JOURNAL OF
EMERGING INVESTIGATORS

Article

Ribosome distribution affects stalling in amino-acid starved cancer cells

Mark Deng[1], Michel Nofal[2]
[1] Las Lomas High School, Walnut Creek, California
[2] Princeton University, Princeton, New Jersey

SUMMARY

Protein synthesis is a process central to all life on Earth, including mammalian cells. During this process, ribosomes attach to mRNA strands and translate them into proteins using amino acids. Under stress (for example, when the supply of circulating amino acids has been disrupted by tissue injury), ribosomes can stall. Ribosome profiling, a technique that creates a snapshot of active ribosomes in a cell by sequencing ribosome-protected mRNA fragments, captures a snapshot of ribosomes along transcripts and can detect such stalling events. This method has also revealed that the patterns of ribosomes along transcripts can vary from transcript to transcript in a manner that has not yet been explained. Here, we analyzed ribosome profiling data from amino acid-starved pancreatic cancer cells to explore whether the pattern of ribosome distribution along transcripts under normal conditions can predict the degree of ribosome stalling under stress. We hypothesized that ribosomes would stall more along "elongation-limited" transcripts that have fewer ribosome footprints near the start and stop codons than "initiation-limited" transcripts that have a large fraction of footprints at the start codon. Indeed, we found that ribosomes in amino acid-deprived cells stalled more along elongation-limited transcripts. By contrast, we observed no relationship between read density near start and stop and disparities between mRNA sequencing reads and ribosome profiling reads. This research identifies an important relationship between read distribution and propensity for ribosomes to stall, although more work is needed to fully understand the patterns of ribosome distribution along transcripts in ribosome profiling data.

INTRODUCTION

Protein synthesis is fundamental to life as we know it. It is ongoing in all mammalian cells, even those that are not growing and dividing. Because life is dynamic, proteins need to be synthesized continuously to accommodate cellular needs in a changing environment. Proteins are also biomolecules and are thus somewhat reactive. They can be damaged over time, so many proteins must be degraded and replaced. Proteins are synthesized by ribosomes, which use messenger RNA (mRNA) as a template. Within ribosomes, transfer RNAs (tRNAs) translate the 4-letter language of DNA and RNA into the 20-letter language of proteins by mapping each three-nucleotide codon to its corresponding amino acid. As tRNAs bind within ribosomes, the amino acid is transferred to the nascent polypeptide chain; this is how proteins are synthesized.

A common way to measure gene expression is through mRNA sequencing. However, mRNA levels do not necessarily reflect protein abundances, as some mRNAs may be translated at higher rates than others. Thus, there may be sizable discrepancies between mRNA abundances, which are relatively easy to measure, and true protein levels, which determine phenotype (1). A technique known as ribosome profiling measures translation of transcripts, not just transcript abundance, and may be a better tool to measure gene expression (2). Assuming protein degradation is roughly equal for all proteins, protein abundances should be closely related to translation and thus measurable by ribosome profiling (2).

Ribosome profiling captures a snapshot of the active ribosomes in a population of cells. Ribosomes are bound to a small piece of the mRNA strand that they are translating – approximately 30 bases. These stretches of mRNA are essentially protected by ribosomes, which can be frozen in place if a drug called cycloheximide is added. When a ribonuclease is added to the extracted cell lysate, it degrades all mRNA except the ribosome-protected "footprints". These footprints are subsequently converted into a DNA library that can be deep-sequenced. Through this deep-sequencing, the abundance of these different ribosome-protected fragments is measured. To date, most implementations of ribosome profiling have assumed that all footprints mapping to the same mRNA are equivalent – in other words, if two transcripts have the same density of ribosome footprints as measured by ribosome profiling, it is assumed that they are being translated at the same rate. However, closer investigation reveals that the distribution of footprints varies dramatically between transcripts (2). Here, we investigate these differences to further our understanding of ribosome profiling data and, more generally, translation in mammalian cells.

Translation, as opposed to transcription, is particularly important to study in settings where amino acids may be limited. This is because amino acids are direct substrates for translation (and not transcription). Should synthesis at some transcripts respond differently to amino acid starvation than synthesis at other transcripts, RNA sequencing would be an insufficient method of sequencing; ribosome profiling is needed. One example of an amino acid-deprived tissue is pancreatic tumor tissue, which is densely fibrotic. Amino acid scarcity limits protein synthesis. Each time a ribosome binds to an mRNA and initiates translation then becomes critical in terms of resource allocation. Too much translation initiation

can lead to ribosome stalling, which is toxic to the cell. Thus, although a ribosome may be bound to a specific mRNA, the translation of that mRNA may be paused because of the lack of amino acids in the cell. This inappropriate pausing due to lack of resources is referred to as ribosome stalling. The transcripts that are paused at start or stop codons are not considered to be stalled as they are not pausing in direct response to a lack of amino acids (3, 4)

In these amino acid-depleted environments, a specific kinase, GCN2, becomes much more important to the protein synthesis process. GCN2 kinase activity is stimulated by binding to uncharged tRNAs, which become more abundant under amino acid deprivation. GCN2 suppresses translation initiation in amino acid-deficient conditions, thus limiting protein synthesis at the beginning rather than during elongation. It adapts cells to an environment where essential amino acids levels are low by suppressing initiation of gene translation rather than allowing stalling in the middle of translation (5).

Because of the connection between ribosome stalling, translation initiation, and amino-acid allocation, we hypothesized that there are two general kinds of transcripts: transcripts whose rate of translation is primarily limited by translation initiation rate – "initiation-limited" transcripts – and transcripts whose translation is primarily limited by elongation – "elongation-limited" transcripts. These two kinds of transcripts can be identified using ribosome profiling data, which shows how many ribosomes are bound near start codons and how many are bound elsewhere. We expected that more stalling events would occur in elongation-limited transcripts relative to initiation-limited transcripts. Our analysis suggests that this is at least somewhat true. To confirm whether elongation-limited transcripts have more than the expected number of ribosomes, we compared the number of total reads mapping to these transcripts from ribosome profiling and RNA sequencing datasets. We expected that elongation-limited transcripts would have a higher ratio of ribosome profiling reads to RNA sequencing reads than their initiation-limited counterparts, but we found no evidence that this is the case. This relationship between read distribution and propensity for ribosomes to stall can be important in gathering a better understanding of the inner workings of a mammalian cell, but more research needs to be done in this area to fully understand the implication of this relationship.

RESULTS

To investigate whether patterns of ribosome profiling reads along transcripts can predict ribosome stalling, we used a dataset previously generated using a pair of murine pancreatic cancer cell lines (WT and Gcn2 knock-out) derived from spontaneously arising mouse tumors (6). The cell lines were cultured in amino acid-rich or leucine-free medium for one hour before ribosome footprints were extracted and analyzed by ribosome profiling (7).

The data from these experiments can be summarized either by gene (each row contains read counts in various samples for one gene) or by codon (each row contains read counts for that specific codon). Because only transcripts with many proper reads are useful when attempting to understand the patterns of reads along transcripts, it was necessary to filter out genes with i) misannotated start sites and ii) insufficient read depth – meaning there were too few reads to conduct proper analysis. We focused first on the data summarized by codon. To focus on transcripts with sufficient read density, we used only the top 1000 transcripts in terms of ribosome profiling reads out of the original ~10,000; this effectively removed all transcripts with insufficient read depth. Though high read density could be attributed to longer genes, these top 1000 transcripts had similar lengths to the original gene set (560.151 vs. 617.755), minimizing gene length as a confounding factor (**Figure 1**). We found that of these transcripts, many had improperly annotated start and stop sites. For example, no reads map to the annotated start of *Sdc4* (**Figure 2B**, right). This could be because the start codon is truly misannotated or because data processing that generated the data frame that we started with was faulty. It also could be because ribosomes move so quickly over the first few codons of *Sdc4* that the start codons are not detected by ribosome profiling. Using genes like *Sdc4* in an analysis involving read density at start as a key parameter confounds such an analysis because their start sites are likely to be misannotated. Of the 1000 top genes, 752 had correctly annotated start sites (**Figure 2A**, left), and 427 had correctly annotated stop sites (**Figure 2A**, right). After removing all but proper start and stop sites, we were left with 320 genes from the original top 1000 genes.

With a list of correctly annotated genes, we were able to make comparisons between the fraction of reads near the termini of the open reading frame and other aspects of the data. We took a table that included mRNA reads and ribosome profiling reads for each gene and measured the density of ribosome profiling reads near the start and stop codons by finding the percentage of all reads that mapped to one of the first five codons or last five codons. As an example, we delineate these boundaries for *S100a6*. *S100a6* has quite a high fraction of reads mapping near the termini of the open reading frame, with obvious accumulation of ribosome footprints at both start and stop codons (**Figure 3**).

To identify a potential relationship between accumulation near start and stop sites and stalling in amino acid-deficient conditions, we analyzed data from the *Gcn2* knock-out (KO) cells. GCN2 prevents stalling by slowing translation initiation when amino acids become scarce (5). When plotting *Gcn2* KO cell stalling against footprint accumulation near transcript termini in the amino acid-rich condition, we found that there is a modest but significant correlation between the two (**Figure 4**). The amino acid-rich condition serves as a baseline indicator of how much accumulation near the termini occurs when a cell

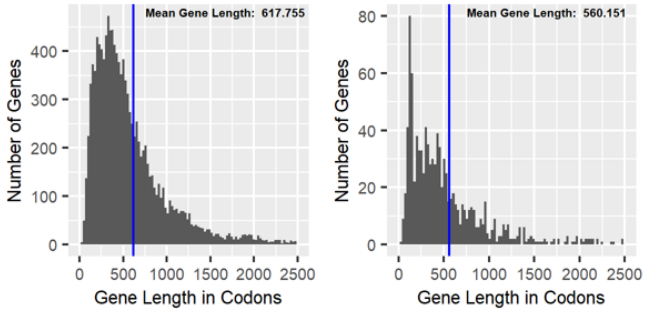

Figure 1: Gene lengths of the full gene set and the top 1000 genes selected for analysis. Histogram showing that gene lengths are similar in the original data set (left) and the truncated data set (right).

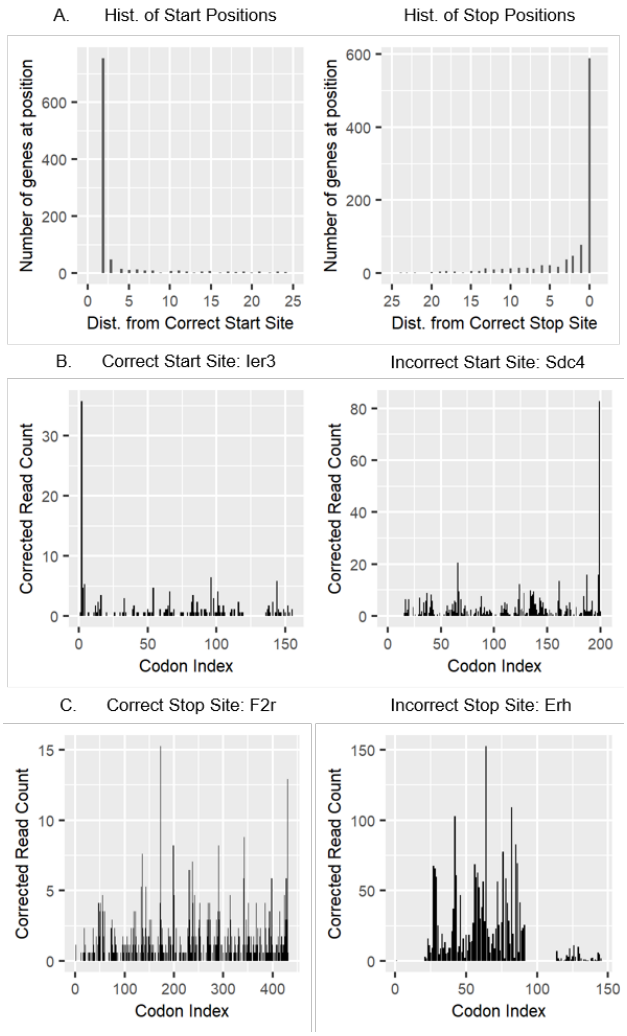

Figure 2: Figure 2: Correct and incorrect stop and start sites visualized. A) Histogram of start and stop codon site distribution. B) Plot of a correctly annotated start site (*Ier3*) and an incorrectly annotated start site (Sdc4). C) Plot of a correctly annotated stop site (*F2r*) and an incorrectly annotated stop site (*Erh*).

is in a healthy state. We used this condition because it reveals the ribosome binding pattern in cells that are unperturbed by nutrient deprivation. As more transcripts pause at start and stop sites, the stalling we observed in leucine-starved *Gcn2* KO cells decreased. Thus, GCN2 may be more important in protecting ribosomes that translate transcripts which are primarily elongation-limited rather than transcripts which are initiation or termination-limited.

We further analyzed this data by isolating the reads near the start and performing a correlation analysis. Here, the correlation was much clearer (**Figure 4A**). Though the Pearson Correlation coefficient is relatively low (R^2 = 0.0422), the statistical significance is clear as p = 0.0008. However, when this analysis was performed using only the reads near stop, there was no clear correlation with a low $R2$ value and no statistical significance (p = 0.638) (**Figure 4B**). This shows that transcripts with a higher fraction of ribosomes pausing at the start codon tend to have less stalling, while ribosomal pausing near the stop codon does not correlate with ribosomal

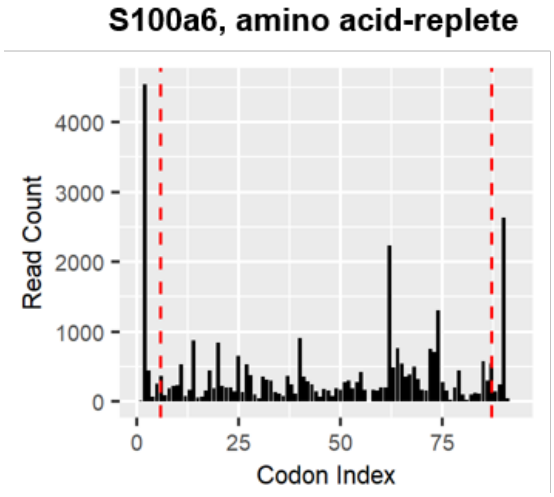

Figure 3: Graphical representation of start five and last five groups used in calculation. The example gene plotted is *S100a6*, and the data shown is from GCN2 WT cells in amino acid-rich medium.

stalling.

We also conducted a similar analysis of the *Gcn2* WT cell line and determined that there was no correlation (R^2 = 0.0019) between reads near stop or start and ribosomal stalling in these cells (**Figure 4C**). This indicates that the statistically significant correlation between the fraction of reads at the start site of *Gcn2* KO cells and ribosomal stalling arises only in cells without GCN2.

We next sought to investigate the differences in fractional read counts (reads per kilobase million) across sequencing experiments. More specifically, we noticed that for any given gene in any given cell line and condition, the number of reads per million mapping to that gene sometimes varied dramatically between RNA sequencing and ribosome profiling – an effect which, to our knowledge, has not been explained. A read number that is much lower in ribosome profiling compared to RNA sequencing could potentially indicate that a transcript is translated at a rate much lower than average, and vice versa. We hypothesized that this behavior related to the pattern of ribosome footprints along transcripts. Leveraging the analysis above, we decided to compare the fraction of reads mapping near start and stop codons and the discrepancies between the absolute number of sequencing reads in RNA sequencing versus ribosome profiling. To do this, we calculated the difference between the ribosome profiling reads and the RNA sequencing reads for each gene in the amino acid-replete conditions (sum of WT and KO) and plotted that against the read density near start and stop. This analysis revealed that the fraction of reads near transcript termini has no obvious relationship to the difference between ribosome profiling reads and RNA sequencing reads: the R^2 value is low (R^2 = .036) and not statistically significant (p = .1364) (**Figure 5**). Thus, the dramatic discrepancies that we observe between RNA sequencing reads and ribosome profiling reads cannot be explained by this study.

DISCUSSION

In this study, we explored the patterns of ribosome distribution along transcripts and how they relate to the

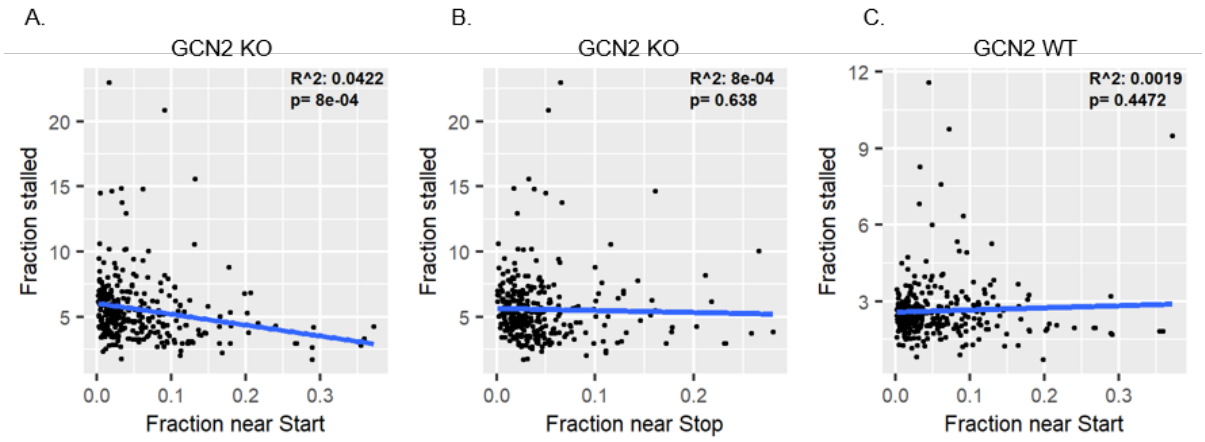

Figure 4: Ribosomal profiling reads compared to fraction of reads near transcript termini in both KO and WT cell lines. Ribosomal profiling reads in GCN2 KO cells after one hour of leucine deprivation is plotted against fraction of reads near start codons (A) and stop codons (B). We repeated this analysis only for start codons in GCN2 WT cells (C). We measured the fraction of reads near start and stop in amino

degree of ribosome stalling under amino acid deprivation in pancreatic cancer cells. First, we isolated a subset of highly expressed transcripts whose start and stop sites were properly annotated. We used this subset of genes to explore the idea that transcripts with higher read densities near start and stop codons (initiation-limited transcripts) might be less prone to stalling than transcripts where the majority of reads map in the middle (elongation-limited transcripts). Indeed, we determined that stalling occurs more often along transcripts which have a low fraction of reads mapping near start codons. The strength of the relationship was not strong, suggesting that there are other important factors at play. We found that this relationship was only true in amino acid-starved *Gcn2* KO cells (not *Gcn2* WT cells), and we found no such relationship between stalling and reads mapping near stop codons. In addition, we observed no relationship between read density

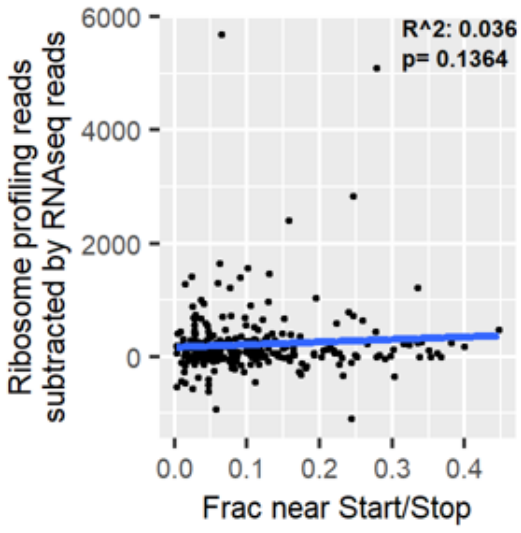

Figure 5: Ribosomal profiling reads compared to RNA sequencing reads. Differences between ribosome profiling reads and RNA sequencing reads for each gene are plotted against the fractions of reads near the start and stop codons of each of those genes. No apparent relationship was observed.

near the termini and disparities between mRNA sequencing reads and ribosome profiling reads.

The idea that initiation-limited transcripts are less prone to stalling than elongation-limited transcripts sheds light on the inner workings of translation in mammalian cells. However, given that the correlation is quite weak, there may well be other factors that contribute to making transcripts prone to stalling. For example, each species of mRNA has a unique localization within cells. Additionally, the density of all required factors for translation around each mRNA species, such as charged tRNAs, initiation factors, and elongation factors, may be different. Further, secondary structures within mRNAs may lead to intentional pauses to allow the protein to fold or the mRNA to be translocated to the endoplasmic reticulum, for example; such mid-transcript pauses would be read out as stalling events, contributing to noise in this data set. Finally, stalling could be the result of irregularities in sample preparations rather than real biological stalling. In other words, ribosomes could "slip" after cells are lysed, creating the appearance of stalling (3). To truly understand stalling events, further investigation accounting for these kinds of factors is required. Alternatively, deeper and higher quality ribosome profiling data may enable better analyses without consideration of all of these factors.

We also analyzed whether there exists a relationship between the difference between ribosome profiling reads and RNA sequencing reads and the degree of ribosome accumulation near the termini. We found that the ribosome profiling data we analyzed does not support a connection between these two phenomena. Thus, the origins of both the variation in the patterns of ribosome footprints along transcripts and the variation between reads per million between RNA sequencing and ribosome profiling cannot be addressed by the results of this study. Further research is required to better understand these mysteries. Because this experiment excluded a large number of genes, it is possible that, if more than just ~350 genes were included in such an analysis, the results could vary. In order to include more genes, the experimental aspects of the ribosome profiling protocol must also be improved, or otherwise deeper sequencing is needed.

Another experiment to supplement this analysis would be to repeat the above analysis in different cell lines. This analysis solely focused on murine pancreatic cancer cell lines. However, a similar experiment could be performed using a variety of different cell lines, including human cell lines originating from different tissues. This would reveal the way in which different types of cells respond to amino-acid deprivation, and consequently, the way ribosome distribution can affect stalling in different cells.

An experiment involving the drug harringtonine could also prove interesting. Harringtonine, another protein synthesis inhibitor, stalls ribosomes only at the start. If harringtonine is added to cells a few minutes before lysis, ribosomes would accumulate at start codons but "run off" the rest of transcripts. If cells are lysed and ribosome profiling is conducted at a series of time points after harringtonine addition, one could calculate their speed in a transcript-specific manner (8). Per-transcript elongation rates could be instrumental in interpreting ribosome profiling data. Eventually, we hope to understand ribosome profiling data well enough to translate it into per-transcript protein synthesis rates, and this work was one step in that direction.

MATERIALS AND METHODS

The computational analysis in this experiment was done using the computer programming and modeling language, R. Specifically, we used the R package dplyr and the ggplot function.

Data Generation

The experiment below was repeated in WT cells and cells in which *Gcn2* was knocked out. Each of these cell lines were propagated in amino acid-rich conditions, then switched to either amino acid-rich medium (+AA) or identical medium lacking the essential amino acid leucine (-Leu). Both media were supplemented with 50 g/L bovine serum albumin, which allows the leucine-starved cells to synthesize protein – albeit at a much slower rate – by taking up this albumin from the medium and degrading it within lysosomes (6). Leucine deprivation is a model for amino acid deprivation as a whole and enables cell growth, unlike complete amino acid deprivation (3). After an hour in their respective media, cells were lysed in the presence of cycloheximide, a commonly used protein synthesis inhibitor, mRNA-bound ribosomes were extracted, and ribosome profiling libraries were prepared and sequenced (7). In addition, standard mRNA sequencing was performed using these lysates. The read metric used was reads per kilobase million (RPKM).

Determining Correct Start Sites

To systematically identify genes with potentially misannotated start sites, we exploited the fact that ribosomes typically accumulate at start codons. We reasoned that the absence of such an accumulation may indicate a misannotation. We calculated the average number of reads at each codon in a gene transcript by taking the sum of reads in the Aminoacyl-tRNA site (A-site) of the ribosomes and dividing by the gene length in codons. We mapped the A-site by using the psite function of the plastid package (9). We then identified the first codon of every gene with more reads than this "codon average". Because it is expected that there is a peak in reads at the start, the first codon where the A-site reads are greater than the codon average should be the second codon in the sequence and the codon before that should be the true start codon. The reason that the first codon with many reads is the second codon, as opposed to the first, is because the start codon sits in the Peptidyl-tRNA site (P-site) during translation initiation, and when the start codon is in the P-site, the second codon is in the A-site. In contrast, we expect stalling to occur at the A-site, which is where charged tRNAs bind when available. We also removed any gene whose first codon (presumably the start codon) was not one of the four possible start codons (AUG, CUG, GUG, and UUG).

Determining Correct Stop Sites

Similarly, for some transcripts, the stop codon was misannotated. For example, *Erh* has an improperly annotated stop site, giving the inaccurate appearance of nearly half the gene as having little to no reads, while the true stop codon was further upstream (**Figure 2C**). This data, which we removed, gives the impression that there are leucine codons where no stalling happens and thus interferes with the analysis. We removed all genes without an accumulation of ribosomes at the annotated stop codon as well as those with incorrect codons in that position (the stop codons are UAA, UAG, and UGA).

Removing Incorrect Start and Stop Sites

To remove the incorrect start and stop sites, we created a threshold of A-site ribosome profiling reads. This threshold was the average codon reads across the gene. We then compared each codon starting from the front and back until we reached a codon where the A-site reads was greater than the average codon reads. This became the second site due to the start codon binding directly to the P-site. From the back, the first codon with A-site reads above the average became the stop codon. Any genes that did not match these criteria were eliminated. Any determined start or stop codons that did not have the correct base pairs were also eliminated. This left us with 752 correct start sites and 427 correct stop sites. The intersection of these two sets left use with 320 out of the original 1,000 genes.

Calculating GCN2 Knock-out stalling

To calculate the KO stalling (**Figure 4**), we determined the two leucine codons where stalling most likely occurred (CUC, CUU). Then we calculated the ratio of reads of these codons from the KO cell line and the WT cell line by dividing the KO reads by the WT reads. This gave us a normalized value for the quantity of stalling in each gene after GCN2 was knocked out.

Calculating *p*-values

To calculate the significance of the R^2 values for the plots in **Figure 4** (plotting GCN2 KO stalling against fraction of reads at start or stop), we first generated 5000 randomized samples of the KO stalling data. We then calculated the R^2 value for each sample (randomized KO stalling against original fraction of reads at start or stop) and determined the fraction of the 5000 that were lower than our original R^2 value. If that fraction (the *p*-value) was lower than 0.05, then we considered it statistically significant.

ACKNOWLEDGEMENTS

I would like to thank the Polygence program for providing me with the opportunity to work with an experienced mentor. I would also like to thank the Harvard Journal of Emerging Investigators for providing a platform to share my work.

Received: October 29, 2020
Accepted: May 24, 2021
Published: January 07, 2022

REFERENCES

1. Liu, Yansheng, *et al*. "On the Dependency of Cellular Protein Levels on MRNA Abundance." *Cell*, vol. 165, no. 3, Elsevier Inc., 2016, pp. 535–50, doi:10.1016/j.cell.2016.03.014.
2. Ingolia, Nicholas T., *et al*. "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling." *Science*, vol. 324, no. 5924, 2009, pp. 218–23, doi:10.1126/science.1168978.
3. Kamphorst, Jurre J., *et al*. "Human Pancreatic Cancer Tumors Are Nutrient Poor and Tumor Cells Actively Scavenge Extracellular Protein." *Cancer Research*, vol. 75, no. 3, 2015, pp. 544–53, doi:10.1158/0008-5472.CAN-14-2211.
4. Nedialkova, Danny D., and Sebastian A. Leidel. "Optimization of Codon Translation Rates via TRNA Modifications Maintains Proteome Integrity." *Cell*, vol. 161, no. 7, The Authors, 2015, pp. 1606–18, doi:10.1016/j.cell.2015.05.022.
5. Ye, Jiangbin, *et al*. "The GCN2-ATF4 Pathway Is Critical for Tumour Cell Survival and Proliferation in Response to Nutrient Deprivation." *EMBO Journal*, vol. 29, no. 12, Nature Publishing Group, 2010, pp. 2082–96, doi:10.1038/emboj.2010.81.
6. Nofal, Michel I., *et al*. "GCN2 Adapts Protein Synthesis to Scavenging-Dependent Growth." Cell Systems (in press).
7. McGlincy, Nicholas J., and Ingolia, Nicholas T. Transcriptome-wide measurement of translation by ribosome profiling. Methods. 2017 Aug 15;126:112-129. doi: 10.1016/j.ymeth.2017.05.028.
8. Ingolia, Nicholas T. *et al*. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nature Protocols. 2012 Jul 26;7(8):1534-50. doi: 10.1038/nprot.2012.086.
9. Dunn, Joshua G., and Jonathan S. Weissman. "Plastid: Nucleotide-Resolution Analysis of next-Generation Sequencing and Genomics Data." *BMC Genomics*, vol. 17, no. 1, BMC Genomics, 2016, pp. 1–12, doi:10.1186/s12864-016-3278-x.