Surveying the relative impact of mRNA features on local ribosome profiling read density in 28 datasets.

Patrick BF O'Connor[1], Dmitry E. Andreev[1,2], Pavel V. Baranov[1*].

[1]School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland.

[2]Belozersky Institute, Moscow State University, Moscow, 119234, Russia.

*To whom correspondence should be addressed.

Running title: RUST statistics

Key words: ribosome profiling, ribo-seq, rates, translation elongation, elongation rates, RUST.

Abstract

Ribosome profiling is a promising technology for exploring gene expression. However, ribosome profiling data are characterized by a substantial number of outliers due to technical and biological factors. Here we introduce a simple computational method, Ribo-seq Unit Step Transformation (RUST) for the characterization of ribosome profiling data. We show that RUST is robust and outperforms conventional normalization techniques in the presence of sporadic noise. We used RUST to analyse 28 publicly available ribosome profiling datasets obtained from mammalian cells and tissues and from yeast. This revealed substantial protocol dependent variation in the composition of footprint libraries. We selected a high quality dataset to explore the mRNA features that affect local decoding rates and found that the amino acid identity encoded by the codon in the A-site is the major contributing factor followed by the identity of the codon itself and then the amino acid in the P-site. We also found that bulky amino acids slow down ribosome movement when they occur within the peptide tunnel and Proline residues may decrease or increase ribosome velocities depending on the context in which they occur. Moreover we show that a few parameters obtained with RUST are sufficient for predicting experimental densities with high accuracy. Due to its robustness and low computational demand, RUST could be used for quick routine characterization of ribosome profiling datasets to assess their quality as well as for the analysis of the relative impact of mRNA sequence features on local decoding rates.

Introduction

The advent of ribosomal profiling (ribo-seq) provided the research community with a technique that enables the characterization of the cellular translatome (the translated fraction of the transcriptome). It is based on arresting translating ribosomes and capturing the short mRNA fragments within the ribosome that are protected from nuclease cleavage. The high throughput sequencing of these fragments provides information on the mRNA locations of elongating ribosomes and thereby provides a quantitative measure of ribosome density across each transcript. This local ribosome density depends on a number of variables, one of which is the time that a ribosome dwells at each corresponding location. Accordingly, ribosome profiling data contains information that could be used to infer the properties that affect ribosome decoding (or elongation) rates. Unsurprisingly, a large number of studies analysing ribosome profiling for this purpose have been published recently (Tuller et al. 2010a; Tuller et al. 2010b; Ingolia et al. 2011; Stadler and Fire 2011; Tuller et al. 2011; Dana and Tuller 2012; Li et al. 2012; Qian et al. 2012; Charneski and Hurst 2013; Shah et al. 2013; Woolstenhulme et al. 2013; Artieri and Fraser 2014; Dana and Tuller 2014a; Dana and Tuller 2014b; Gardin et al. 2014; Lareau et al. 2014; Li et al. 2014; Pop et al. 2014; Yang et al. 2014).

There is a considerable discordance among some of the findings in these works. For example, some studies report that the availability of tRNAs cognate to the codon in the A-site strongly correlates with decoding rates in yeast (Dana and Tuller 2014a; Gardin et al. 2014) while others report the lack of such a correlation (Pop et al. 2014). These contradictions are unlikely to be wholly caused by differences in the experimental datasets. They may also be attributed to the computational methods used for estimating local decoding rates. These methods are often based on elaborate models of translation that use certain assumptions regarding the process. The abstraction required for modelling necessitates the generalization of the process across all mRNAs, although we are aware of numerous special cases. Even if the generalized models provide a reasonably accurate representation of the physical process of translation in the cell, they do not model the ribosome

profiling technique itself, which may introduce various technical artefacts. Oft-cited potential artefacts include the method used to arrest ribosomes (the result is affected by the choice (Lareau et al. 2014) and the timing (Ingolia et al. 2011; Gerashchenko and Gladyshev 2015; Jackson and Standart 2015) of antibiotic treatment), the sequence preferences of enzymes involved in the library generation (Artieri and Fraser 2014) and the quality of alignment. These artefacts may distort the output and it may not be easily to disentangle their effects in the presence of biologically functional and sporadic alterations in translation.

We set out to design a simple computational technique for the characterization of ribosome profiling data that would be based on the minimal number of assumptions regarding the process. We reasoned that it is important that such a technique would be resistant to the sporadic noise (deterministic or stochastic) imposed by the process of sequence library generation and by the erratic presence of strong biologically motivated distortions in the signal. With this in mind we developed a smoothing method for ribo-seq profile data which we term RUST (Ribo-seq Unit Step Transformation).

Here we describe the method, its performance and the impact of mRNA features on the distribution of ribosome densities in ribosome profiling data estimated with RUST. We evaluated the method on both simulated and real data and show that RUST is resistant to sporadic noise. We applied this method to 28 publicly available ribosome profiling datasets obtained from biological samples (cells or tissues) in human (Guo et al. 2010; Stadler and Fire 2011; Hsieh et al. 2012; Lee et al. 2012; Stern-Ginossar et al. 2012; Liu et al. 2013; Loayza-Puch et al. 2013; Rooijers et al. 2013; Shalgi et al. 2013; Stumpf et al. 2013; Gonzalez et al. 2014; Rubio et al. 2014; Andreev et al. 2015) and mice (Ingolia et al. 2011; Thoreen et al. 2012; Howard et al. 2013; Shalgi et al. 2013; Reid et al. 2014), as well as in yeast (Ingolia et al. 2009; Brar et al. 2012; Artieri and Fraser 2014; Lareau et al. 2014; McManus et al. 2014; Pop et al. 2014). We found a considerable protocol dependent discordance between the inferred local decoding rates in these datasets. Consistent with a previous report (Martens et al. 2015) sequences near the ends of ribosome protected fragments strongly

influence the distribution of ribosomes suggesting a substantial effect of sequencing biases, which we find to vary substantially among different datasets. In addition, we found that the experimental protocol has a major influence on ribosomal footprint distribution across different codons and even on whether the A-site codon or the P-site codon is a major determinant of footprint densities. This is consistent with earlier reports that the footprint density in ribosome profiling studies does not reflect local decoding rates adequately (Dana and Tuller 2012).

We brought forward the ribo-seq study (Andreev et al. 2015) for more detailed analysis of sequence characteristics determining footprint densities. We chose this dataset because it is one of the two with the lowest sequence bias at the ends of footprints. The other dataset is from (Rubio et al. 2014). We found for (Andreev et al. 2015) study that the major factor contributing to the distribution of ribosome densities is the identity of the amino acid (rather than of the codon) in the ribosome A-site. Four of the five slowest decoded amino acids are polar charged. Mutual synergistic affects from adjacent residues that have an appreciable influence on ribosome decoding rates are observed, but are very rare. Such interactions usually contain Proline and are observed close to the peptidyl transferase center. Finally, we demonstrate that we can reconstruct experimental ribosome density profiles of individual mRNAs with high accuracy using the parameters extracted with RUST. An important caveat is that we do not know the extent of protocol-dependent distortion of local decoding rates in this study. Nonetheless, we believe that owing to its simplicity and resistance to noise, RUST is a useful technique for the characterization and comparison of ribosome profiling studies.

Results

Ribo-seq Unit Step Transformation (RUST).

The number of ribosomes decoding a particular codon of an mRNA (and by extension the expected number of corresponding ribo-seq reads in a library) depends on three variables: the

mRNA expression level, the translation initiation rate for the corresponding open reading frame (ORF) and the time that the ribosome spends at that codon (dwell time). The latter (as an invert) is usually described as a codon elongation rate or a codon decoding rate. Note that this may include not only the time required for the accommodation of a cognate tRNA, but also may include the time required for the peptidyl transferase reaction and translocation. Setting aside (for now) technical factors involved in the generation of libraries, estimating the true decoding rates with ribo-seq is made difficult by the absence of precise measurements of initiation rates. A frequent and intuitive approach of circumventing this problem is the normalisation of the local ribo-seq signal by the average signal across the coding region (Li et al. 2012; Dana and Tuller 2014a). This approach has been described as conventional (Dana and Tuller 2014a) and we will refer to it as CN for conventional normalization. It is based on a simple and to some extent reasonable assumption that the transcript expression levels and ORF initiation rates are the same for all codons from that ORF. Supposing that ORFs have only one initiation and one termination site, then CN has two major shortcomings: it is very sensitive to outliers which frequently occur due to functional ribosome pauses (Dana and Tuller 2014a) (Fig. 1A) and it typically is applied only to the transcripts with high ribosome coverage, as the relative impact of a single read alignment on CN is excessive with sparse profile data (Fig. 1A).

We reasoned that a practical approach for the analysis of ribosome profiling data should be (i) simple; (ii) robust to outliers and sporadic noise; (iii) able to use all available data (i.e. no restriction to genes with high read coverage). This could be achieved with the smoothing of ribosome profiling densities. For this purpose we use a procedure that we term Ribo-seq Unit Step Transformation (RUST). In this procedure the ribosome densities (the number of reads corresponding to the position of the A-site codon) are converted into a binary step unit function (also known as Heaviside step function) where codons are given a score of 1 or 0 depending on whether the density exceeds the average for the corresponding ORF (Fig. 1A).

The average RUST value across all mRNAs for each putative determinant of decoding rates

may be compared to the expected value (which would occur if ribosome footprints were distributed across coding regions equiprobably) to measure its effect. Upon RUST conformation every site has a limited influence on the final score, thus it is influenced primarily by the consistent presence of reads at numerous sites. For example no differentiation is made between a stall site where the ribosome density just exceeds the average to one where the average is grossly exceeded (Fig. 1A,B). The potential disadvantage of this approach is a homogenization of the profiles which may obscure weak but real determinants of decoding rates. The method therefore is conservative and the features detected with RUST are likely to affect decoding rates considerably and consistently.

To explore whether this reasoning holds in practice we tested RUST on real and simulated data and compared it with the conventional normalization (CN).

Evaluation of RUST with simulated data.

As we do not know the true decoding rates, we cannot directly test the performance of different methods on real data. In order to evaluate RUST (and CN) we proceeded to simulate data by setting the decoding rates and sporadic noise then test their ability to estimate these rates (see Methods). Usually alignment data from only highly expressed genes are used with the CN method, therefore we also decided to explore how this affects the performance of the CN method. For this purpose we used the CN method on all transcripts with aligned footprints which we will refer to as CN>0, whereas on the transcripts with an average footprint density of >1 read/nt, we will refer to as CN>1.

The simulated alignment data were modelled using real human mRNA sequences obtained from the RefSeq database and with the average transcript read density similar to that of real ribosome profiling data obtained from HEK293T in a recent study (Andreev et al. 2015). We simulated the data under the simplistic model where the local decoding rate depends exclusively on the identity of a decoded codon (A-site codon). The number of footprints at each codon position

was determined by sampling from the following Poisson probability mass function:

$$p_{m,c,d} = \frac{\left(\frac{t_c D_m}{\sum_{C=(AAA\ldots TTT)} n_{c,m} t_c}\right)^d e^{\frac{-t_c D_m}{\sum_{C=(AAA\ldots TTT)} n_{c,m} t_c}}}{d!} \qquad [1]$$

Where $p_{m,c,d}$ is the probability of finding d number of footprints at a specific location at mRNA m at a codon c from the set of 61 sense codons C. D_m is the total number of footprints aligning to mRNA m; $n_{c,m}$ is the number of codons c in the coding region of mRNA m; and t_c is the relative dwell time for the codon c. To model the irregular noise representing sporadic ribosome pauses, technical sequencing artefacts and misalignments, the number of reads at a certain percentage of randomly selected coordinates was substituted with a 3x value of the highest footprint density for the original simulated profile (see Methods).

The only unknown parameters in the equation [1] are t_c values for each of the 61 codons. We conjectured that the accuracy of the normalisation approaches may depend on codon specific properties, such as their frequency or average read density. Therefore we simulated the data under three different sets of t_c parameters. In the first two simulations the range of t_c values were set to rank-correlate with the codon usage (see Methods), i.e. the lowest t_c was set for the rarest codon and the highest r_c for the most abundant codon. In one set the t_c range spans one order of magnitude and in the other, two orders of magnitude. In the third set, the t_c parameters were set to anticorrelate with the codon usage.

Figure 1C compares the performance of the three methods (CN>0, CN>1 and RUST) for three different sets of t_c parameters and different levels of sporadic noise. When the codon decoding dwell times have a range that spans 1 order of magnitude and an optimal (unadulterated) signal, all three approaches infer t_c values accurately (Fig. 1C, top plots). As may be expected, their accuracy is reduced as the number of coordinates affected by sporadic noise increases. While they all underestimate t_c values in the presence of the noise, the performance of RUST, however, is the least affected by the increased noise. For the scenario where the range of decoding rates spans two orders

of magnitude (Fig. 1C, middle and bottom plots) the effect of noise on the accuracy of t_c inference is much higher. Counterintuitively CN>0 performs similar or even better than CN>1 and suggests that the requirement for a minimal ribo-seq coverage threshold may not be necessary and could be a even harmful step during conventional normalization under conditions of high noise. The probable reason for CN>0 superiority over CN>1 is the signal aggregation from a larger number of transcripts, which provides a more consistent signal despite the nonsensical results within a context of individual mRNAs. Nonetheless, in all scenarios RUST is less affected by the noise in comparison with CN methods (Fig. 1C).

Encouraged by the performance of RUST on simulated data we carried out RUST analysis on real data with the purpose to infer the sequence features of mRNAs that affect local decoding rates and to see if we can predict experimental densities from data parameters extracted with RUST.

Technical artefacts result in significant variation in the composition of footprint libraries.

The velocity of a ribosome could be influenced by the sequence of mRNA in several ways (outlined in the scheme in Fig. 2A).Codons in the E-, P- and A-sites of the ribosome determine the identity of corresponding tRNAs (and amino acid) inside the ribosome. The mRNA sequence in the cavity between subunits could affect ribosome movement by directly interacting with its components. Such interactions and their effect on ribosome progression are well documented in bacteria (Li et al. 2012; O'Connor et al. 2013). In addition, the sequence upstream of the A-site codon (up to 90 nucleotides) could influence the progressive movement of the ribosome through the interactions between the peptide it encodes and ribosome peptide chanel. There are many examples of ribosome pausing mediated by the nascent peptide (Ramu et al. 2009; Ito et al. 2010; Yanagitani et al. 2011; Wei et al. 2012; Woolstenhulme et al. 2013). Lastly, the sequence downstream of the ribosome could alter its velocity through the formation of stable RNA secondary structures (Kontos et al. 2001; Tholstrup et al. 2012) or the presence RNA-protein complexes (including other

ribosomes).

In addition to these intrinsic factors affecting ribosome velocities, there are technical factors that influence the distribution of sequencing reads in ribo-seq datasets. First, particular drugs used to block elongating ribosomes could act on ribosomes only at a specific conformation (Lareau et al. 2014). Second, the timing of antibiotic treatment could also alter the distribution of ribosomes along mRNAs, i.e. when cells are pre-treated with cycloheximide progression of recently initiated ribosomes is blocked by elongating ribosomes already arrested on mRNA (Gerashchenko and Gladyshev 2014). Third, various enzymes used to cleave mRNA and generate and sequence cDNA libraries of ribosome footprints could result in the enrichment of reads with specific sequence constraints especially at the 5' and 3' boundaries of ribosome footprints where RNase cleavage and ligation reactions occur (Artieri and Fraser 2014). Fourth, the alignment step relies on how well the reference sequence matches to the genotype of the experimental system. The accuracy of alignment also depends on the existence of paralogs and transcript sequence complexity and the way how ambiguous alignments are treated.

We used an approach similar to the one recently used by Artieri and Fraser (Artieri and Fraser 2014)to analyse how much various mRNA positions (relative to the ribosome) affect ribo-seq read density. We calculated RUST scores for 60 codons for each codon position within a window of 60 codons (from -40 to +20 relative to the A-site) and compared these scores to the values that would be expected if footprint distribution across coding regions was equiprobable (see Methods and Supplemental Fig. S1). To measure the contribution of local mRNA positions to the density of footprints correspondingly derived from a ribosome decoding a particular A-site codon, we measured the relative entropy at each codon position using the Kulback-Leibler (K-L) divergence:

$$D_l = \sum_c \frac{ro_{cl}}{\sum_{C=(AAA\ldots TTT)} ro_{cl}} log_2 \left(\frac{{ro_{cl}}/{\sum_{C=(AAA\ldots TTT)} ro_{cl}}}{{re_c}/{\sum_{C=(AAA\ldots TTT)} re_c}} \right) \qquad [2]$$

Where D_l is the K-L at location l, ro_{cl} is the observed RUST value for codon c at location l and re_c is the expected RUST value for codon c. The higher the K-L, the less uniform the distribution of RUST values is in the corresponding position. Thus, K-L indicates how much the corresponding position relative to the A-site contributes to the abundance of footprints.

Figure 2B shows the relative entropy and normalized *ro*/*re* RUST ratios for each individual codon for some of the ribosomal profiling datasets explored in this work. By analogy with metagene profiles we refer to the plots of *ro/re* RUST ratios as metafootprint profiles. It can be seen that the areas of reduced entropy are mostly contained within a window of 10 codons upstream and downstream of the A-site, approximately matching to the position of the actual ribosome footprint itself. This is observed in all ribo-seq datasets (Supplemental Figs S2 and S3). In almost all cases three local K-L maxima are observed, one corresponding to the position of the A-site or the P-site codon (indicated as -1 in Figure 2B), the other two maxima roughly corresponding to the 5' and 3' ends of ribosome footprints. The same procedure carried out on mRNA-seq libraries reveals decreased entropy in the same area but only with two maxima corresponding to the mRNA fragment ends (Fig. 2B). This strongly suggests that the main contributing factors to footprint composition in the library are the identity of the codons in the A- and/or P-sites and the sequence-specificity of the enzymes used during library construction. The degree of variation in the relative impact of these factors among different datasets is surprising. In some of the ribo-seq datasets, the density of footprints depends more on the identity of the codon at the ends of footprint more than on the identity of the codon in the A- or P-sites.

An important factor in this analysis is the application of the correct offset for inferring the position of the A-site from the 5' end. This is typically estimated with a metagene profile of either initiating or terminating ribosomes. This may not always allow a precise estimation of the offset and it is possible that initiating or terminating ribosomes do not protect mRNAs in the same way as elongating ones. With the premise that the A-site should have the greatest influence on decoding rates we set out to estimate the offset using RUST codon metafootprint profiles with different

offsets. For one of the datasets (with low sequencing bias) we found that the maximal K-L divergence for reads of 32 nucleotides corresponded to the offset determined with initiating ribosomes (Supplemental Fig. S4).

The estimation of sequencing biases introduced by cDNA library generation protocols is comparatively easy, as they act predominantly on positions distant from the A- and P-sites of the ribosome. The effect of the lysis protocols and antibiotic treatments is more difficult to estimate. A particular codon may be enriched at the A-site of the ribosome because either the ribosome decodes this codon slowly or the translation inhibitor blocks the ribosome preferentially at this codon.

To explore how specific conditions of the ribosome profiling experiment may affect cDNA library compositions we surveyed RUST ratio values for the codons in the A-site for the 28 ribosome profiling datasets and some of their mRNA-seq controls (Fig. 2C). The conditions of the experimental protocols (which we believe may affect the distribution of ribosomes or their footprints across the transcriptome) are given in the Table 1. Most apparent is the high reproducibility for most ribosomal profiling datasets produced in yeast under cycloheximide pretreatment, which is not so surprising given no apparent variations in the protocol (Table 1). The variance across the datasets obtained from mammalian sources is more substantial as are the differences in the protocols (Table 1).

Some of the studies produced the data under very similar protocols with a single parameter being different: the samples were either pretreated with cycloheximide before lysis or no drug was used (Ingolia et al. 2011; Stadler and Fire 2011; Stern-Ginossar et al. 2012; Lareau et al. 2014). Comparing these datasets to each other should, in theory, allow us to estimate the effect of individual experimental factors on the composition of footprint cDNA libraries. However, we found that ‘Stadler’ and ‘Stern-Ginnoassar’ datasets are similar for both types of treatments, while ‘Lareau’ and ‘Ingolia’ are different (Fig. 2C). Supplemental Figure S5 provides the analysis of RUST ratios for ‘Lareau’ and ‘Ingolia’ datasets under both conditions, clearly indicating that cycloheximide

substantially alters the distribution of footprints on mRNA. The difference is even more pronounced when RUST ratios are compared for the P-site codons (Supplemental Fig. S6). This is consistent with the observation that cycloheximide blocks ribosomes in a specific conformation and this ribosome arrest has certain codon preferences (Lareau et al. 2014).

Prior studies explored the effects of different antibiotic treatments in mammalian cells (Ingolia et al. 2011) and in yeast (Gerashchenko and Gladyshev 2014; Lareau et al. 2014). The effect of buffer conditions on triplet periodicity was also explored to some extent and lower concentrations of di- and monovalent ions were found to improve it (Ingolia et al. 2012; Stern-Ginossar et al. 2012). It is clear, however, that more systematic studies of experimental protocol conditions effect on local footprint distributions are needed before we could confidently use ribo-seq data for inferring local decoding rates from ribo-seq densities. At present it is tempting to conclude that there are important factors in the ribosomal profiling experimental protocol that researchers do not describe routinely in their research articles and perhaps are even unaware of.

Detailed characterization of sequence-specific factors affecting the composition of ribo-seq libraries.

(Artieri and Fraser 2014) attempted to correct for technical sequencing biases by adjusting ribosome footprint densities to that of mRNA densities. We believe that such an approach only partially solves the problem, as the mRNA fragmentation is done under different protocols for mRNA-seq and ribo-seq, alkaline digestion for the former and enzymatic digestion for the latter. Differences in RNA isolation protocols, ribosome-bound RNA vs polyadenylated RNA may also contribute to certain differences. To consider that any bias at the footprint ends to be of a technical nature would also be incorrect. The codons at the 5’ end of a footprint could affect the velocity of the ribosome through the nascent peptide within the tunnel while the footprint sequence at the 3’ end could affect it through RNA secondary structures.

Therefore, in the absence of a straight forward approach for sequence bias correction we chose to concentrate on a dataset with a low sequencing bias. We chose the dataset obtained in HEK293T under control conditions in the Andreev et al study (Andreev et al. 2015) because the K-L divergence for the RUST values is high at the decoding center, and low at the footprint ends (Fig. 3A and Supplemental Figure S2). Interestingly its A-site RUST ratio was also found to strongly correlate to the only dataset with a comparable low sequencing bias (Rubio et al. 2014), see Figure 2C. First we explored in greater detail how the identities of particular codons in the A-site affect the density of corresponding footprints. Displayed below the metafootprint profiles on Fig. 3A is the relative RUST ratios (with the lowest fixed as 1) for each of the 61 sense codons sorted by the amino acids that they encode. The RUST ratios for individual codons vary ~15 fold. The codons encoding four of the five polar charged amino acids were found to be the slowest residues, the exception was Arg codons which were found to have comparatively low RUST ratios (Fig.3A). The highest variation between synonymous codons was observed for Ser codons, ~4 fold difference between the slowest and the fastest. We do not observe the previously reported slower decoding of wobble base-pair decoded codons (Stadler and Fire 2011) (Fig. 3A). We also explored the relationship of RUST ratios of the A-site to codon usage and tRNA availability (dos Reis et al. 2004) which were reported to correlate with codon decoding rates on several occasions (Dana and Tuller 2014a; Gardin et al. 2014). We have not found a correlation between codon usage and RUST ratios (Fig. 3B). The codon usage comprise of statistics of codon frequencies in the genes irrespective of their level of expression or even whether they are expressed. Therefore we also calculated a translatome usage that calculates the frequency of codons in the translatome that takes into account transcriptional/translational levels of each gene (see Methods) and in effect is a measure of how frequently an elongating ribosome encounters a particular codon. Nonetheless, no correlation was observed for this index either, actually the translatome codon usage was very similar to the codon usage (Supplemental Fig. S7). We also did not observe a strong correlation between RUST ratios at the A-site obtained with amino acids and amino acid usage (Fig. 3B).

We explored the potential effect of RNA secondary structure on the density of footprints. We calculated the RUST ratios for RNA sequences that can form RNA secondary structures with a particular free energy as calculated with RNAfold (Lorenz et al. 2011), see Methods. Figure 3C shows the distribution of RUST values for RNA secondary structures with 80 nucleotides window with different free energy for both mRNA-seq and ribo-seq reads. It can be seen that sequences predicted to contain stable structures are avoided (low RUST values) in windows that overlap with sequencing reads. This is observed for both ribo-seq and mRNA-seq reads and therefore is likely to be an artefact related to cDNA library generation and sequencing. The nucleotide bias is unlikely to explain this since according to RUST metafootprint profile shown at Supplemental Figure S8A the distribution of individual nucleotides at the footprint location does not deviate significantly from locations remote from it (the exception is the location of the A-site).

To explore the effect of the nascent peptide on footprint densities we analysed RUST ratios for 30 amino acids upstream of the A-site codon and found bulky amino acids (large Tyrosine and Tryptophan as wells as the inflexible Proline) to have higher RUST ratios than other amino acids (Fig. 3D). This was not observed for the mRNA-seq controls (Fig. 3D).

The nascent peptide interactions with the ribosome components may be facilitated by specific physicochemical properties of the peptide. For example it has been proposed previously that positively charged amino acids slow ribosome movement (Lu and Deutsch 2008; Charneski and Hurst 2013). In this case the RUST ratio of individual amino acids may not provide an accurate representation of the of the nascent peptide effect on ribosome movement. For example the inhibitory influence of a positive charge in a nascent peptide could be mitigated by an adjacent negatively charged amino acid. Therefore we also measured RUST ratios for peptide fragments (as a sliding window of 10 residues) with particular physicochemical properties (number of positive charges, net charge and number of hydrophobic amino acids). We observed only minor deviations for the distributions of these physicochemical properties that may not necessarily be caused by their effects on decoding rates (Supplemental Fig. S8B).

Synergistic effects of amino acid interactions within the nascent peptide.

We next set out to examine whether particular di- or tripeptides could affect ribosome velocity more significantly than it would be expected from their individual components. We detect such synergetic effects by comparing the frequencies of dipeptides and tripeptides to the expected values based on the frequencies of corresponding residues at their respective positions by using the standard score (Z-score) as the following:

$$S_{ij} = \frac{ro_{ij}/re_{ij} - ro_i ro_j / re_i re_j}{std}, S_{ijk} = \frac{ro_{ijk}/re_{ijk} - ro_i ro_j ro_k / re_i re_j re_k}{std} \quad [3]$$

Where S_{ij} and S_{ijk} are synergy indexes for dipeptide *ij* or tripeptide *ijk* and *ro/re* are corresponding RUST ratios. *std* is the standard deviation of the differences observed at regions from -40 to -30 and +7 + 17 relative to the A-site. This is based on the assumption that at these positions the sequence of the encoded peptide should not significantly affect velocities. Thus differences between RUST ratios for residues and for combinations of residues should represent the level of stochasticity in the data. This approach could be carried out for a pair of amino acids at any distance from each other. We carried out this analysis for adjacent amino acids only. Out of the 23,600 possible dipeptides (20x20 codons at 59 positions in the interval from -40 to +20) the 2ndstrongest synergism was observed for the Proline dipeptide, di-P, at positions -1 and -2, i.e. two C-terminal residues of the growing peptide. The RUST ratio for di-P is about 25% larger than what would be expected from individual contributions of two Proline residues in the corresponding positions (Fig. 4A). We also explored synergism for 464,000 tripeptides three codon motifs (20x20x20 residues x 58 positions) (Fig 4B, 4C). Only 0.15% (703) of these tripeptides were found to have a standard score greater than 5 (S_{ijk}>5 or S_{ijk}< -5). These synergistic interactions were found to occur mostly near the decoding center with 26% of them occurring at the positions where the first two residues of the tripeptide are the last C-terminal residues of the growing nascent peptide and the last amino acid is

the one attached to the A-site tRNA (Fig. 4D).They also had a relatively small influence with the majority of interactions having less than a 2 fold change between observed and expected values (Fig 4E).

The amino acid Proline was also found to be a mainstay in the positive synergetic interactions in tripeptides with it occurring at least once in 16 of the top 20 candidates. Ten of these contain di-P, all with the 2ndProline being in the P-site of the ribosome. Proline is a poor substrate for peptide bond formation and frequently occurs at known pause sites. Di-P has been found to induce ribosome stalling in bacteria which is relieved by EF-P(Peil et al. 2013).Therefore it was quite surprising to find that the PPP was found among the top tripeptides of negative synergism (decoded faster than what would be expected from individual Proline contributions). Perhaps this is because of the action of elongation factor, eIF5A, a functional homologue of EF-P, required to relieve the associated pausing at polyproline regions (Gutierrez et al. 2013).

RUST parameters accurately predict experimental footprint densities.

We know neither the true local decoding rates nor how various technical factors may distort the relationship between ribosome densities and observed footprint densities. Therefore the performance of a computational method cannot be evaluated based on the prediction of local decoding rates. A good characterization of the method would be the ability to accurately reconstruct the original dataset from a small number of parameters.

Therefore, we decided to test whether we can accurately reconstruct ribosome densities using RUST parameters. We started with the simplest 1st order model incorporating the least number of parameters corresponding to the most significant factor affecting footprint densities, the identity of a codon in the A-site. According to this model the density at a particular codon will be predicted to be an average of the footprint density for a specific mRNA multiplied by the RUST ratio for the corresponding codon. The 2nd model is based on the same concept except that it uses

the paired amino acid information of amino acids in the A and P sites. The third model incorporates RUST values for all codon sites along the entire footprint. The predicted profile can be represented as a discrete probability density function

$$p_k = \frac{\prod_{i=1}^{N} r_{oik}/r_{eik}}{\sum_{j=1}^{M}(\prod_{i=1}^{N} r_{oij}/r_{eij})} \qquad [4]$$

Where p_k is the probability of finding a footprint at position k of the mRNA coding region consisting of M codons.r_{oik}/r_{eik} is the RUST ratio for the codon at site i (relative to codon k) from the total of N sites used.

We compared the results of these models to the actual ribo-seq profiles of genes with read density greater than 1 read/nt. It was found to have impressive predictive power (Fig. 5A) Predictions made based only on the A-site RUST values correlate with the real profiles (r^2=0.451, ρ=0.503). Predictions based on the pair of amino acids in the P- and A-sites are slightly more powerful (Fig. 5A). Thus, the gain of information on the decoding rates of the P-site appears to equate that of replacing the codon information at the A-site to the amino acid. The incorporation of RUST ratios for all codon sites in the footprint improves the predictive power even further (r^2=0.622, ρ=0.640). This improvement most probably is due to better representation of the influence of sequence biases on the cDNA library generation and sequencing rather than to improvements of estimated ribosomal decoding rates. An example of a profile predicted with this model is shown in Figure 5B.

Discussion

In this work we described a simple computational technique RUST for the characterization of ribosome profiling data based on a simple smoothing transformation of ribosome density profiles into a binary function. Using simulated data we show that this technique is robust in the presence of sporadic noise and outperforms conventional methods for normalisation of ribosome profiling data.

Using experimental data, we show that the characteristics of ribosome profiling data extracted with RUST can be used for the accurate prediction of experimental ribosome footprint densities.

We applied this technique to almost thirty publicly available ribo-seq datasets (yeast, mammalian cultured cells and tissues) and uncovered substantial protocol dependent variability among them. The most similar datasets are those obtained with cycloheximide pre-treatments of yeast cells and probably reflects the low variation among protocols in these studies. A strong reproducibility does not necessarily indicate that ribosome footprint densities accurately reflect decoding rates of the ribosomes as can be suggested based on strong dissimilarity with the dataset obtained with no cycloheximide pre-treatment. For the datasets obtained in mammalian systems we found substantial variation that is likely to be related to the timing of cycloheximide treatments as well as conditions of buffers used for lysis and nuclease digestion. The position specificity of sequencing biases (they affect the boundaries of ribosome footprints) enabled us to discriminate (at least partially) the influence from genuine decoding rate differences from that of sequence biases. These technical biases may have a limited effect on ribosome profiling experiments when those are applied for gene expression analysis, since they are independent of physiological conditions in which the experiments are carried out. However, they could have a very large effect on the distribution of footprints within the same transcripts and thus could complicate the analysis of local decoding rates. We found that sequence biases related to cDNA library generation and sequencing varies substantially among datasets. In some datasets, the identity of the codon at the position of a footprint boundary is a stronger predictor of footprint occurrence in the library than the identity of the codon in the A-site.

RUST allows for the estimation of the degree of technical bias across different datasets. We selected a dataset with a low sequencing bias to further explore the sequence factors that influence local decoding rates. We found that the identity of the codon in the ribosomal A-site is the most powerful predictive factor of the local decoding rates. Similar to previous studies(Lareau et al. 2014), the variation observed for non-synonymous codons is substantially higher (~10 fold) than

what is observed for synonymous codons (~3 fold). Moreover, we showed that the identity of the amino acids in the P- and A-sites is as powerful a predictor of decoding rates as the identity of the A-site codon. If we assume that experimental conditions did not affect the distribution of the ribosomes on the transcriptome, this may suggest that the speed with which tRNAs are accommodated into the ribosomes is affected by the identity of amino acids participating in the peptidyl transferase reaction. We also found that large (tyrosine and tryptophan) and bulky (Proline) amino acids in the nascent peptide slow down ribosomes. The synergetic effects between neighbour amino acids in the nascent peptide are rare, insubstantial and frequently involve Proline residues.

Our results suggest that sites other than at the decoding center have a relatively minor influence the decoding rate in general. This is perhaps not too surprising when we consider that the primary function of mRNA is to encode a protein. If there were pervasive and significant inhibitory interactions from multiple sources, the productivity of elongating ribosomes would be significantly decreased. It may also constrain the possible codon combinations that may be decoded. This would significantly outweigh the advantages that may be gained from such a system such as translational regulation. Our results do not suggest that local decoding rates could not be affected significantly by other factors. Rather, they suggest that such factors are not general. They could be highly specific. Indeed, we know many examples where RNA secondary structures (Somogyi et al. 1993; Tholstrup et al. 2012) or nascent peptide signals cause exceptionally long pauses. Such signals could even alter the standard decoding, see (Baranov et al. 2002; Namy et al. 2004) for reviews. Moreover RUST could be used to find such specific signals by comparing the real footprint densities to those predicted based on general factors.

In conclusion, we believe that RUST will be a valuable tool for characterizing ribosome profiling data due to its simplicity and resistance to sporadic noise (technical and biological). RUST metafootprint analysis can be used effectively to estimate the degree of sequencing bias in ribo-seq datasets and we hope it will help to improve and standardise ribosome profiling protocols. In addition RUST can be used for determining sequence factors affecting local decoding rates.

Methods

Ribosome profiling datasets used in this study and their processing

The datasets (and SRA repository accession numbers) are summarized in Supplemental Table S1. The processing of the reads consisted of clipping the adapter sequence and removal of ribosomal RNA reads followed by the alignment of the mammalian reads to the RefSeq transcriptome and the yeast reads to the sacCer3 genome. The Human RefSeq catalogue was downloaded on 13th Aug 2014 and the mouse RefSeq catalogue was downloaded on 18th March 2014 from the NCBI ftp website ftp://ftp.ncbi.nlm.nih.gov/refseq/ (Pruitt et al. 2014). The sacCer3 genome and annotation data were downloaded on 13th Aug 2014 from the UCSC genome browser website http://genome-euro.ucsc.edu (Karolchik et al. 2014).

Bowtie version 1.0.0 was used to carry out the alignments(Langmead et al. 2009). The mammalian reads were aligned to the RefSeq catalogue using the same approach as in (Andreev et al. 2015).Reads were aligned using bowtie to the entire human or mouse catalogue with the following parameters (-a, -m 100 –norc).The reads that mapped unambiguously to a gene (but not necessarily to a single transcript) were brought forward for further analysis. For the yeast datasets, reads were aligned to the yeast genome allowing only unambiguous alignments (-a, -m 1).

Ribo-seq simulation

The simulated ribo-seq profiles were designed such that the final ribo-seq density was similar to that observed in (Andreev et al. 2015). The dwell times t_c for the 61 sense codons were set to span either a 10 or 100 fold range with equal increments of 0.15 or 1.5, (the fastest codon was given a score of 1, the slowest was 10.15 or 101.5). In order to create noisy alignment data, a number of codons were randomly selected and the number of alignments to them was replaced with a constant. The number of codons selected was calculated as a percentage (either 1.5% or 10%) of the number

of codons with a mapped read.

Normalisation approaches

The alignment data to the longest coding transcript of every expressed gene was used. Owing to possible atypical translation at the beginning or end of the coding regions, the analysis was carried out on coding regions with the A-site position within 120 nucleotides (40 codons) downstream of the start codon and 60 nucleotides upstream of the stop codon. Generally the analysis was carried out solely by using reads of 32 nucleotides in length for the human and mouse datasets and 30nt for the yeast dataset. An offset of 17 nucleotides was used to indicate the A-site. The exclusive selection of reads of one length was done to minimise the attenuation of the enrichment signal at specific sites of the ribosome such as the termini. In this analysis all reads were used irrespective of the subcodon position to which they aligned. The exclusive selection of reads that align to a particular subcodon position may further improve the signal at the expense of sequencing depth.

An implementation of our RUST algorithm is provided in a custom script written in python ("RUST_script.py" in supplemental material).

Indexes of usage and adaptiveness of codons

Codon usage frequency and relative codon usage frequency were obtained from the GenScript website (http://www.genscript.com/cgi-bin/tools/codon_freq_table). The *wi* or relative adaptiveness value of codon was obtained from supplementary table S1 from(Tuller et al. 2010a).

The translatome codon usage was calculated as a measure of the demand of each codon to translating ribosomes based on ribo-seq data. For each codon it was taken to be the cumulative value of the average read density across the coding region in each coding region of each transcript.

RNA structure free energy prediction

The computational prediction of RNA binding free energy was predicted using RNAfold in the

ViennaRNA package(Lorenz et al. 2011). Using a sliding window of 80 nucleotides and a step size of 10 nucleotides the free energy was recorded across each transcript. The quantitative free energy value calculated for every window was classified based on whether it was lower than -40.1, -32.8 or -29.0 kcal/mol which correspond to the 1st, 5th and 10th percentiles. This classification was applied to all 10 nucleotides in one step size.

The comparison of predicted and experimental footprint densities.

The comparison was carried out on 3,319 transcripts with density greater than 1 read /nt. To maximise read density the ribo-seq profile reads of length 29 to 35 inclusive were used to produce the real profile. Only unambiguously mapped reads were used. For genes with multiple transcript isoforms, this analysis was carried out on the transcript isoform with the longest coding region. A custom script, "Profile_prediction.py" is provided in supplementary files enabling the prediction of ribosome profiles using data obtained from "RUST_script.py" with codons -6 to +6.

Acknowledgments. We would like to thank Audrey Michel for critical reading of the manuscript. We also are grateful to Nicholas Ingolia, Noam Stern-Ginossar, Shu-Bin Qian and Jonathan Weissman for providing us with details of experimental protocols that were used to generate the datasets surveyed in this work.

This work was supported by grants from Science Foundation Ireland (12/IA/1335) and the Wellcome Trust (094423) to P.V.B.

Author Contributions. P.B.F.O'C and P.V.B. conceived the study. P.B.F.O'C developed the method and carried out the data analysis. D.E.A. surveyed ribosomal profiling protocols. All authors participated in interpretation of the data. P.B.F.O'C and P.V.B. wrote the manuscript.

Conflict of interests. The authors wish to declare no conflict of interests.

References.

Andreev DE, O'Connor PB, Fahey C, Kenny EM, Terenin IM, Dmitriev SE, Cormican P, Morris DW, Shatsky IN, Baranov PV. 2015. Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *eLife* **4**: e03971.

Artieri CG, Fraser HB. 2014. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome research* **24**(12): 2011-2021.

Baranov PV, Gesteland RF, Atkins JF. 2002. Recoding: translational bifurcations in gene expression. *Gene* **286**(2): 187-201.

Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS. 2012. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335**(6068): 552-557.

Charneski CA, Hurst LD. 2013. Positively charged residues are the major determinants of ribosomal velocity. *PLoS biology* **11**(3): e1001508.

Dana A, Tuller T. 2012. Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. *PLoS computational biology* **8**(11): e1002755.

-. 2014a. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic acids research* **42**(14): 9171-9181.

-. 2014b. Properties and determinants of codon decoding time distributions. *BMC genomics* **15 Suppl 6**(Suppl 6): S13.

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

Gardin J, Yeasmin R, Yurovsky A, Cai Y, Skiena S, Futcher B. 2014. Measurement of average decoding rates of the 61 sense codons in vivo. *eLife* **3**.

Gerashchenko MV, Gladyshev VN. 2014. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic acids research* **42**(17): e134.

-. 2015. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic acids research* **42**(17): e134.

Gonzalez C, Sims JS, Hornstein N, Mela A, Garcia F, Lei L, Gass DA, Amendolara B, Bruce JN, Canoll P et al. 2014. Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **34**(33): 10924-10936.

Guo H, Ingolia NT, Weissman JS, Bartel DP. 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**(7308): 835-840.

Gutierrez E, Shin BS, Woolstenhulme CJ, Kim JR, Saini P, Buskirk AR, Dever TE. 2013. eIF5A promotes translation of polyproline motifs. *Molecular cell* **51**(1): 35-45.

Howard MT, Carlson BA, Anderson CB, Hatfield DL. 2013. Translational redefinition of UGA codons is regulated by selenium availability. *The Journal of biological chemistry* **288**(27): 19401-19413.

Hsieh AC, Liu Y, Edlind MP, Ingolia NT, Janes MR, Sher A, Shi EY, Stumpf CR, Christensen C, Bonham MJ et al. 2012. The translational landscape of mTOR signalling steers cancer initiation and metastasis. *Nature* **485**(7396): 55-61.

Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. 2012. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature protocols* **7**(8): 1534-1550.

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**(5924): 218-223.

Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**(4): 789-802.

Ito K, Chiba S, Pogliano K. 2010. Divergent stalling sequences sense and control cellular physiology. *Biochemical and biophysical research communications* **393**(1): 1-5.

Jackson R, Standart N. 2015. The awesome power of ribosome profiling. *Rna* **21**(4): 652-654.

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA,

Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser database: 2014 update. *Nucleic acids research* **42**(Database issue): D764-770.

Kontos H, Napthine S, Brierley I. 2001. Ribosomal pausing at a frameshifter RNA pseudoknot is sensitive to reading phase but shows little correlation with frameshift efficiency. *Molecular and cellular biology* **21**(24): 8657-8670.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**(3): R25.

Lareau LF, Hite DH, Hogan GJ, Brown PO. 2014. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *eLife* **3**: e01257.

Lee S, Liu B, Huang SX, Shen B, Qian SB. 2012. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of America* **109**(37): E2424-2432.

Li GW, Burkhardt D, Gross C, Weissman JS. 2014. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157**(3): 624-635.

Li GW, Oh E, Weissman JS. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484**(7395): 538-541.

Liu B, Han Y, Qian SB. 2013. Cotranslational response to proteotoxic stress by elongation pausing of ribosomes. *Molecular cell* **49**(3): 453-463.

Loayza-Puch F, Drost J, Rooijers K, Lopes R, Elkon R, Agami R. 2013. p53 induces transcriptional and translational programs to suppress cell proliferation and growth. *Genome biology* **14**(4): R32.

Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms for molecular biology : AMB* **6**: 26.

Lu J, Deutsch C. 2008. Electrostatics in the ribosomal tunnel modulate chain elongation rates. *Journal of molecular biology* **384**(1): 73-86.

Martens AT, Taylor J, Hilser VJ. 2015. Ribosome A and P sites revealed by length analysis of ribosome profiling data. *Nucleic acids research* **43**(7): 3680-3687.

McManus CJ, May GE, Spealman P, Shteyman A. 2014. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome research* **24**(3): 422-430.

Namy O, Rousset JP, Napthine S, Brierley I. 2004. Reprogrammed genetic decoding in cellular gene expression. *Molecular cell* **13**(2): 157-168.

O'Connor PB, Li GW, Weissman JS, Atkins JF, Baranov PV. 2013. rRNA:mRNA pairing alters the length and the symmetry of mRNA-protected fragments in ribosome profiling experiments. *Bioinformatics* **29**(12): 1488-1491.

Peil L, Starosta AL, Lassak J, Atkinson GC, Virumae K, Spitzer M, Tenson T, Jung K, Remme J, Wilson DN. 2013. Distinct XPPX sequence motifs induce ribosome stalling, which is rescued by the translation elongation factor EF-P. *Proceedings of the National Academy of Sciences of the United States of America* **110**(38): 15265-15270.

Pop C, Rouskin S, Ingolia NT, Han L, Phizicky EM, Weissman JS, Koller D. 2014. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Molecular systems biology* **10**: 770.

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM et al. 2014. RefSeq: an update on mammalian reference sequences. *Nucleic acids research* **42**(Database issue): D756-763.

Qian W, Yang JR, Pearson NM, Maclean C, Zhang J. 2012. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS genetics* **8**(3): e1002603.

Ramu H, Mankin A, Vazquez-Laslop N. 2009. Programmed drug-dependent ribosome stalling. *Molecular microbiology* **71**(4): 811-824.

Reid DW, Chen Q, Tay AS, Shenolikar S, Nicchitta CV. 2014. The unfolded protein response triggers selective mRNA release from the endoplasmic reticulum. *Cell* **158**(6): 1362-1374.

Rooijers K, Loayza-Puch F, Nijtmans LG, Agami R. 2013. Ribosome profiling reveals features of

normal and disease-associated mitochondrial translation. *Nature communications* **4**: 2886.
Rubio CA, Weisburd B, Holderfield M, Arias C, Fang E, DeRisi JL, Fanidi A. 2014. Transcriptome-wide characterization of the eIF4A signature highlights plasticity in translation regulation. *Genome biology* **15**(10): 476.
Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. 2013. Rate-limiting steps in yeast protein translation. *Cell* **153**(7): 1589-1601.
Shalgi R, Hurt JA, Krykbaeva I, Taipale M, Lindquist S, Burge CB. 2013. Widespread regulation of translation by elongation pausing in heat shock. *Molecular cell* **49**(3): 439-452.
Somogyi P, Jenner AJ, Brierley I, Inglis SC. 1993. Ribosomal pausing during translation of an RNA pseudoknot. *Molecular and cellular biology* **13**(11): 6931-6940.
Stadler M, Fire A. 2011. Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA* **17**(12): 2063-2073.
Stern-Ginossar N, Weisburd B, Michalski A, Le VT, Hein MY, Huang SX, Ma M, Shen B, Qian SB, Hengel H et al. 2012. Decoding human cytomegalovirus. *Science* **338**(6110): 1088-1093.
Stumpf CR, Moreno MV, Olshen AB, Taylor BS, Ruggero D. 2013. The translational landscape of the mammalian cell cycle. *Molecular cell* **52**(4): 574-582.
Tholstrup J, Oddershede LB, Sorensen MA. 2012. mRNA pseudoknot structures can act as ribosomal roadblocks. *Nucleic acids research* **40**(1): 303-313.
Thoreen CC, Chantranupong L, Keys HR, Wang T, Gray NS, Sabatini DM. 2012. A unifying model for mTORC1-mediated regulation of mRNA translation. *Nature* **485**(7396): 109-113.
Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010a. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**(2): 344-354.
Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. 2011. Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome biology* **12**(11): R110.
Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010b. Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences of the United States of America* **107**(8): 3645-3650.
Wei J, Wu C, Sachs MS. 2012. The arginine attenuator peptide interferes with the ribosome peptidyl transferase center. *Molecular and cellular biology* **32**(13): 2396-2406.
Woolstenhulme CJ, Parajuli S, Healey DW, Valverde DP, Petersen EN, Starosta AL, Guydosh NR, Johnson WE, Wilson DN, Buskirk AR. 2013. Nascent peptides that block protein synthesis in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **110**(10): E878-887.
Yanagitani K, Kimata Y, Kadokura H, Kohno K. 2011. Translational pausing ensures membrane targeting and cytoplasmic splicing of XBP1u mRNA. *Science* **331**(6017): 586-589.
Yang JR, Chen X, Zhang J. 2014. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS biology* **12**(7): e1001910.

Table 1. Ribosome profiling protocol conditions for the studies described in this work. CHX – cycloheximide, MN – micrococcal nuclease, GR – gradient, CS –cushion.

Description	PMID	SRA accession	Biological source	Lysis buffer				Lysis method	RNase	Separation	RNase digestion stage
				CHX pre-treatment, mins	Mg^{2+} mM	M^{+} mM	Drugs				
			Human								
Andreev	25621764	SRR1173909 SRR1173910	HEK293T	no	1.5	250NaCl	CHX	Detergent	I	GR	Lysate
Gonzalez	25122893	SRR1562539	Brain	no	15	250 NaCl	CHX	Dounce homogenizer, freeze	I	GR	Lysate
Guo	20703300	SRR057512	HeLa	8	5	100 KCl	CHX	Detergent	I	GR	Lysate
Heish*	22367541	SRR403883	PC3								
Lee	22927429	SRR618771	HEK293	30	5	100 KCl	CHX	Detergent	I	GR	Polysome
Liu	23290916	SRR619083	HEK293	1	5	100 KCl	CHX	Detergent	I	GR	Polysome
Loayzo-Puch	23594524	SRR627620	BJ fibroblast	8	10	100 KCl	CHX	Detergent	I	GR	Lysate
Rooijiers	24301020	SRR935448	BJ fibroblast	5	10	100 KCl	CHX	Detergent	I	GR	Lysate
Rubio	25273840	SRR1573934	MDA-MB-231	no	15	220 NaCl	CHX	Detergent	I	CS	Lysate
Shalgi	23290915	SRR648667	HEK293T	5	5	100 KCl	No	Freeze	I	CS	Lysate
Stadler CHX.	22045228	SRR407637	HeLa	no	1.5	140 KCl	CHX	Freeze	I	GR	Lysate
Stadleruntr.	22045228	SRR407643	HeLa	no	1.5	140 KCl	No	Freeze	I	GR	Lysate
Stern-Ginossar, CHX	23180859	SRR609197	human foreskin fibroblasts	1	15	250 NaCl	CHX	Detergent	I	CS	Lysate
Stern-Ginossar, untr	23180859	SRR592961	human foreskin fibroblasts	no	15	250 NaCl	No	Detergent	I	CS	Lysate
Stumpf*	24120665	SRR970561	Hela								
			Mouse								
Howard	23696641	SRR826795	Liver	no	10	300 KCl	CHX	Homogenizer	I	CS	Lysate
Ingolia, CHX	22056041	SRR315601	Embryonic stem cell	1	15	250 NaCl	CHX	Detergent	I	CS	Lysate
Ingolia, untr.	22056041	SRR315616	Embryonic stem cell	no	15	250 NaCl	No	Detergent	I	CS	Lysate
Reid	25215492	SRR1066893	Embryonic fibroblast	no	15	100 KoAc	CHX	Detergent (digitonine)	MN	CS	Lysate
Shalgi	23290915	SRR648667	3T3	5	5	100 KCl	No	Freeze	I	CS	Lysate
Thoreen	22552098	SRR449467	Embryonic fibroblast	5	7.5	300 KCl	CHX	Detergent	I	GR	Lysate
			Yeast								
Artieri	25294246	SRR1049093		2	1.5	140 KCl	CHX	Freeze	I	GR	Lysate
Brar	22194413	SRR387871		2	1.5	140 KCl	CHX	Freeze	I	GR	Lysate
Ingolia	19213877	SRR014374 SRR014375 SRR014376		2	1.5	140 KCl	CHX	Freeze	I	GR	Lysate
Lareau, CHX	24842990	SRR1363415 SRR1363416		yes	1.5	140 KCl	CHX	Freeze	I	GR	Lysate
Lareau, untr.	24842990	SRR1363412 SRR1363413 SRR1363414		no	1.5	140 KCl	No	Freeze	I	GR	Lysate
McManus	24318730	SRR948555		5	1.5	140 KCl	CHX	Freeze	I	CS	Lysate
Pop	25538139	SRR1688547		2	1.5	140 KCl	CHX	Freeze	I	CS	Lysate

* These authors did not provide protocol conditions in the original publications and did not respond to a specific query regarding the protocols used.

Figure legends

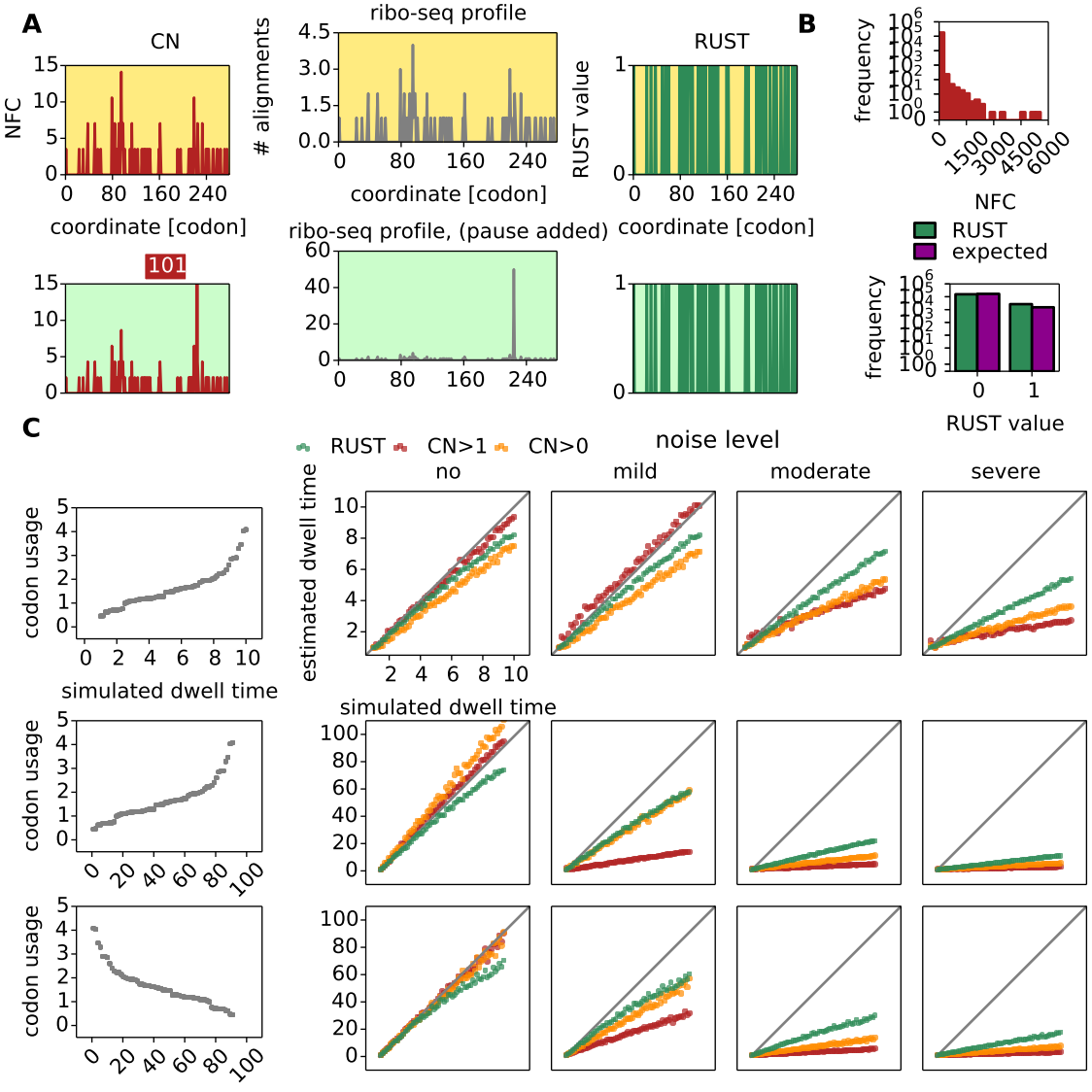
Figure 1. Shortcomings of conventional normalisation lessened with RUST. (A) Ribo-seq alignment profile (top center) and its normalisation profiles with either conventional normalisation (CN) (left) or RUST (right). The addition of one high density peak (bottom center) to the profile influences the normalised footprint counts (NFC) considerably with CN whereas this is less pronounced with the RUST protocol. The value of the peak produced with CN is indicated with text highlighted in red background. (B) The comparatively large influence of such peaks at individual mRNA sites is illustrated by the skewness of the NFC frequency distribution for the AAA codon with CN of real human data (top) whereas every site has an equal influence in the RUST frequency distribution (bottom). (C) Relationship between the actual and observed codon dwell times on simulated ribosome profiling data with three approaches; RUST, CN of transcripts with average gene density >1/nt (CN>1) and CN of all expressed transcripts (CN>0). The input parameters (relationship dwell times with codon usage as well as the fold difference of dwell times) are shown on the left, and the response to increasing noise in the data are on the right.

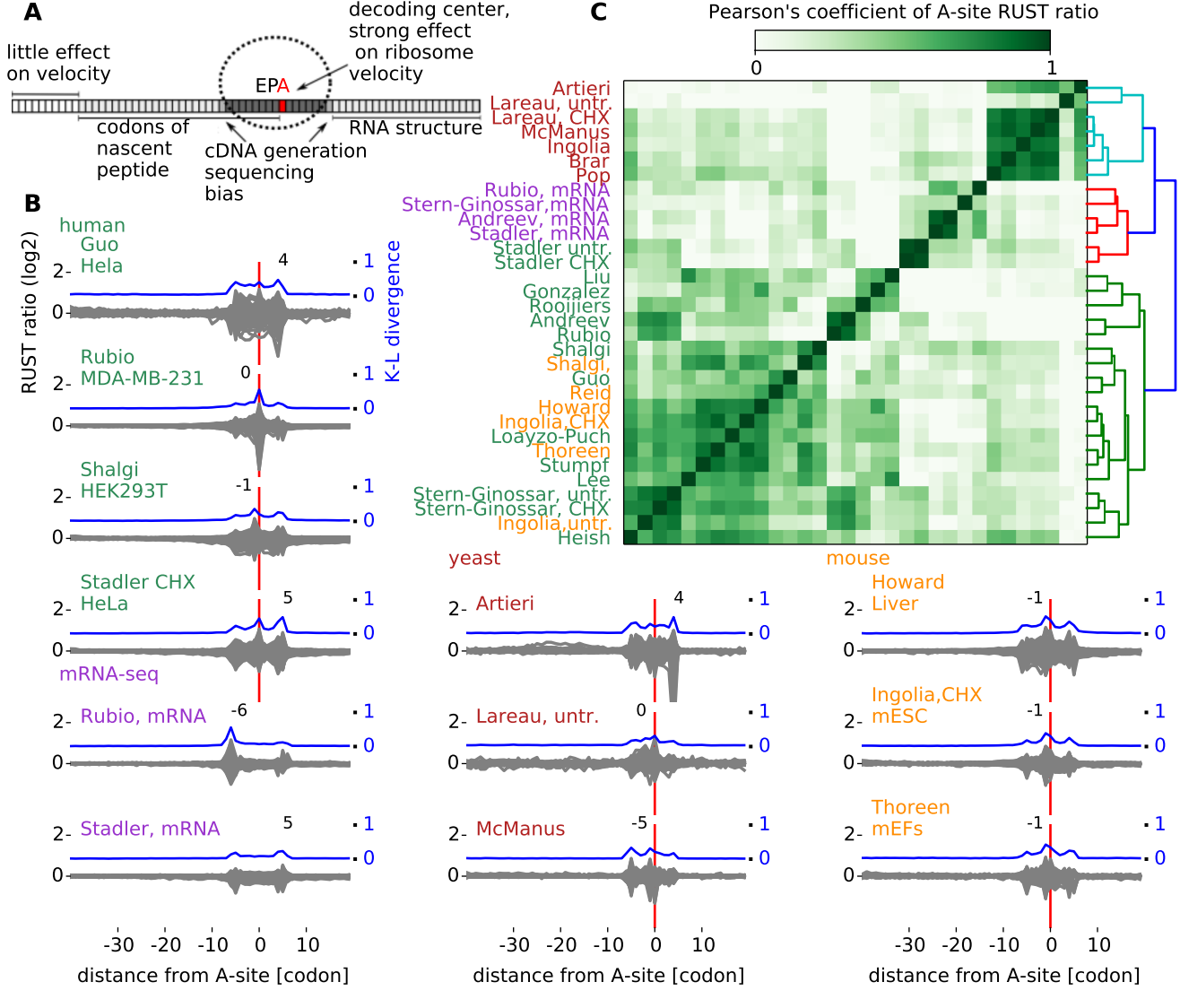
Figure 2. Evaluation of ribo-seq alignment data with RUST. (A) Anatomy of the ribosome footprint displaying position-specific mRNA influence on ribo-seq read density. (B) RUST codon metafootprint profiles of selected ribo-seq and mRNA-seq datasets used in this study. The individual RUST ratio values of 61 sense codons across the ribosome are displayed in grey. The corresponding Kullback-Leibler divergence (K-L) is shown in blue. The position relative to the A-site with the greatest K-L is also indicated. See Supplemental Figure S2 and S3 for the metafootprints of the other datasets. The datasets are described as in Table 1. (C) Heatmap displaying the pairwise similarity of codon RUST ratio at the A-site, as measured by the Pearson's correlation, for ribo-seq datasets of human (green), yeast (red) and mouse (orange). Also included are human mRNA-seq data (violet). The associated dendrogram was created with scipy using "Euclidean" distance metric with "average" linkage clustering.

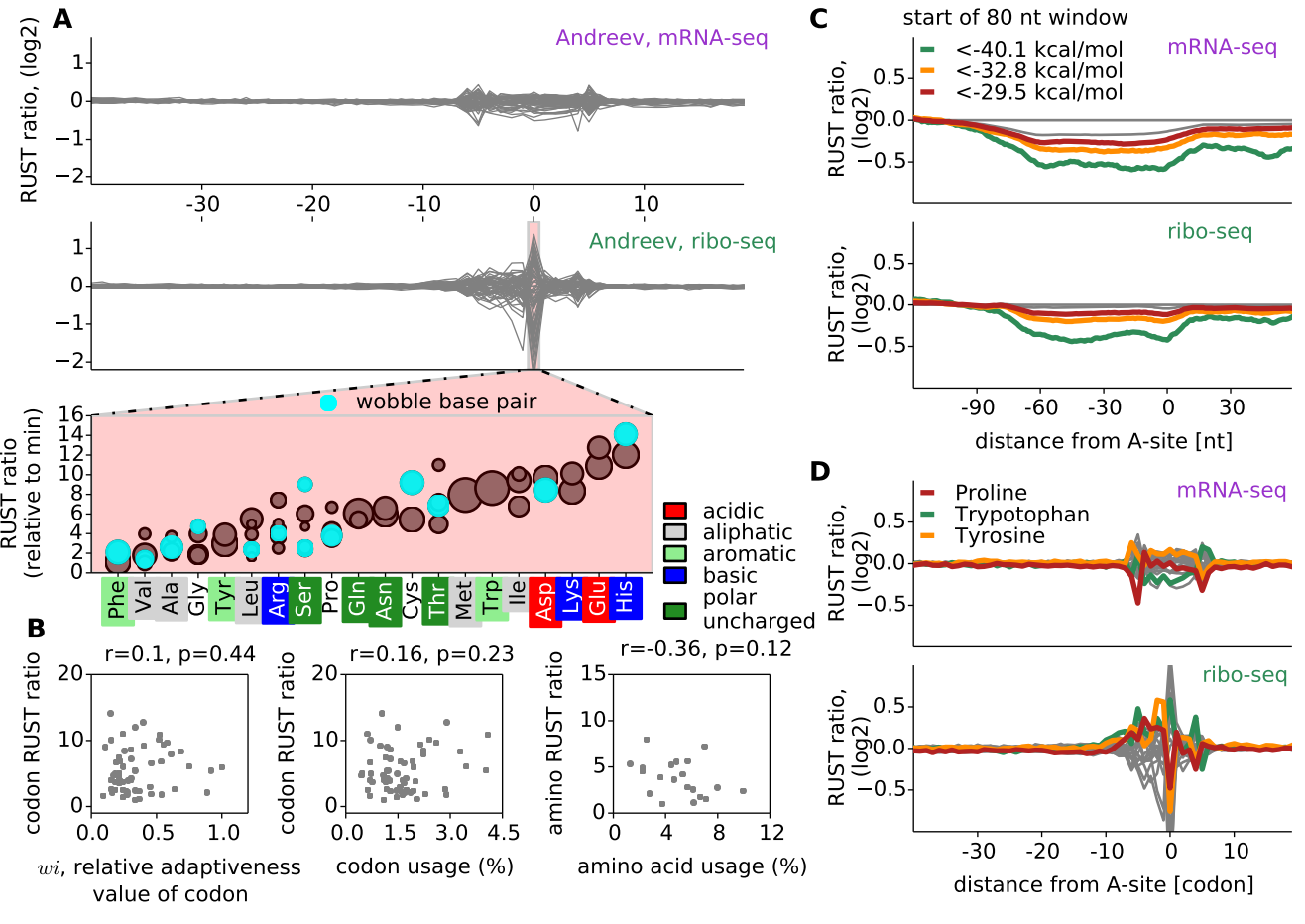
Figure 3. Detailed characterization of sequence-specific factors affecting the composition of ribo-seq libraries. (A) RUST codon metafootprint profiles of a selected dataset. The normalised RUST ratios of the A-site (below) is also displayed; codons are grouped according to the amino acid that they encode, those requiring wobble base interactions are displayed in cyan and their relative codon usage is indicated with the size of each dot. (B) Relationships between normalised A-site codon RUST ratios and relative adaptiveness of codons (*wi)* and codon usage. Also shown is the relationship between the normalised A-site amino acid RUST ratios and the amino acid usage. (C) The RUST metafootprint profile of regions predicted to contain strong RNA structures. Each position of the metafootprint profile indicates the start of a 80 nucleotides (nt) window. (D) RUST amino acid metafootprint profiles.

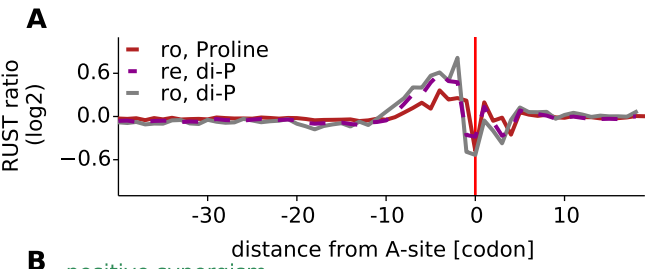
Figure 4. Synergistic effects of amino acids interactions within the nascent peptide. (A) Comparison of expected and observed metafootprint profiles at di-P with the observed profile of Proline. (B) Examples of strongest cases of positive or negative synergism detected with tripeptides. The position of the first amino acid of the tripeptide is indicated. (C) Stronger candidates of tripeptide synergism including the position of the synergetic interaction and the *ro/re* fold change (see Supplemental Table S1). (D) The relative frequencies of synergism detected across different positions of the ribosome. The position of the first residues is indicated (E) The fold change between the expected and observed RUST ratio for cases of synergism with tripeptides.

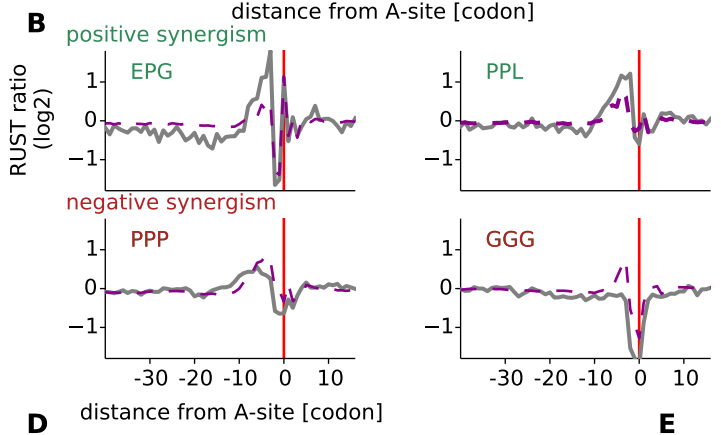
Figure 5. RUST parameters accurately predict experimental footprint densities. (A)Distributions of Pearson's correlation coefficients for experimental and predicted footprint densities for individual transcripts. Correlations were measured only for coding regions of highly expressed transcripts from 120 nucleotides downstream of the start codons upstream of the stop codons. (B) An example of experimental (solid grey) and predicted (based on RUST values for the codons -6 to +6 relative to A-site) ribosome densities (broken purple) for the same transcript.


A
CN
NFC
coordinate [codon]
ribo-seq profile
alignments
coordinate [codon]
RUST
RUST value
coordinate [codon]
101
ribo-seq profile, (pause added)
B
frequency
NFC
RUST
expected
frequency
RUST value
C
RUST
CN>1
CN>0
noise level
no
mild
moderate
severe
estimated dwell time
codon usage
simulated dwell time
simulated dwell time
codon usage
codon usage


A
little effect on velocity
decoding center, strong effect on ribosome velocity
EPA
codons of nascent peptide
cDNA generation sequencing bias
RNA structure
B
RUST ratio (log2)
K-L divergence
human
Guo
Hela
Rubio
MDA-MB-231
Shalgi
HEK293T
Stadler CHX
HeLa
mRNA-seq
Rubio, mRNA
Stadler, mRNA
distance from A-site [codon]
C
Pearson's coefficient of A-site RUST ratio
Artieri
Lareau, untr.
Lareau, CHX
McManus
Ingolia
Brar
Pop
Rubio, mRNA
Stern-Ginossar,mRNA
Andreev, mRNA
Stadler, mRNA
Stadler untr.
Stadler CHX
Liu
Gonzalez
Rooijiers
Andreev
Rubio
Shalgi
Shalgi,
Guo
Reid
Howard
Ingolia,CHX
Loayzo-Puch
Thoreen
Stumpf
Lee
Stern-Ginossar, untr.
Stern-Ginossar, CHX
Ingolia,untr.
Heish
yeast
Artieri
Lareau, untr.
McManus
distance from A-site [codon]
mouse
Howard
Liver
Ingolia,CHX
mESC
Thoreen
mEFs
distance from A-site [codon]

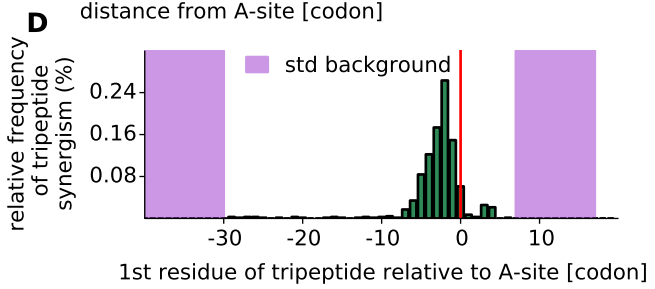

A
Andreev, mRNA-seq
RUST ratio, (log2)
Andreev, ribo-seq
wobble base pair
RUST ratio
(relative to min)
Phe
Val
Ala
Gly
Tyr
Leu
Arg
Ser
Pro
Gln
Asn
Cys
Thr
Met
Trp
Ile
Asp
Lys
Glu
His
acidic
aliphatic
aromatic
basic
polar
uncharged
B
r=0.1, p=0.44
r=0.16, p=0.23
r=-0.36, p=0.12
codon RUST ratio
amino RUST ratio
wi, relative adaptiveness value of codon
codon usage (%)
amino acid usage (%)
C
start of 80 nt window
<-40.1 kcal/mol
<-32.8 kcal/mol
<-29.5 kcal/mol
mRNA-seq
ribo-seq
RUST ratio, (log2)
distance from A-site [nt]
D
Proline
Trypotophan
Tyrosine
mRNA-seq
ribo-seq
RUST ratio, (log2)
distance from A-site [codon]

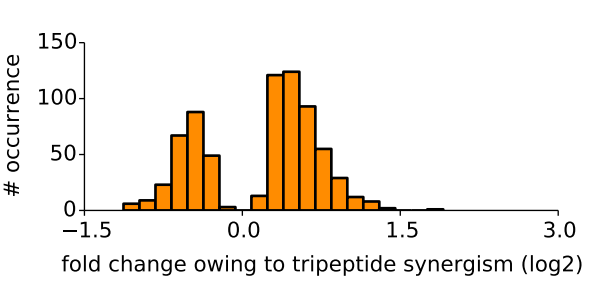


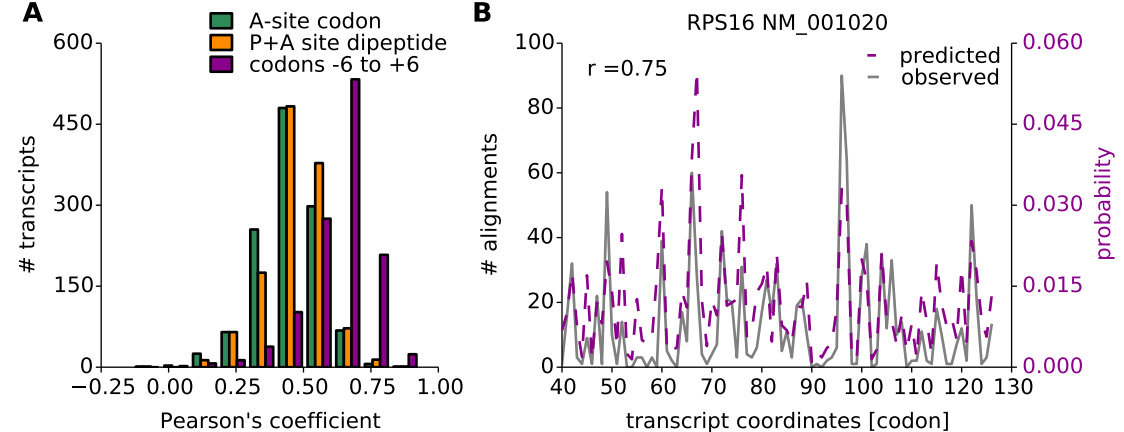


C

tripeptide	Sijk (Z-score)	1st residue relative A-site	ro/re fold change
LPP	23.8	-3	1.9
PPL	21.8	-2	2.5
DPG	21.0	-3	2.3
EPG	20.9	-3	2.6
RPP	17.6	-3	1.8
KPP	17.1	-3	2.4
RLK	17.1	-2	1.7
PPA	16.4	-2	2.2
PPE	15.8	-2	1.8
PEE	15.6	-2	1.5
PGP	-11.2	-3	0.7
GGG	-11.5	-4	0.6
PPP	-11.9	-3	0.7
PGP	-11.9	-4	0.7
AAA	-12.0	-4	0.7
PAP	-13.0	-3	0.7
PAP	-14.2	-4	0.7
PPP	-14.5	-4	0.7
LDE	-15.4	-1	0.6
GGG	-17.8	-3	0.5






A
A-site codon
P+A site dipeptide
codons -6 to +6
transcripts
Pearson's coefficient
B
RPS16 NM_001020
r =0.75
predicted
observed
alignments
probability
transcript coordinates [codon]