on the features with high $\rho^2$ Spearman. The figure also shows that the features with non-monotonic relations tend to have low MIC levels (*i.e.*, to have weak relations) close to the non-significant relations. Furthermore, the regression line indicates there is a significant linear relation between MIC scores and Spearman ($\rho^2$) scores both when considering only the monotonic significant relations and when considering all the MIC significant relations. See more details in the next subsection.

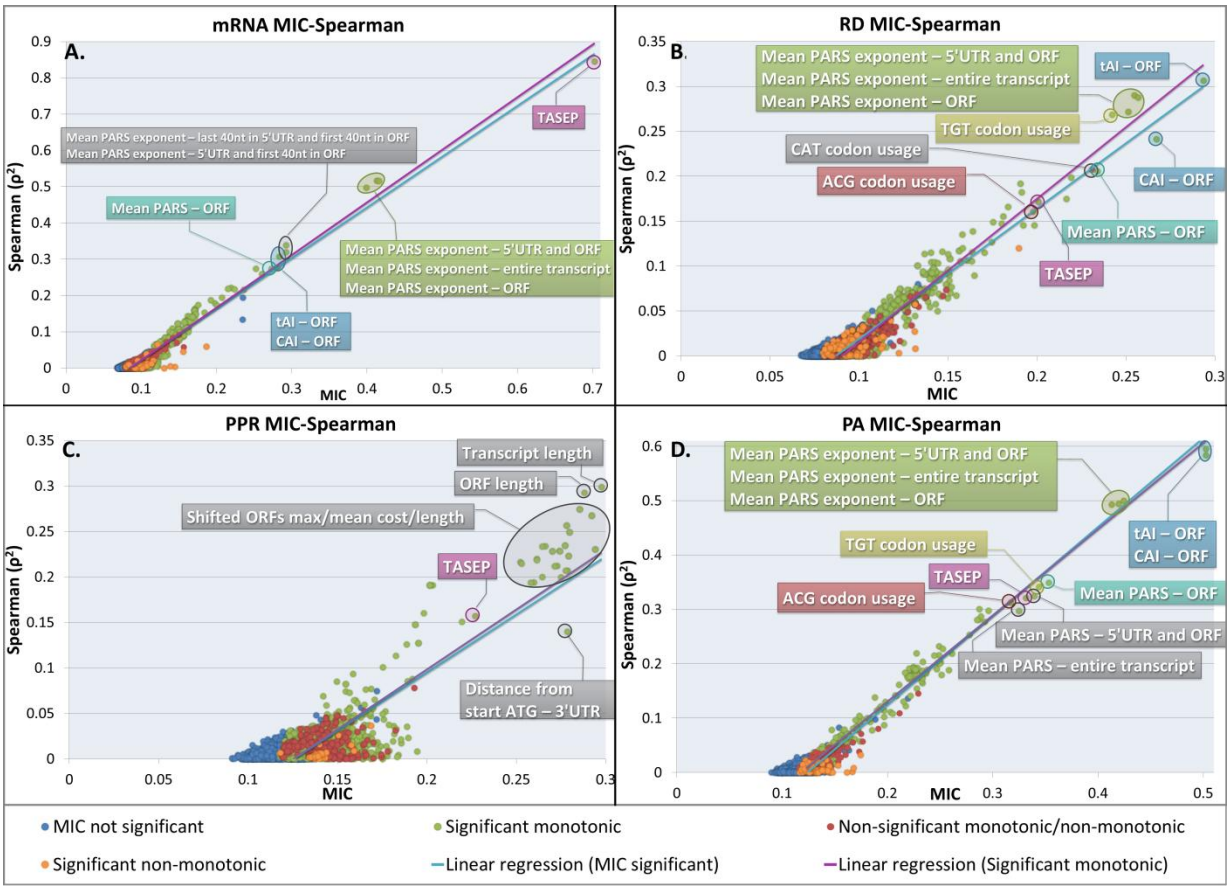



**Figure 3** – Comparison between MIC and Spearman $(\boldsymbol{\rho^2})$ per feature. Blue indicates non-significant MIC (empirical p-value > 0.05). For significant MIC, orange indicates $\boldsymbol{MIC - \rho^2}$ is significantly high (a non-monotonic relation) and green indicates that the difference is not significant (some overlap causes some points not to be shown in the graph). The purple line is linear regression for the significant monotonic associations and the light blue line shows the linear regression for all significant MIC features. Features with top MIC scores and comparison to Spearman correlation are also presented (Ordering was done by

selecting the lower p-value and, in case it's the same, by the Z-score). Results are presented for **A.** mRNA, **B.** RD, **C.** PPR, and **D.** PA.

Comparison of the MIC rankings of features for mRNA, RD, PA, and PPR using Spearman correlation is available in **Table 1**. As can be seen, there is significant correlation among the MIC ranking of features for mRNA, RD, and PA. On the other hand, the MIC ranking of features based on PPR is less correlative with the MIC ranking based on mRNA, RD, and PA The results suggest that the effect of transcript features on protein per mRNA (or the relation between transcript features and PPR) is relatively different than their effect on mRNA levels and ribosomal densities.

| | mRNA | PA | RD | PPR |
|---|---|---|---|---|
| **mRNA** | --- | | | |
| **PA** | 0.2855 ($< 4.9 \cdot 10^{-324}$) | --- | | |
| **RD** | 0.5604 ($< 4.9 \cdot 10^{-324}$) | 0.2664 ($< 4.9 \cdot 10^{-324}$) | --- | |
| **PPR** | 0.1570 ($1.7 \cdot 10^{-21}$) | 0.0082 (0.6203) | 0.1480 ($3 \cdot 10^{-19}$) | --- |

**Table 1** – Spearman correlation ($\boldsymbol{\rho^2}$) between the rankings of features per expression level, the p-value is provided in parenthesis. The ranking was calculated by first sorting according to the MIC p-value and if equal the Z-score is used. As evident, mRNA, RD and PA have stronger similarity in ranking, while PPR MIC ranking has weaker correlation or non-significant correlation with the other rankings.

Since the features with top MIC scores have high spearman correlation, most (but not all) of them have been reported in previous studies[1,3–5,7–9,11,12,27]. This result supports the previous conclusions and also the MIC method. In most gene expression databases, the tRNA adaptation index (tAI)[8] and Codon Adaptation Index (CAI)[9] rank high; these features depict the association of codon usage bias with the expression levels. The strong relation between codon usage bias and expression levels has been reported in many previous papers[3,12,25,28]; as was suggested in previous studies, it can be a result of causal/direct relation (codon usage bias improves various aspects of gene expression), non-causal relation (the relation between codon usage bias and expression is not direct and or with opposite direction: expression levels contribute to higher codon usage bias and not vice versa), or a superposition of both explanations[5,29].

Various features that are based on Parallel Analysis of RNA Structure (PARS)[30], which are related to the experimental measurements of the strength of the folding of the mRNA sequence in various regions (Methods), are also in the top ranks; the strong relation between PARS and expression levels has been reported in the past[27]. Predictions of the translation rate by a model based on the Totally Asymmetric Simple Exclusion Process (TASEP) of ribosomal movement (see details in the Methods section) has a very high score in the case of mRNA for both MIC and Spearman; this may be explained by global/indirect selection for translation in highly expressed genes to improve ribosomal allocation[31]; TASEP is also evident in RD and PA. PPR gives different results, it is calculated by dividing PA by mRNA which reduces the score of features that have similar correlation for PA and mRNA, and emphasizes features that have strong association with PA (a super position of translation and protein degradation), but weak association with mRNA (a super position of transcription and mRNA degradation).

Specific top ranked features in the case of mRNA are as follows (**Figure 3 A.**) – TASEP has the strongest association followed by mean PARS exponent on: 1. 5'UTR and ORF, 2. ORF, 3. Entire transcript; tAI and CAI are next in rank. Subsequently, mean PARS exponent of the last 40nt in the 5'UTR and the first 40nt in the ORF as well as mean PARS exponent of the 5'UTR and the first 40nt in the ORF, although their MIC and Spearman scores are somewhat higher than some of the previously mentioned features.

In the case of RD (**Figure 3 B.**), tAI has the strongest association followed by CAI. The next features are: the frequency of the codon TGT, the frequency of the codon CAT, the TASEP, and the mean PARS score on various parts of the transcript (Methods). The next feature is the frequency of the codon ACG.

In the case of PPR (**Figure 3 C.**), the highest rank is the entire transcript length followed by the mean metabolic cost for frame shifted ORF in the main ORF (sORF); the metabolic cost ('cost' in Figure 3) of the main ORF, short ORF that appears in the UTR (uORF), or sORF is the sum of the metabolic/energetic cost for biosynthesis of all the amino acids encoded in these ORFs/uORFs/sORFs in *S. cerevisiae* (see

details in the Methods section). There are overall 32 features related to the mean and maximum length and metabolic cost of the sORF that are ranked high; finally, aside from the transcript length, the ORF length is also highly ranked.

In the case of PA (**Figure 3 D.**), both tAI and CAI present the strongest association with similar scores followed by features related to the PARS score for various parts of the transcript. We then have a diverse group of features with strong association which includes: the frequency (codon usage) of codon TGT, the mean PARS of ORF, TASEP, the mean PARS of 5'UTR and ORF, the frequency of codon ACG, and the mean PARS of the entire transcript.

### Most MIC significant features have monotonic relations with expression levels

At the next step, we compared the MIC value and spearman correlation of the features (Figure 3). Since MIC performs an exhaustive search its value is not absolute zero for pairs of variables with no association. However, by calculating p-values based on a permutation test (Methods) we can find all the associations that are significant. **Figure 3** includes the features per gene expression measurement, marked as monotonic, non-monotonic and those that did not have significant association with the gene expression measurement (see explanations in the previous section and the Methods section). This figure shows that the strongest associations are monotonic; the non-monotonic features have lower statistical significance and lower scores.

We then performed linear regression for monotonic features with significant MIC between their MIC score and their corresponding Spearman correlation $(\boldsymbol{\rho^2})$. The results are presented in **Figure 3** and in **Table 2.** Very similar results were obtained when considering all features with significant MIC ($R^2 > 0.71$ in all cases).

| | mRNA | RD | PPR | PA |
|---|---|---|---|---|
| **Total features** | 4773 | 4733 | 3646 | 3681 |
| **Significant MIC** | 1891 | 1033 | 1178 | 404 |
| **Significant non-monotonic** | 48 | 115 | 32 | 69 |
| **Significant monotonic** | 1289 | 434 | 951 | 195 |
| **Significant monotonic linear regression ($R^2$)** | 0.9391 | 0.9301 | 0.755 | 0.9807 |
| **Significant MIC linear regression ($R^2$)** | 0.9175 | 0.9004 | 0.7196 | 0.9678 |
| **Significant MIC that pass FDR (q=0.05)** | 1537 | 645 | 840 | 227 |
| **Significant non-monotonic that pass FDR (q=0.05)** | 8 | 9 | 0 | 7 |
| **Significant monotonic that pass FDR (q=0.05)** | 1051 | 313 | 687 | 168 |
| **Z score-based Significant MIC that pass FDR (q=0.05)** | 1657 | 758 | 939 | 269 |
| **Z score-based Significant non- monotonic that pass FDR (q=0.05)** | 8 | 17 | 1 | 12 |
| **Z score Significant monotonic that pass FDR (q=0.05)** | 1067 | 405 | 741 | 170 |

**Table 2** – MIC and Spearman comparison for empirical p-value < 0.05. For each expression level, the vast majority of the features are monotonic. As can be seen, the $\boldsymbol{R^2}$ of the linear regression between the features marked as monotonic or MIC significant is relatively high. For detailed results see tables S1-4.

As evident, in all gene expression measurements there is a very high correlation between the MIC score and Spearman correlation ($R^2 > 0.71$ in all cases; **Table 2**). Thus, high Spearman correlation is usually an indication of high MIC scores and a high MIC score is usually an indication for a strong monotonic relation.

Furthermore, there are very few features with significant $MIC - \rho^2$ which have relatively high MIC values and relatively low Spearman values; these features are depicted by an orange color in **Figure 3**. The opposite does not exist – relatively high Spearman values with relatively low MIC values; this is expected, as MIC should be able to report non-monotonic relationships that can be detected via Spearman correlation.

Moreover, most of the non-monotonic associations are associated with less significant p-values: According to **Table 2**, for a p-value cutoff of 0.05 2%-18% of the features are non-monotonic. However, for a stricter p-value cutoff of 0.01 (**Figure 4**) 1%-8% of the features are non-monotonic, and for a p-value cutoff of 0.001 (**Figure 4**) only 0%-5% of the features are non-monotonic.

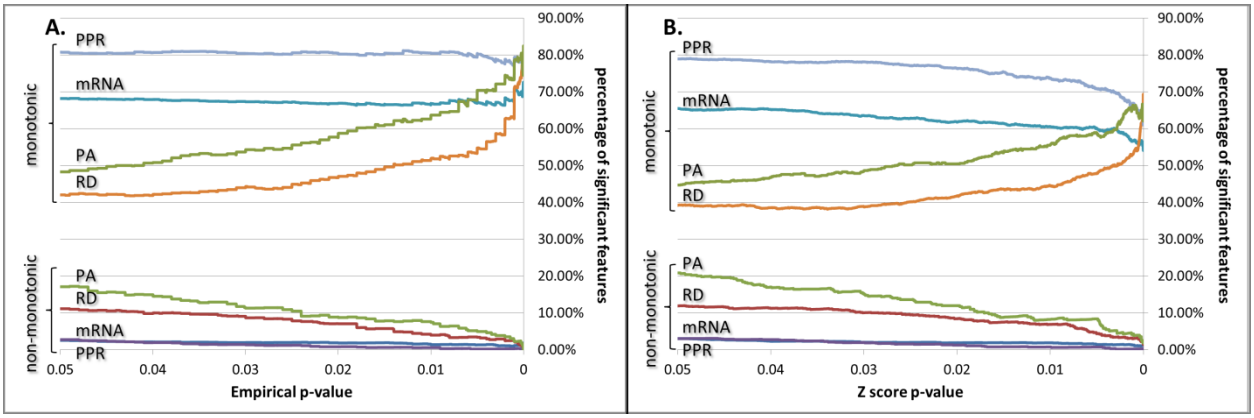



**Figure 4** – Measurements per p-value of monotonic and non-monotonic features for each expression type. The graphs show that the more strict the p-value the higher the percentage of monotonic features as opposed to non-monotonic features. This result demonstrates that the more significant features are monotonic. **A.** Using an empirical p-value, the percentage of monotonic and non-monotonic features out of all the significant features. **B.** Using a Z score p-value, the percentage of monotonic and non-monotonic features out of all the significant features.

The number of features after running False Discovery Rate is also available in Table 2. This result agrees with all the data presented above where the portion of the monotonic features is much greater than the portion of the non-monotonic features. For an empirical p-value only 0%-3% of the MIC significant features that pass FDR filtering are non-monotonic, and when using a Z score p-value (Methods) only 0.1%-5% of the MIC significant features that pass FDR filtering are non-monotonic.

Some selected results of statistically significant monotonic features are presented in **Figure 5**. These features have been discussed in the previous section, the figure illustrates the monotonicity of the relations.

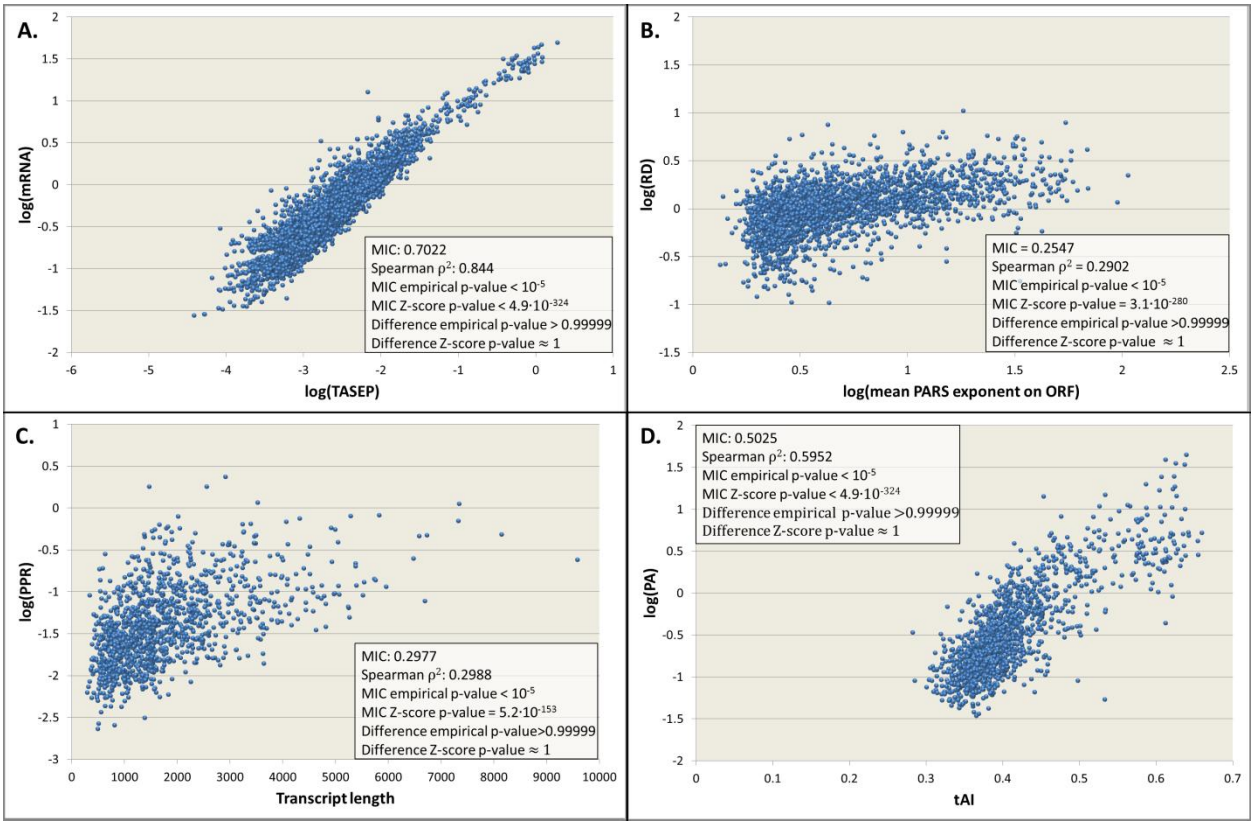



**Figure 5** – Some examples of several relations found to be significantly monotonic (all presented results have MIC empirical p-value $< \mathbf{10^{-5}}$, very low Z-score p-value and difference Z-score p-value ≈ 1). Results are presented for **A.** TASEP and mRNA. **B.** RD and mean PARS exponent of ORF. **C.** PPR and Transcript length. **D.** PA and tAI.

## Examining some significant non-monotonic features

As mentioned above, the vast majority of the analyzed features/gene-expression measurements do not exhibit significant non-monotonic relations.

There are a few features that have significant non-monotonic association with the expression level; in the remainder of this subsection we report a number of them. All results presented here are with empirical p-value $< 4 \cdot 10^{-4}$ and pass Z score p-value FDR < 0.03. For example, the relation of PARS (measurements of mRNA folding strength) of the first 40nt in the ORF *vs.* mRNA levels has MIC of 0.1864, which is very significant (Z-score p-value is $3.2 \cdot 10^{-97}$). RD also has significant associations with several PARS-related features that are non-monotonic, and the maximum metabolic cost of the sORF (small ORFs that appear in the coding region frame shifted relatively to the main ORF) across all frames for the first 30 codons is not monotonic (a MIC score of 0.1133 and the deviation from monotonicity Z-score based p-value is $1.7 \cdot 10^{-15}$). See the Methods section for more details on the aforementioned features.

PPR has fewer relations that are highly significant non-monotonic; the feature with the highest non-monotonic ranking is the CAI of 40nt in the ORF starting from nt 33 (with MIC score 0.1592 and Z-score p-value of $2.2 \cdot 10^{-5}$). PA also has several non-monotonic relations with PARS-related features and with CAI of 40nt in the ORF starting from nt 47 (which has MIC of 0.148 and Z-score p-value $1.1 \cdot 10^{-5}$). For a complete list of the results – see tables S1-4.

Some select results of statistically significant non-monotonic features are displayed in **Figure 6**. The few features that are significant non-monotonic usually exhibit relatively low MIC scores (in all these cases the MIC scores < 0.19); nevertheless here we will briefly review some of the significant non-monotonic features (**Figure 6 A-D**).

The non-monotonic relation between PARS (Parallel Analysis of RNA Structure), an experimental measure of the tendency of a nucleotide to be base paired when the mRNA sequence fold, near the beginning of the ORF and mRNA measurement **(Figure 6 A.**) may be explained by the fact that, in these regions, mRNA folding plays an important role in gene translation regulation and more generally its expression levels. However, the exact direction of the relation between the mRNA folding and the translation efficiency in this region is strongly related to the position within the ORF. In addition, the strength of the selection positively correlates with the expression levels of the gene. It is important to emphasize that since there is strong correlation between mRNA levels and protein levels (genes undergo in parallel selection for transcription and translation efficiency), some of the reported signal may be related to translation and not only to transcription. Specifically, it was observed that the very first codons of the ORF (less than the first 10 codons) and the 5' end of the 5'UTR are under selection for *weak* mRNA folding (low PARS score), probably for improving the recognition of the START codon by the pre-initiation complex[29,31,32].Thus, in this region, PARS score should have *negative* correlation with expression levels. However, it was shown that subsequent codons (codons ~10-25) are under selection for *strong* folding of the mRNA, presumably to prevent for the pre-initiation complex from continuing scanning after the start codon and from initiating translation from wrong alternative start codons downstream from the START codon[33–35]. Thus, in this region, the PARS score should have *positive* correlation with expression levels. In addition, strong folding was shown to be negatively correlated with translation elongation speed[33] (it may also affect *transcription* elongation in a similar manner), negatively contributing to protein levels and mRNA levels.

It was also suggested that strong folding over the entire transcript is generally *positively* correlated with mRNA levels and translation efficiency, presumably to prevent aggregation of mRNA molecules[27], and it may also be related to mRNA half life[36]. Thus, while folding at the beginning of the ORF is clearly related to the expression levels of the gene, the direction of the relation may vary among genes as the exact boundaries between the regions mentioned above are probably gene/context depended, resulting in a significant but non-monotonic relation.

Glutamine is the most abundant amino acid and one of the 2/5 amino acids with lowest metabolic cost in respiratory/fermentative conditions respectively[37], thus we expect a negative relation between the expression levels of a gene and the frequency of Glutamine in the protein it encodes. However, the fact that the relation between Glutamine frequency and ribosomal density is not monotonic (**Figure 6 B.**) may suggest that there are additional interactions between Glutamine and ribosomal density, translation efficiency, and protein functionality. For example, it is possible that the ribosomes tend to translate transcripts that encode more Glutamine efficiently since this amino acid has no charge and hence does not interact with its exit tunnel [33,38,39]; it is also possible that certain groups of highly/lowly expressed genes tend to have high frequency of Glutamine due to its effect on their function[40].

The non-monotonicity of the *CAI* near the beginning of the ORF (Figure 6 C.) may be explained by the fact that this region undergoes selection to include codons with *relatively* lower *CAI* in comparison to the codons downstream of this region [41]. Specifically, it was shown that when compared to the subsequent codons, the first 30-50 codons in *S. cerevisiae* tend to be relatively less adapted to the tRNA pool, and that this signal is under stronger selection in highly translated genes[41]. The fact that genes with high PPR tend to have higher *CAI* in general, but also a strong signal of relatively lower *CAI* at the beginning, probably contributes to this non-monotonic relation.

Moreover, it was reported that additional signals related to gene expression tend to be encoded near the beginning of the ORF and undergo stronger selection in highly expressed genes, thereby affecting the codon usage bias and the *CAI* in this region [5,29,31–33,35]. It is possible that these signals also contribute to the non-monotonic relation between *CAI* at the beginning of the ORF and PPR

Furthermore, it is important to mention that the relation between the *CAI* of the entire coding region and PPR is monotonic (see supplementary table 3); this is not surprising since it was suggested that highly expressed genes tend to undergo selection for specific (presumably "optimal") codons (see, for example, Plotkin and Kudla[5]).

Finally, the non-monotonic relation between the *CAI* at the beginning of the ORF and other measurement of expression levels (*e.g.*, PA and mRNA) were less significant than in the case of PPR. This fact supports the conjecture that the effect of CAI in this region on *translation* (and not, for example, transcription and/or mRNA degradation) mostly contributes to the evolution of this region; specifically, translation is related to PPR, and transcription/mRNA degradation, which are related to mRNA are also partially related to PA (since PA is a result of both transcription and mRNA levels). See Figure S2 for comparison of the CAI in the same window per different gene expression measurements.

The non-monotonic relation between the PARS of the 3' UTR and PA (**Figure 6 D.**) can be explained by the various (positive and negative) relations between folding strength and expression levels mentioned above[27,29,31–33]. In addition, it was suggested that signals related to mRNA transport and degradation are encoded in this region and thus may be affected by its folding[42].

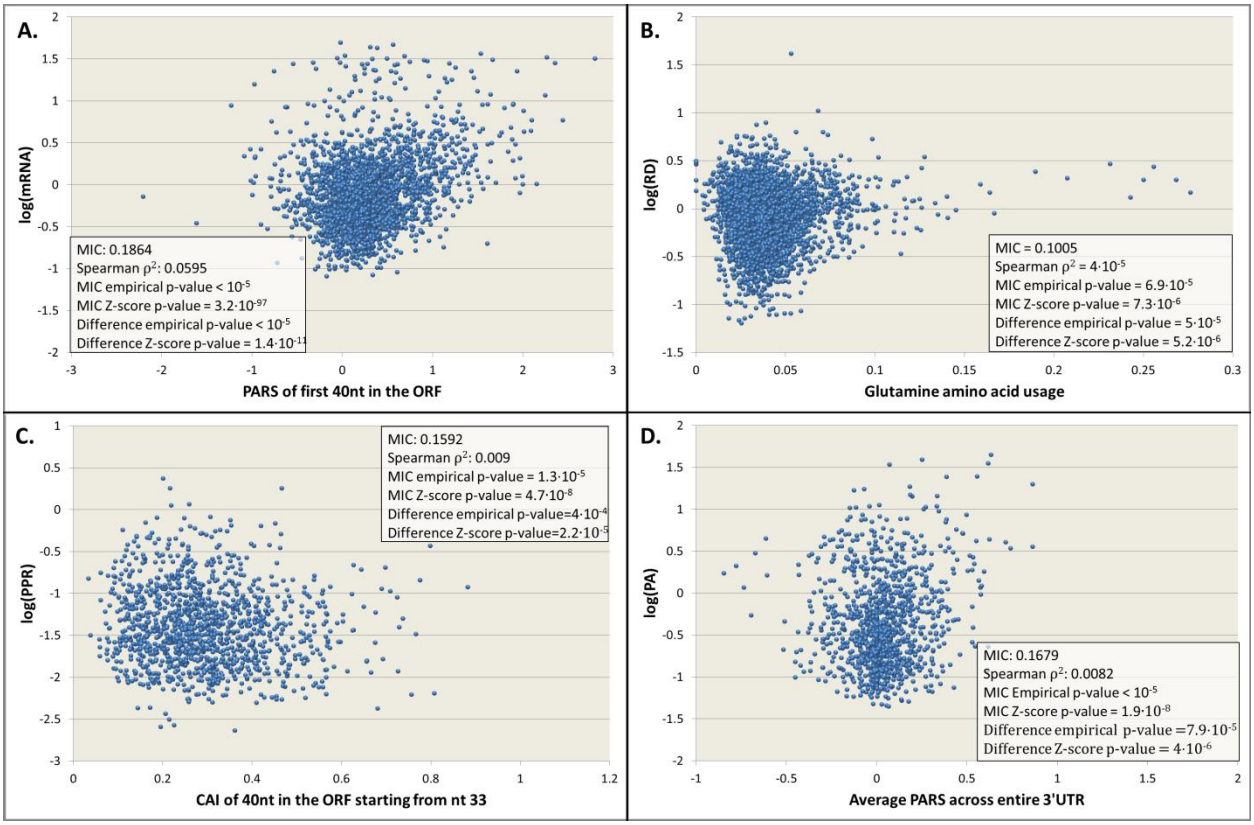



**Figure 6** – Some examples of relations found to be significantly non-monotonic (all presented results have MIC p-value $\leq \mathbf{10^{-5}}$). Results are presented for **A.** PARS calculation of the first 40nt of the ORF with its relation to mRNA; in this case, the MIC score is 0.1864 and Spearman ($\boldsymbol{\rho^2}$) is 0.0595; the difference p-value is low suggesting a non-monotonic association (empirical p-value $< 10^{-5}$, Z-score p-value $\mathbf{1.4 \cdot 10^{-11}}$). **B.** Glutamine amino acid usage (defined as the percentage of Glutamine in the protein) *vs.* RD. has a MIC score of 0.1 and Spearman ($\boldsymbol{\rho^2}$) of only $\mathbf{4 \cdot 10^{-5}}$ (the difference empirical p-value $= \mathbf{5 \cdot 10^{-5}}$ and Z-score p-value $= \mathbf{5.2 \cdot 10^{-5}}$). **C.** CAI of a windows size of 40nt in the ORF starting from nt 33 compared to PPR; in this case, the MIC is 0.1592 and the Spearman ($\boldsymbol{\rho^2}$) is 0.009 (the difference empirical p-value is $\mathbf{4 \cdot 10^{-4}}$, Z-score p-value $\mathbf{2.2 \cdot 10^{-5}}$). **D.** Average PARS across entire 3'UTR compared to PA; in this case, the MIC is 0.1679 and the Spearman 0.0082 (the non-monotonic empirical p-value $= \mathbf{9.2 \cdot 10^{-5}}$, Z-score p-value $= \mathbf{4 \cdot 10^{-6}}$).

## Discussion

The results reported in current study support the conjecture that in *S. cerevisiae* the strong relation between transcript features and various steps of gene expression are monotonic: *i.e.*, increase/decrease in the value of the transcript feature corresponds to higher/lower levels of gene expression. It is important to emphasize that in the current study we focused on *S. cerevisiae*, the organism with most abundant large scale measurements of all gene expression stages. It is possible that in other organisms (*e.g.*, prokaryotes and/or multi-cellular organisms) the associations between transcript features and gene expression are more complex. In addition, *S. cerevisiae* is known to undergo translational selection[8,43]. Thus, it is possible that the results reported here will be different for organisms with weaker selection for translation and non-synonymous aspects of the transcript[43]. Further studies on the topic for other organisms, such as mammals, will help understanding if the reported results are unique to *S. cerevisiae*. The answer to this question is deferred to future studies.

In addition, here we analyzed a large set of 5,432 features; some of these features were based on prior knowledge of this topic and have been suggested in previous studies, and others were not based on specific prior knowledge (details in the Methods section). It is possible that there are simple, yet to be

discovered, transcript features with non-monotonic relations with gene expression. We would like to emphasize that in many of the features default values and/or the zero value are over-represented. This may bring about a Spearman $\rho^2$ score that is not significant, to relations that are actually monotonic, if these values were not included. We also want to mention the possibility that some non-monotonic relations are more sensitive to noise, contributing to lower number of detected non-monotonic associations.

The results reported here have important applications to various disciplines:

For example, they should aid in understanding the way transcript features interact with the intracellular machinery to affect the expression of the transcript. Our results support the conjecture that for *S. cerevisiae*, in most cases, each transcript feature tends to affect expression in the same direction no matters what the expression levels of the transcript, and what is the level and the direction of the effect of the feature on expression.

In addition, the reported results should help developing computational predictive models of gene expression aspects based on transcript features. Specifically, our results suggest that relatively 'simple' monotonic machine learning models (*e.g.*, regressors) that can be inferred in a computationally efficient manner may achieve similar performances as more 'complicated' models.

Furthermore, in relation to the previous point, the reported results should help developing efficient synthetic biology based approaches for engineering gene expression based on manipulation of transcripts' nucleotide composition; if we know that the relations between features of the transcript and their protein levels are usually monotonic, then one can reduce the search space of the transcript sequences that optimize a certain gene expression objective function.

Finally, it was suggested that various types of mutations (point mutations, deletions/duplications of genes, etc.) affect transcript evolution via their effect on its gene expression[5,44–48]. The analyzed features include both discrete features (that obtain a value from a small set of possible values; *e.g.* GC content) that can measure the effect of point mutations, and also features that are continuous and may be related to the accumulation of many mutations. Thus, our result should contribute towards developing novel models of molecular evolution and population genetics that connect mutations to fitness via their effect on expression. Specifically, the reported results support the conjecture that at least when considering mutations that affect gene expression, a monotonic relation between mutations and fitness can usually be assumed.

## Methods

### Gene expressions data sources

All gene expression levels are taken from [11], the data sources used to obtain these expression levels are available here to provide complete description.

**mRNA levels:** Large scale measurements for *S. cerevisiae* are available from Ingolia *et al.*[23]. There are measurements for 5,295 genes, and included values for 4,176 /4,367 genes participating in the study.

**Ribosomal densities:** Large scale Ribosomal Densities (RD), defined as the number of ribosomes occupying the transcript divided by its length, were taken from two datasets, each generated by a different technology. The first dataset was generated more recently by Ingolia *et al.*[23] (4,648 genes, and included values for 3,954 /4,367 genes participating in the study), and the second by Arava[24] (5,181 genes, and included values for 4,015/4,367 genes participating in the study). The two RD datasets were averaged (after normalizing each dataset by its mean), in order to minimize experimental noise (resulting with 4316 genes, which included values for 3,682/4,367 genes participating in the study).

**Protein Abundance:** Four large scale datasets were used to calculate Protein Abundance (PA): Ghaemmaghami *et al.*[14] (3,839 genes, and included values for 3,263/4,367 genes participating in the study), two large scale measurements in two conditions from Newman *et al.*[21] (2,508/2,433 genes, and included values for 2,250 and 2,187/4,367 genes participating in the study respectively), and large scale protein abundance from Lee *et al.*[22] (2,360 genes, and included values for 2,117/4,367 genes participating in the study). Similarly to the RD, the four datasets were averaged to reduce experimental noise (resulting with 1,448 genes, which included values for 1,343/4,367 genes participating in the study).

**Proteins per mRNA molecule (PPR), (Protein Abundance)/(mRNA levels):** is the number of proteins produced on average from an mRNA molecule and termed proteins per mRNA molecule, or PPR for short. We added this feature as it is directly related to the translation stage, unlike PA which is related to both mRNA levels (transcription and mRNA degradation) and translation. The final number of proteins (PA) is related to mRNA levels the output of the transcription stage (and also related to mRNA degradation rate), and the post-transcriptional regulatory stages related to generating proteins from the mRNA sequence (*e.g.*, gene translation and protein degradation).

### Transcript features

All the features analyzed in this study, a total of 5,432, were taken from[11], below is a brief summary of the features for which we presented results herein. Additional features that were not mentioned in the Results section are available in the supplementary material

**Measured folding energy**: Recently, a new technology for measuring folding strength of RNA sequences at single nucleotide resolution was developed by Kertesz *et al.*[49]. The product of this method, named the Parallel Analysis of RNA Structure (PARS) score, includes the estimated ratio between the probability that each nucleotide in the transcript is in a double-stranded conformation and the probability that it is in a single-stranded conformation. The PARS score was computed in vitro for transcripts devoid of any ribosomes. The PARS is a global feature and thus relevant only to the combined predictor, and includes the mean PARS over each respective segment and over the entire transcript, first 40nt of the ORF, last 40nt of 5'UTR, last 40nt of 5'UTR and first 40nt of ORF, entire 5'UTR and first 40nt of ORF, entire 5'UTR and entire ORF, first 40nt of 3'UTR, first 40nt of ORF and first 40nt of 3'UTR, first 40nt of ORF entire 3'UTR, entire ORF and entire 3'UTR. Additionally the exponent of each of these features was used, in which the ratio of probabilities is used (instead of the original log ratio) of each nucleotide. For each transcript, The PARS score of its nucleotides is average, and the result represents the PARS (mF strength) exponent score.

**The number, mean and maximum length and metabolic cost of uORFs in the UTRs**: An Upstream Open Reading Frame (uORF) is a very short Open Reading Frame (ORF) within the UTR. The 5'UTR predictor includes the number, mean and maximum length and metabolic cost of uORFs across the entire

and last 30 codons of the 5'UTR. The 3'UTR predictor includes the number, mean and maximum length and metabolic cost of uORFs across the entire and first 30 codons of the 3'UTR. The features are calculated for each of the 3 frame shifts and across all frames. We also considered the following features: the number, mean and maximum length and metabolic cost of uORFs across the entire and last 30 codons of the 5'UTR allowing ending in the ORF. The total metabolic energy costs of amino acids in *S. cerevisiae* under respiratory conditions were taken from Wagner[37], and the metabolic cost of a peptide was calculated as the sum of the energy cost of the amino acids composing it. Peptide length was measured as the number of nucleotides composing it.

**The number mean and maximum length and metabolic cost of shifted ORFs (sORFs) in the coding sequence**: An sORF is a *frame shifted* truncated ORF, starting with an alternative ATG (START codon) in the ORF and terminating with a stop codon (all are frame shifted relatively to the main ORF). We considered features such as the number, mean and maximum length and metabolic cost of sORFs across the entire first 200 codons and first 30 codons ORF. In addition, the number, mean and maximum length and metabolic cost of sORFs across the entire and first 200 codons of the ORF allowing ending in the 3'UTR. The features are calculated for each of the 3 frame shifts and across all frames.

**The Codon Adaptation Index (CAI)**: a technique for analyzing codon usage bias. The CAI[9] measures the deviation of a given protein coding gene sequence with respect to a reference set of genes. Ideally, the reference set in CAI is composed of highly expressed genes, so that CAI provides an indication of gene expression levels under the assumption that there is translational selection to optimize gene sequences according to their expression levels. The CAI is simply defined as the geometric mean of the weight associated to each codon over the length of the gene sequence (measured in codons):

$$CAI = \exp\left(\frac{1}{L}\sum_{l=1}^{L} ln\big(w_i(l)\big)\right)$$

For each amino acid, the weight of each of its codons, in CAI, is computed as the ratio between the observed frequency of the codon ($f_i$) and the frequency of the synonymous codon ($f_j$) for that amino acid:
$w_i = \frac{f_i}{\max(f_j)}, ij \in [synonymous\ codons\ for\ amino\ acid]$

The ORF predictor includes the mean CAI across the entire ORF, and the first 100 sliding windows of length 40nt of the ORF. The combined predictor includes these features.

**Coding sequence tRNA Adaptation Index (tAI)**: a statistical model for measuring adaptation of codons to the tRNA pool. It assumes that the relative concentrations of the tRNA molecules that recognize a codon have a strong effect on the codon translation efficiency. This measure is determined by combining thermodynamic properties of the codon-anticodon interaction, taking into account that due to wobble interactions, several anti-codons can recognize the same codon, with different efficiency weights. The tAI[8] gauges the availability of the different tRNA molecules for each codon along an mRNA.

To calculate tAI we define the absolute adaptiveness, $W_i$ , for each codon $i$ as

$$W_i = \sum_{j=1}^{n_i}\big(1 - S_{ij}\big)tCGN_{ij}$$

where $tCGN_{ij}$ the copy number of the j[th] tRNA that recognizes the i[th] codon, and let $S_{ij}$ be a parameter corresponding to the efficiency of the codon-anticodon coupling between codon i and tRNA j. The $S_{ij}$ are inferred optimizing the correlation between the tAI and gene expression measurements.

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

The final tAI of a gene, $g$, is the following geometric mean:

$$tAIg = \left( \prod_{k=1}^{lg} w_{ikg} \right)^{1/lg}$$

Where $ikg$ is the codon defined by the k[th] triplet on gene $g$; and $lg$ is the length of the gene (excluding stop codons). Thus, the tAI of a gene is a number between 0 (extremely non-efficient codons) and 1 (utmost efficiency).

The ORF predictor includes them mean tAI across the entire ORF, and the first 100 sliding windows of length 40nt of the ORF. The combined predictor includes these features.

**Totally Asymmetric Exclusion Process (TASEP)**: The TASEP is a stochastic flow model of translation elongation, whose output is the predicted translation rate[50–52]. In the TASEP, initiation time as well as the time a ribosome spends translating each codon are exponentially distributed with a codon dependent rate. In addition, ribosomes span over several codons and if two ribosomes are adjacent, the trailing one is delayed until the ribosome in front of it has proceeded onwards.

**Codon Usage**: in this context is the frequency of each codon per gene. These 64 features are related to the ORF.

**Codon Usage Pairs**: in this context, is the frequency of pair of codons in a gene. These 4096 features (since there are 64*64 possible pairs of codons) are related to the ORF.

**Amino Acid Usage**: in this context is the frequency of each amino acid per gene. These 20 features are related to the ORF.

**Amino Acid Usage Pairs**: in this context is the frequency of each amino acid pair per gene. These 400 features are related to the ORF.

### General description of the calculations of MIC scores

As mentioned, we took from Zur & Tuller[11] the 5,432 transcriptional features of 4,367 *S. cerevisiae* endogenous genes, as well as measurements of Protein Abundance (PA), Ribosomal Density (RD), mRNA levels and Proteins per mRNA molecule (PPR).

For each pair of expression type and feature we calculated the maximal information coefficient (MIC) statistic[16]. This coefficient enables detecting non-trivial relationships between variable pairs. One advantage of MIC is that it can detect relationships without prior knowledge of the type of relationship.

For example, MIC can detect linear, mutually exclusive, cubic and sinusoidal relations, and will generally have a greater score for these relations than just random data.

For each expression and feature pair, we first discarded the measurements that had missing values (denoted by NaN) in either the expression or the feature, leaving us with a fully defined pair. Furthermore, we discarded features that had less than 40 genes with a value that is either non-zero or not default. For each pair we recorded the resultant number of variables along with the MIC score.

### Using p-values and Z-scores to rank results via permutation tests

A MIC score that corresponds to no signal has a score which is slightly above zero. This value greatly depends on the number of data points (number of genes with measurements in our case). For example, in the case of RD with 3,682 measurements, we see the non-significant MIC values ($\geq 0.05$) range between 0.068 and 0.0926 and the minimal value for which MIC is significant ($< 0.05$) is 0.079. Thus, in order to find which MIC results are significant we calculated an empirical p-value based on a permutation test of the results. For each feature-expression pair we calculate the empirical p-value by choosing random permutations of the feature and expression data (we permute the actual values and not the transcript sequences). Choosing a permutation preserved many characteristics of the data while still allowing us to verify if the MIC results are significantly higher than random. Besides the MIC calculation, we also calculated Spearman correlations for the same permutation, and the difference between MIC and Spearman ($MIC - \rho^2$). We then referred to the MIC as significant if its value was higher for more than 95% of the permutations, and we referred to the results as non-monotonic if the $MIC - \rho^2$ value was higher for more than 95% of the permutations.

The task of calculating many random permutations and calculating MIC for each permutation is CPU intensive and therefore we divide the task to smaller sub-tasks and performed the calculations using parallel computing. To obtain random independent values we used the Mersenne Twister random number generator and for each sub-task we choose a randomly generated seed.

The empirical p-value is then calculated by counting the number of cases the MIC score of the random permutation is greater or equal to the MIC score of the real data (without permutation). By dividing this value by the number of total permutations we performed, we obtain a likelihood ratio for a MIC value to appear in random data. The same process is also performed on the difference between the MIC score and the Spearman $\rho^2$ value, thus obtaining the $MIC - \rho^2$ empirical p-value.

The empirical p-values were computed as follows: First, we estimated a p-value based on 1,000 permutations. In the cases that this initial p-value did not give high enough resolution (*i.e.*, the MIC p-value was 0; or the $MIC - \rho^2$ p-value was 0 or 1) we performed additional permutations until we obtained enough permutations for determining if the feature passed FDR filtering; for example, in the case of PA we have 3681 features, thus if the MIC p-value is still 0 or the $MIC - \rho^2$ p-value is 0 or 1, the number of permutations required is $> 3681/0.05 = 73{,}620$.

In case a p-value remained 0 after completing all the permutations, we continue to randomly permute the data and calculate the MIC until we reach at least 100,000 permutations. If the p-value is still 0 we consider it to be $< 10^{-5}$ and if the p-value of $MIC - \rho^2$ is 1 we consider it to be $> 0.99999$.

In many instances the real values were higher than all the values obtained in the permutation, only allowing us to give an upper bound for the p-value, but not being able to rank it. Increasing the number of permutations helped slightly, but we still got a large portion of the associations in the top ranks without the ability to actually rank them. We therefore calculated the standard score (Z-score) of the same random data by calculating the results average and standard deviation where

$$z = \frac{x - \mu}{\sigma}$$

Where $x$ is the value obtained for the real data, and $\mu$ is the estimator for the mean obtained by calculating the average of all the sampled randomized data. For given $N$ measurements it is calculated by:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$\sigma$ is the estimator for the standard deviation of all the sampled randomized data. Its calculation is:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2}$$

We then verified that we can indeed use Z-score (that is, randomization by calculating the normal cumulative distribution value for each Z-score and comparing it to the empirical p-value). To better compare the Z-score to the empirical p-value we first transformed the Z-score to a p-value (a value between 0 and 1 instead of an unbounded Z-score). This is done by calculating the probability that an observed value is above the Z-score using the standard cumulative distribution function. In practice, this is done in Matlab.

We will denote this value as Z-score p-value and the p-value we obtained through permutation as empirical p-value.

Results showed a clear linear relation between the two p-values, after excluding cases where the feature had less than 40 non-zero or default values (which we deemed features that do not contain enough information to be useful to evaluate the association with the expression level). An example of this relationship is presented in **Figure 7**.

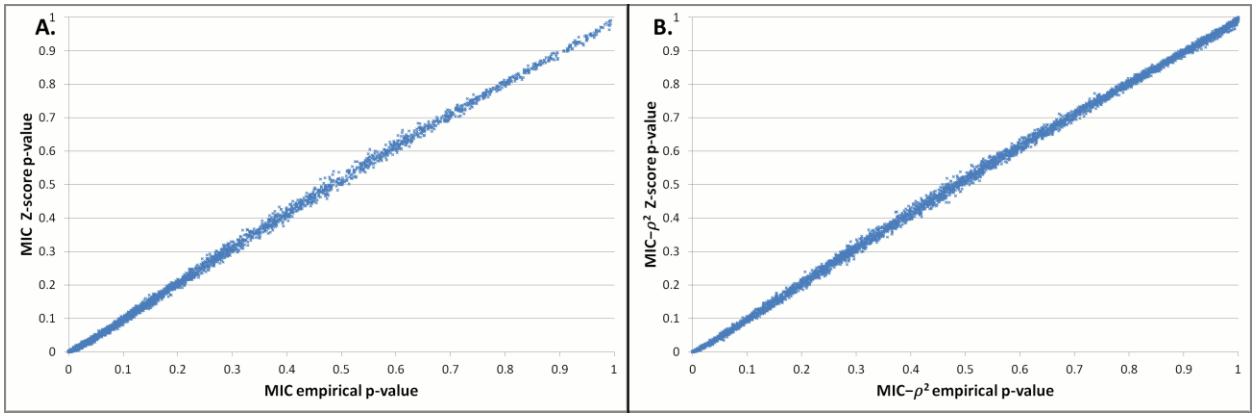


**Figure 7** – Results comparing the empirical permutation p-value to the Z-score p-value for mRNA. In both cases there is a strong linear association of the two ($\boldsymbol{R^2 = 0.9988}$). **A.** MIC empirical p-value compared to Z-score p-value. **B.** Empirical p-value of $\boldsymbol{MIC - \rho^2}$ compared to its Z-score p-value. The figure does not include features with less than 40 non-zero/non-default values.

## Controlling the false discovery rate

We calculated FDR using Matlab with the procedure introduced by Benjamini and Hochberg[53]. Prior to calculating FDR, we first ignore the features that have less than 40 non-zero/non-default values. To control the FDR of $MIC - \rho^2$ we run FDR twice. The first time we run with the difference (non-monotonic) p-value ($p$value), and in the second we calculate it for $1 - p$value. The first calculation is used to determine non-monotonic features that pass the FDR, the second is to determine monotonic features that pass the FDR.

Counting the features that pass FDR is done by checking which features have result values of less than 0.05. Just as with the p-value, a feature is significant if both the data MIC FDR and the filtered data (data without zero/default values) MIC FDR are $< 0.05$. We then consider only features that are MIC significant when examining whether they are monotonic or non-monotonic, and calculating only those p-value FDRs. A feature is considered non-monotonic if its MIC FDR is significant and if the $MIC - \rho^2$ FDR is $< 0.05$. A feature is considered monotonic if its MIC FDR is significant and the $MIC - \rho^2$ FDR for $1 - pvalue$ is $< 0.05$. In all other cases we consider the feature to be non-significant monotonic/non-monotonic.

## Detecting non-monotonic features and ranking features

We first check whether the MIC p-value is $< 0.05$ to determine if it is significant or not. To determine monotonicity, we first note that in many cases the feature data includes many data points which have a zero value or given a default value when the feature was devised. For example, if several genes do not have a specific codon, then, for those genes, the feature that measures the incidence of that codon would be zero. However, although the feature value was constant and equal to zero, the expression level was different for these genes. This sometimes creates visually 2 types of relations – the first for non-default or non-zero values and a second distinct vertical line for cases where there was no relevant measurement for the feature; see an example in **Figure S1**. In order to convincingly compare MIC to Spearman we performed a two-phase approach. The first phase was to compare MIC without any change to random permutations and we calculated the p-value. In the second phase we removed from the features the default value (and if no default value was defined, we remove the zero value if it is applicable). We then recalculate MIC and Spearman, recalculate the p-value and also calculate the p-value of the difference between MIC and Spearman ($MIC - \rho^2$).

This last comparison of $MIC - \rho^2$ is a similar approach to Reshef *et al.*[16], who used Pearson correlation to check whether the association is linear or non-linear. Spearman is calculated with Pearson on the ranked data, and MIC is invariant to order preserving transformation[16], such as ranking. We can therefore compare MIC of the data (which is equal to MIC of the ranked data) to Spearman (which is Pearson of the ranked data) and obtain a measurement of monotonicity. We also observed that MIC is roughly equal to $\rho^2$ when the only association is monotonic (see, for example, **Figure 1 B.**). However, when the association exists and it is non-monotonic MIC remains high while Spearman is low, making $MIC - \rho^2$ a very useful measurement.

Since MIC should be robust to such changes we consider only results that have a significant MIC p-value (where p-value < 0.05) in both cases. We use the p-value of the comparison after the manipulation because we expect Spearman to give us more accurate results whether the relationship is monotonic or non-monotonic. In the case where the difference p-value < 0.05, we classify the relationship as non-monotonic since we found a significant difference. In the case where the difference p-value > 0.95, we verify the Spearman p-value is also significant. Therefore, if the difference p-value >0.95 and the Spearman p-value <0.05 we classify the relationship as monotonic since we did not find a significant difference. In all other cases we would say the relationship is non-significant monotonic/non-monotonic.

### MIC implementation details

MIC is a rank order statistic that enables exploration of variable pair relationship out of thousands of variable pairs. MIC ranges between 0 and 1, but it's important to emphasize that MIC doesn't return an absolute zero value for random data. Instead, it will return a lower value for random data compared to data of the same size that has a real association. We therefore need to be certain a MIC value is significant compared to random permutations before declaring two variables are related. It is also worth noting that MIC is more computationally intensive than simpler statistic measurements, such as Spearman. This is necessary because it allows MIC to detect more complex associations than simpler statistics.

All the code was written and executed in Matlab, two specific helper functions were written in C++ since the pure Matlab execution time was inadequate. To verify the calculation yields correct results, we took the PA MIC results of the first 100 features and compared it to the results provided by the implementation presented by Reshef *et al.*[16]. All results used the default values of n=0.6 and clumpFactor=15.

## Conclusion

Most of the selected features and expression levels have a monotonic relationship. Only a few of the features exhibit potential complex associations and in this case the associations are relatively weak. Therefore, in most cases, research and models related to expression levels can focus on examining monotonic associations, and may have little benefit in considering more computationally intensive methods.

## Acknowledgment

We would like to thank Hadas Zur for providing the analyzed features and for her help with the data analysis.

## Supplementary material

For each expression type we provide per relevant features the MIC score and Spearman score, the empirical p-value for MIC and for $MIC - \rho^2$, as well as the Z-score p-value. Results are calculated before any manipulation to the data and after removal of zero/default values. Data that is calculated after removal of zero or default value is marked as "filtered" in the supplementary tables. See the Methods section for more details on these calculations. Specifically:

**Table S1** includes all the mRNA features.

**Table S2** includes all the RD features.

**Table S3** includes all the PPR features.

**Table S4** includes all the PA features.

**Table S5** summarizes the number of features for each category (category definition is also available in the same file)

## References


1. C. Vogel, R. de Sousa Abreu, D. Ko, S.-Y. Le, B. A. Shapiro, S. C. Burns, D. Sandhu, D. R. Boutz, E. M. Marcotte, and L. O. Penalva, *Mol. Syst. Biol.*, 2010, **6**.
2. D. Allan Drummond and C. O. Wilke, *Nat. Rev. Genet.*, 2009, **10**, 715–724.
3. T. Tuller, M. Kupiec, and E. Ruppin, *PLoS Comput. Biol.*, 2007, **3**, e248.
4. H. Gingold and Y. Pilpel, *Mol. Syst. Biol.*, 2011, **7**.
5. J. B. Plotkin and G. Kudla, *Nat. Rev. Genet.*, 2010, **12**, 32–42.
6. C. Dressaire, C. Gitton, P. Loubière, V. Monnet, I. Queinnec, and M. Cocaign-Bousquet, *PLoS Comput. Biol.*, 2009, **5**, e1000606.
7. T. Huang, S. Wan, Z. Xu, Y. Zheng, K.-Y. Feng, H.-P. Li, X. Kong, and Y.-D. Cai, *PLoS ONE*, 2011, **6**, e16036.
8. M. dos Reis, R. Savva, and L. Wernisch, *Nucleic Acids Res.*, 2004, **32**, 5036–5044.
9. P. M. Sharp and W. H. Li, *Nucleic Acids Res.*, 1987, **15**, 1281–1295.
10. J. M. Comeron and M. Aguadé, *J. Mol. Evol.*, 1998, **47**, 268–274.
11. H. Zur and T. Tuller, *BMC Bioinformatics*, 2013, **14**, S1.
12. G. Lithwick and H. Margalit, *Genome Res.*, 2003, **13**, 2665–2673.
13. R. Kawaguchi and J. Bailey-Serres, *Nucleic Acids Res.*, 2005, **33**, 955–965.
14. S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman, *Nature*, 2003, **425**, 737–741.
15. L. Nie, G. Wu, and W. Zhang, *Genetics*, 2006, **174**, 2229–2243.
16. D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, *Science*, 2011, **334**, 1518–1524.
17. J. M. Beman and M. T. Carolan, *Nat. Commun.*, 2013, **4**.
18. C. D. Rau, N. Wisniewski, L. D. Orozco, B. Bennett, J. Weiss, and A. J. Lusis, *Front. Genet.*, 2013, **4**.
19. J. Das, J. Mohammed, and H. Yu, *Bioinformatics*, 2012, **28**, 1873–1878.
20. L. Song, P. Langfelder, and S. Horvath, *BMC Bioinformatics*, 2012, **13**, 328.
21. J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman, *Nature*, 2006, **441**, 840–846.
22. M. V. Lee, S. E. Topper, S. L. Hubler, J. Hose, C. D. Wenger, J. J. Coon, and A. P. Gasch, *Mol. Syst. Biol.*, 2011, **7**.
23. N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman, *Science*, 2009, **324**, 218–223.
24. Y. Arava, *Proc. Natl. Acad. Sci.*, 2003, **100**, 3889–3894.
25. D. Greenbaum, C. Colangelo, K. Williams, and M. Gerstein, *Genome Biol.*, 2003, **4**, 117.
26. T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley-Interscience, Hoboken, N.J., 2006.
27. H. Zur and T. Tuller, *EMBO Rep.*, 2012, **13**, 272–277.
28. O. Man and Y. Pilpel, *Nat. Genet.*, 2007, **39**, 415–421.
29. T. Tuller, Y. Y. Waldman, M. Kupiec, and E. Ruppin, *Proc. Natl. Acad. Sci.*, 2010, **107**, 3645–3650.
30. J.-D. Wen, L. Lancaster, C. Hodges, A.-C. Zeri, S. H. Yoshimura, H. F. Noller, C. Bustamante, and I. Tinoco, *Nature*, 2008, **452**, 598–603.
31. G. Kudla, A. W. Murray, D. Tollervey, and J. B. Plotkin, *Science*, 2009, **324**, 255–258.
32. W. Gu, T. Zhou, and C. O. Wilke, *PLoS Comput. Biol.*, 2010, **6**, e1000664.
33. T. Tuller, I. Veksler-Lublinsky, N. Gazit, M. Kupiec, E. Ruppin, and M. Ziv-Ukelson, *Genome Biol.*, 2011, **12**, R110.
34. T. Ben-Yehezkel, H. Zur, T. Marx, E. Shapiro, and T. Tuller, *Genomics*, 2013, **102**, 419–429.
35. H. Zur and T. Tuller, *PLoS Comput. Biol.*, 2013, **9**, e1003136.
36. G. Lenz, A. Doron-Faigenboim, E. Z. Ron, T. Tuller, and U. Gophna, *PLoS ONE*, 2011, **6**, e28544.
37. A. Wagner, *Mol. Biol. Evol.*, 2005, **22**, 1365–1374.
38. J. Lu and C. Deutsch, *J. Mol. Biol.*, 2008, **384**, 73–86.
39. C. A. Charneski and L. D. Hurst, *PLoS Biol.*, 2013, **11**, e1001508.

40. S. H. Satyal, E. Schmidt, K. Kitagawa, N. Sondheimer, S. Lindquist, J. M. Kramer, and R. I. Morimoto, *Proc. Natl. Acad. Sci.*, 2000, **97**, 5750–5755.
41. T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, and Y. Pilpel, *Cell*, 2010, **141**, 344–354.
42. R. Shalgi, M. Lapidot, R. Shamir, and Y. Pilpel, *Genome Biol.*, 2005, **6**, R86.
43. M. dos Reis and L. Wernisch, *Mol. Biol. Evol.*, 2009, **26**, 451–461.
44. J. V. Chamary, J. L. Parmley, and L. D. Hurst, *Nat. Rev. Genet.*, 2006, **7**, 98–108.
45. K. Fredrick and M. Ibba, *Cell*, 2010, **141**, 227–229.
46. D. A. Drummond and C. O. Wilke, *Cell*, 2008, **134**, 341–352.
47. Y. Xu, P. Ma, P. Shah, A. Rokas, Y. Liu, and C. H. Johnson, *Nature*, 2013.
48. G. Cannarozzi, N. N. Schraudolph, M. Faty, P. von Rohr, M. T. Friberg, A. C. Roth, P. Gonnet, G. Gonnet, and Y. Barral, *Cell*, 2010, **141**, 355–367.
49. M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal, *Nature*, 2010, **467**, 103–107.
50. C. T. MacDonald, J. H. Gibbs, and A. C. Pipkin, *Biopolymers*, 1968, **6**, 1–25.
51. R. Heinrich and T. A. Rapoport, *J. Theor. Biol.*, 1980, **86**, 279–313.
52. L. B. Shaw, R. K. P. Zia, and K. H. Lee, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 2003, **68**, 021910.
53. Y. Benjamini and Y. Hochberg, *J. R. Stat. Soc. Ser. B Methodol.*, 1995, 289–300.