מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Thesis for the degree

Master of Science

Submitted to the Scientific Council of the

Weizmann Institute of Science

עבודת גמר (תזה) לתואר

מוסמך למדעים

מוגשת למועצה המדעית של

מכון ויצמן למדע

By

Tehila Sima Leiman

מאת

תהילה סימא ליימן

מיפוי טעויות תרגום ברקמות אדם ועיצוב רצף רנ"א של חיסון רב-אנטיגני

Mapping of translation errors in human tissues and development of a multi-antigenic RNA vaccine sequence

Advisor: Prof. Yitzhak Pilpel

מנחה: פרופ' יצחק פלפל

27/04/2023

ו' אייר ה'תשפ"ג

Abstract

I investigated the misincorporation of amino acids during translation errors in human tissues. Using a bioinformatics approach, I compared the pattern of translation errors in humans to those in other organisms and found differences, indicating that the pattern of mistranslation is unique to each organism.

Furthermore, I observed that there are differences in the rate and types of substitutions between different tissues, suggesting that mistranslation is not a uniform process in the human body. However, I also discovered that proteins carrying errors tend to appear with errors in different tissues, indicating that specific factors contribute to the occurrence of mistranslation in these proteins, and that mistranslation is not a random process.

Consistent with previous research, I found that mistranslation tends to occur in less conserved parts of proteins across species, suggesting that there may be evolutionary pressures that allow for the occurrence of mistranslation without negatively impacting protein function.

Finally, I also aimed to create multi-strain RNA vaccine to target multiple variants of the virus using synonymous mutations to direct for specific amino acids substitutions to occur and activate the immune system against more than one virus strain. This goal still requires further development, but I did discover amino acid substitution in RNA vaccine protein product in model human cells.

Table of Contents

List of abbreviations

HPA: Human Proteome Atlas

LC-MS: Liquid chromatography-mass spectrometry

LFQ: Label free quantification

BP: Base peptide

DP: Dependent peptide

NeCE: Near cognate substitution

NeCE: Non-cognate substitution

TEL: Translation error low intensity ratio

SNP: Single-nucleotide polymorphisms

tAI: tRNA Adaptation Index

AlaRS: Alanyl-tRNA synthetase

R4S: rate4site score

MSA: multiple sequence alignment

SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2

COVID-19: Coronavirus disease 2019

VOC: variants of concern

MHC: major histocompatibility complex

Introduction

Protein translation is a fundamental process for all living organisms and it plays a crucial role in maintaining all cellular functions. This process involves the interaction of transfer RNA (tRNA) molecules with ribosomes to produce a polypeptide chain according to the genetic code. However, despite the specificity of the genetic code, errors can occur during the translation process, resulting in the incorporation of an incorrect amino acid. The phenomenon of codons being occasionally translated with a different amino acid than that specified by the codon has been previously observed and characterized in researches[1–3], including systematic research in Escherichia coli (E. Coli) in our laboratory[4]. Interestingly, this research has shown that different codons may have varying tendencies to be translated wrongly to different amino acids, and that these substitutions are more frequent in non-conserved regions of proteins. This observation raises important questions about the nature of these substitutions and whether they result only in deleterious effects on cells or if they may provide an advantage. At the heart of this question is the trade-off between the cost of proofreading and the potential benefit of tolerating these errors. It is possible that these substitutions are simply the result of errors, and the cell tolerates them because the cost of proofreading is too high. Alternatively, it could be a directed process that provides the cell with an advantage in specific contexts, such as under stress conditions as was occasionally demonstrated[5–8].

Previous studies showed that translation errors occur much more frequently than mutations or errors in transcription. Ernest et al. was the first to systematically measure error rate across all E. coli proteins. The values he measured varied between 10^{-3}-10^{-4}, i.e. one mis-incorporation for 1,000-10,000 amino acids, other estimations for translation error rate were conducted over specific codons or proteins, using different methods and across species. Overall, translation error rate varies between 10^{-3}-5×10^{-6} for all mentioned variables[1,3,9,10].

As mentioned before, codons, also for same amino acid, may have different tendency to result in translation error toward each amino acid[1]. It has been shown that even the same codon within the same open reading frame (ORF) can have a different tendency to result in non-

cognate amino acid[11]. This hints that the chemical properties of the codon and its surroundings can influence the frequency and the destination of the translation error.

The main objective of this study is to enhance our comprehension of translational errors and investigate their occurrence within the human context. To achieve this goal, I examined Liquid Chromatography-Mass Spectrometry (LC-MS) data from 29 diverse human tissues that were obtained from the Human Proteome Atlas (HPA)[12]. My laboratory's tools were utilized, with adjustments made to accommodate human data, and new branches of the pipeline were created to analyze more portions of the data. However, I faced difficulties when trying to apply the pipeline designed for E. coli to human data. A significant challenge arose from the fact that human cells are diploid (except for the germ line cells, which were not analyzed here). As a result, mutations can be detected as translational errors because reference proteomes do not account for these mutations. To resolve this, I compared the translational errors discovered to mutations identified in the same sample to estimate the proportion of results that were due to mutations. This allowed me to estimate the frequency and characterize the pattern of translation errors and compare them to what is known in literature. I also explored whether these errors occurred randomly or in specific proteins and sites. To do this, I examined the identity of proteins with translation errors across different tissues and evaluated the conservation of positions with translation errors. The findings suggested that certain proteins and sites were more prone to errors than others.

In a more applied aspect of the project, I examined the possibility to harness knowledge on translation error patters to design a new version of the SARS-CoV-2vaccine that will provide immunity towards several strains of the virus at once. Specifically, I was aiming to design a single sequence of a vaccine that can be translated to several Spike protein variants with substitutions similar to virus mutations can cover many variants of the virus at once. The emergence of the COVID-19 pandemic caused by the SARS-CoV-2 virus and the subsequent development of a vaccine, which uses mRNA coding for the virus Spike protein, presented a unique opportunity for investigation. One of the characteristics of the SARS-CoV-2 is it has variants of concern (VOC), meaning potentially dangerous mutants of the virus. This could make it harder to target the virus with a single vaccine.

In past research[1] it was found that different codons can have different rates of translation errors, which result in different amino acids being incorporated into the final protein. By designing the RNA sequence to include synonymous mutations for codons with specific profile of translation errors, the resulting protein could potentially contain multiple variants of the virus, exposing the immune system to a wider range of potential mutations that could arise in the future. By introducing these potential mutations, the mRNA vaccine could offer better protection against a wider range of viral variants, improving its overall efficacy against the virus.

Here I focused on mRNA vaccines for COVID-19. Two different vaccines were developed separately by Pfizer-BioNTech and Moderna, and they work by introducing the genetic sequence for the spike protein of SARS-CoV-2 into the body, which then instructs the cells to produce the protein. This triggers an immune response, where the immune systems create antibodies against the spike protein, preparing it to fight against the actual virus upon future exposure.

Peptidomics is a powerful tool for analyzing the presentation of peptides (short chains of amino acids) on major histocompatibility complex (MHC) molecules, which are central to the immune response. When a virus infects a human cell, its proteins are often processed into peptides and presented on the surface of the cell, bound to MHC molecules. These peptide-MHC complexes serve as the target of the immune system, which recognizes them as foreign and generates an immune response.

In the context of a vaccine, the goal is to elicit an immune response to a specific viral peptide. By using peptidomics to analyze the presentation of viral peptides on MHC molecules, I could gain insight into which peptides are likely to be most immunogenic and therefore effective targets for a vaccine. In this case, detecting translation errors in peptides presented on the MHC can help identify peptides generated with translation errors. By targeting these peptides, a vaccine may be able to elicit a more robust immune response that covers a wider range of viral strains.

Goals

- Comprehensively explore the landscape of translation errors in humans by identifying the specific types of amino acid substitutions that occur during translation, their frequency, and their distribution across different tissues.
- Compare the patterns of translation errors observed in humans with those seen in other organisms to gain insight into the evolutionary conservation of translational fidelity.
- Determine whether translation errors occur randomly throughout the proteome or if there are specific proteins and sites that are more susceptible to misincorporation.
- Use predictions of translation errors to design RNA vaccine that could target multiple strains of SARS-CoV-2, enhance the immune system response to various targets, and even the immune system response power.

Results

Quantifying translation errors in human and comparing to other organisms

I analyzed proteomics data (see Methods) from 29 different human tissues that were taken from different healthy participants[12]. I found 10,375 candidates for translation errors across all the tissues studied. Out of these candidates, I was able to identify 5,510 unique instances of translation errors, meaning that I found unique instances of codon to amino acid substitution within specific positions of the proteins in question.

Quantification of proteins in LC-MS

Protein quantification in LC-MS poses significant challenges due to several factors. The limited dynamic range of LC-MS makes it difficult to detect proteins present at low or high concentrations accurately. Additionally, proteins ionize differently depending on their physicochemical properties, leading to varying ionization efficiencies that can affect quantification. Co-elution of proteins during chromatographic separation, complex mixtures of biological samples, and post-translational modifications further complicate the quantification of individual proteins. These challenges make it difficult to compare protein abundance between different samples accurately in label free quantification (LFQ)[13].

To quantify errors in proteins and understand what their abundance is, I wanted to ask first which way is best to quantify proteins. Spectral counting and intensity-based approaches are two label-free quantification methods used in LC-MS to estimate the abundance of proteins in a sample. Spectral counting estimates protein abundance by counting the number of MS/MS spectra matching peptides from a protein, while intensity-based approaches use the intensity of the MS signal for identified peptides. To test which one fits better to our pipeline, I ran the pipeline on proteomics raw files from A375 cells from Yardena Samuels' Lab. Then for each peptide detected I added annotation whether it is closer to C or N terminal of the protein. If there are no biases, the amount of C terminal of a protein should be equal to the amount of N terminal. I then chose only peptides that were uniquely assigned to one protein only – 695 proteins in total. Then for each terminus I estimated the amount of protein by different approaches: spectra count, max intensity, or sum intensity. For spectra count, I counted the

number of spectra mapped to each protein terminus. For max intensity I used the value of intensity of the spectra with the highest intensity. The rationale is that if peptide were found with specific intensity, it could be underestimated but not overestimated, so the peptide with the highest intensity would be the best estimator of the protein's abundance. Lastly, sum of intensities of all peptides would give an estimator that incorporates both the intensities and the count of spectra per terminus of protein.

Out of the three approaches, quantifying by spectra showed the best match between the termini of the protein, with slope coefficient of 0.6, and Pearson correlation of 0.59. The sum of the intensities had worse correlation between the termini, but visually it seems that it is more accurate the higher the values are.

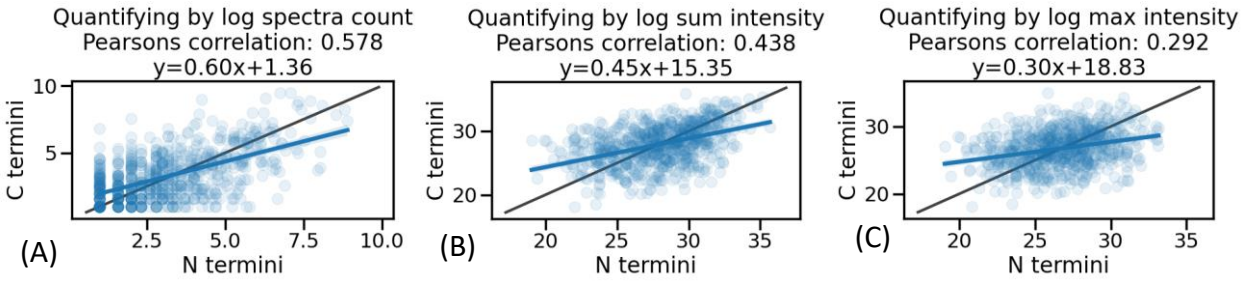

Figure 1: Counting spectra of protein is a better way to quantify than taking the sum or the max of intensities. Regression plots show each protein as a dot, with the values of spectra count (A), sum intensities (B) or max intensities (C) for N and C terminus of the protein in axes X and Y respectively. 695 proteins in total in each plot. Blue line is the linear regression model fit. Black line is x=y line.

Keeping this result in mind, I still used different approaches to quantify translation errors.

DP/BP intensity ratios of peptides with translation errors are different than in other organisms

An estimator for the substitution rate for each translation error detected is the ratio of intensity of base peptide (BP) and the intensity of the dependent peptide (DP) that can be written as $Intensity_{DP}/Intensity_{BP}$. This estimator can be biased because of potential difference in ionization efficiency between peptides that are composed of different amino acids. The ionization efficiency of a peptide can vary depending on its composition, and differences in ionization efficiency can range from as much as one order of magnitude between completely different tryptic peptides[14] to less for peptides that differ by a single amino acid. This difference in ionization efficiency can impact the accuracy of the estimator for the substitution rate by

affecting the intensity of the peptides in question. For example, if the ionization efficiency of a peptide containing a translation error is significantly lower than that of the corresponding peptide without the error, the estimator will underestimate the frequency of the error in the population. However, we considered this bias when we analyzed the results.

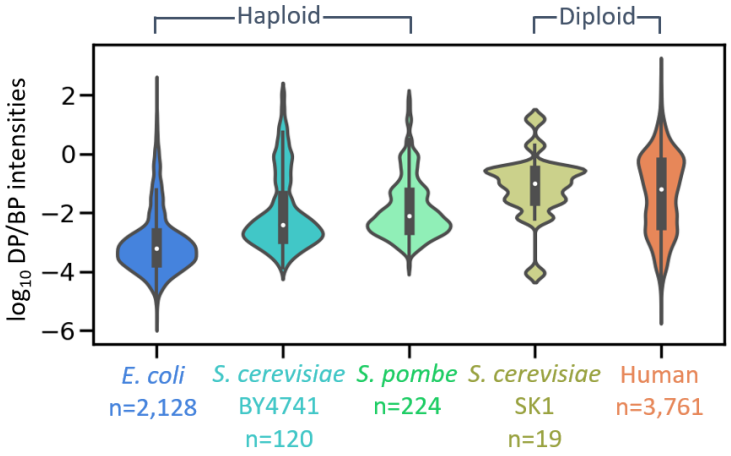

Figure 2: Distributions of the DP/BP intensity ratio, as measured by proteomics data, across E. coli, S. pombe, different strains of S. cerevisiae, and 29 human tissues. The x-axis represents each source, while the y-axis shows the frequency of each log10 ratio in the dataset.

I used proteomics data from various sources to compare the translation error rate between *E. coli*[4], *S. pombe*[15], 2 different strains of *S. cerevisiae* – BY4741[15] and SK1[16] – and substitutions observed in all 29 human tissues – see *Figure 2*. See Methods for intensity ratios calculation.

The distribution of the DP/BP intensity ratio in human is bimodal and has the highest mode among all species studied. The observed ratio of DP to BP intensity in many human translation errors is slightly lower than 1, which is highly unusual based on current knowledge of translation errors and raises suspicion. This may be due to variations between alleles caused by mismatch SNPs or mutations. The second highest mode is observed in diploid yeast, which also support this explanation. Ionization efficiency cannot fully explain the orders of magnitude difference between humans and haploid organisms, nor can it account for why the high peak in the human data is slightly below 1. Additionally, there is no apparent reason ionization efficiency would vary between humans and other species, or why the DP intensity is consistently lower than the BP intensity.

To determine if the substantial increase in human data is due heterozygosity in homologous chromosomes, I examined the coincidence of translation error sites with SNP sites. The list of SNPs found in tonsil tissue obtained through exome sequencing, which was obtained from the same source as the mass spectrometry (MS) data, was published alongside the MS data (there was no such data on SNPs for the other 28 tissues and organs).

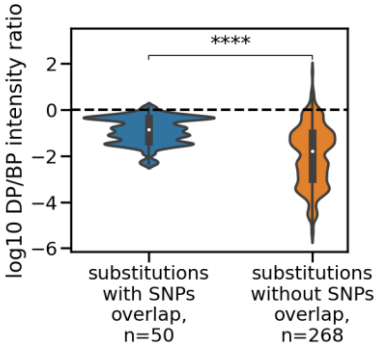

Figure 3: Substitutions in SNP sites have a higher DB/BP intensity ratio than all other substitutions. Y axis is log10 DP/BP intensity ratio values in tonsil tissue. Blue: substitutions that were identified and overlapped with SNPs found in exome sequencing of the same individual. Orange: all other substitutions *identified in the same individual. P –value by Mann–Whitney U test, ****: p <= 1.00e-04*

Out of 9,847 mismatched SNP sites, 50 coincided with translation errors detected by our pipeline in both position and resulting amino acid change (Figure 3). In the original analysis of this data, the detected SNPs were included in the reference proteome, and 724 of these SNPs were identified through proteomics. The discrepancy in detection rates is due to the "open search" approach employed in our analysis, which enables the identification of any substitution at any position but increases the false discovery rate (FDR) compared to the "close search" method that targets specific substitutions.

The difference between the two distributions is significant, as indicated by the Mann-Whitney U test (p-value = 1.7e-07). Even after filtering for SNPs, the remaining translation error candidates still show a bimodal distribution with a high median value.

To see if wider SNPs annotations would explain the SNPs unseen with the annotation of SNPs from exome sequencing, I looked also on the overlap between the translation errors candidates to all SNPs in dbSNP[17] (Figure 4). Another use of this approach is that it can be applied to all tissues and samples from humans and would not require matched exome sequencing.

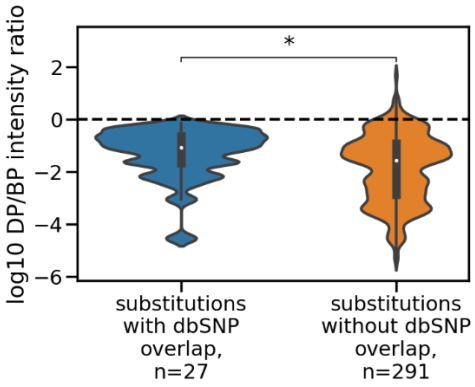

Figure 4: Substitutions in SNP sites according to dbSNP have a higher DB/BP intensity ratio than all other substitutions, but not as significant as for SNPs detected in exome sequencing of the same sample. Y axis is log10 DP/BP intensity ratio values in tonsil tissue. Blue: substitutions that were identified and overlapped with from dbSNP. Orange: all other substitutions identified in the same individual. P –value by Mann–Whitney U test, *: 1.00e-02 < p <= 5.00e-02

Here too, the DB/BP intensity ratio was higher among the sites that overlap with SNPs, suggesting again that some of the high ratios might result from heterozygosity and not from translation errors. Yet, in this analysis the difference between the distributions (p-value: 4.1e-02) is less significant. There are less matches between the MS observed substitutions to the SNP data than in the date that was acquired from the exome sequencing. Because it poorly explains the high error rate and not necessarily means that a heterozygous SNP was in the donor DNA sequence, I decided to keep the substitutions that overlap with SNPs from dbSNP for further translation error related analysis, keeping in mind that some percent of the data is from allele's variation.

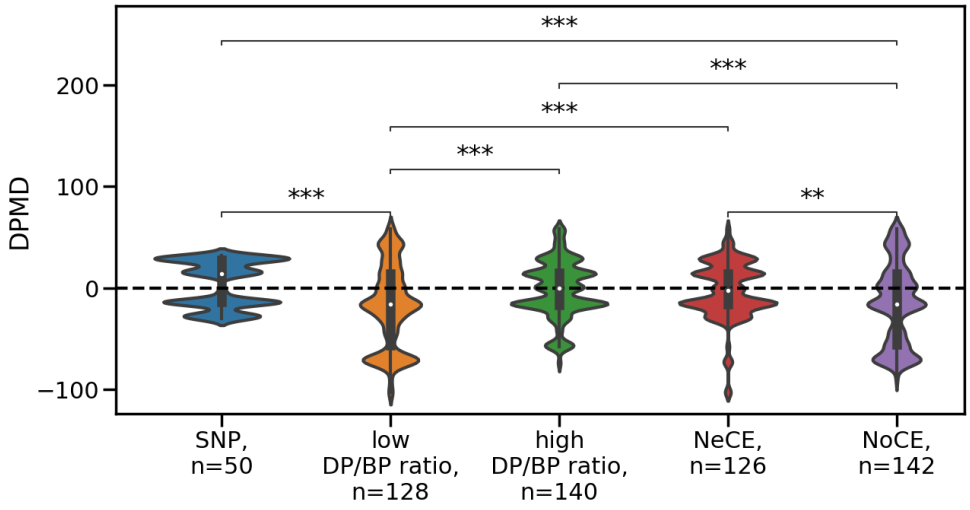

Figure 5: dependent peptide mass difference (DPMD), which is ($Mass_{DP}$ – $Mass_{BP}$), shows different trends for separate groups of substitutions in tonsil. Blue: substitutions that overlap with SNPs found in exome. All the rest of the violins do not include those sites. Orange and green: substitutions with low and high DP/BP intensity ratios, respectively. Red and violet: NeCE (near cognate) and NoCE substitutions, respectively. One substitution could be assign to more than one group in the plot (but SNPs). P –value by Mann–Whitney U test, **: 1.00e-03 < p <= 1.00e-02, ***: 1.00e-04 < p <= 1.00e-03. Pairs that were not significantly different from each other were not marked.

The use of SNP annotations from exome sequencing of the matching sample was expected to support the theory that the high peak of DP/BP intensity was caused by allele variation and to eliminate those substitutions from the data. However, the distribution of mass differences (Mass DP – Mass BP) for SNPs that were identified as translation errors was found to be significantly different (p-value = 2.5e-03) from substitutions with low intensity ratios but not from substitutions with high ratios (p-value = 0.16), which may suggest that some mutations or SNPs were not detected through exome sequencing, see Figure 5.

Another explanation for the high intensity ratios is that the substitutions are caused by transcription errors, but in this scenario the phenomenon would be less prevalent in human humans than in bacteria[18,19].

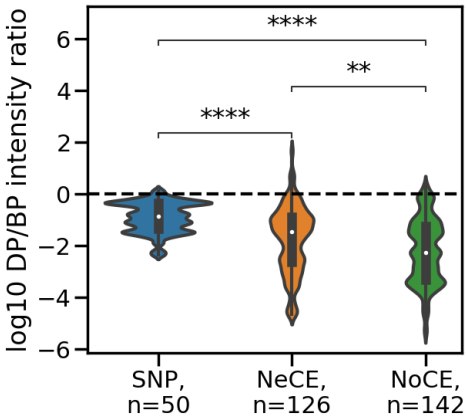

Figure 6: Distribution of log10 DP/BP intensity ratio values in tonsil tissue is different for NeCE, NoCE and substitutions sites that overlaps with SNPs. Blue: substitutions that were identified and overlapped with SNPs from exome sequencing. Orange and green: NeCE and NoCE substitutions that were identified in the same individual, not including substitutions that overlapped with SNPs. P –value by Mann–Whitney U test, *: 1.00e-02 < p <= 5.00e-02

Another method of categorizing translation errors is to differentiate between near-cognate and non-cognate substitutions. Our pipeline uses this approach to broadly distinguish between errors caused by a mismatch between codon and anti-codon during translation and those resulting from misloading errors by tRNA aaRS. Mordret et al. made an assumption that substitutions between near-cognate (NeCE) codons results from mispairing while non-cognate (NoCE) results from misleading. Mutations leading to amino acid changes would fall under the near-cognate category. The lack of significant differences in the distribution of mass shifts

between near-cognate and SNPs is therefore reassuring. The high similarity between substitutions with high DP/BP intensity ratios to NeCE indicates that NeCE substitutions have high intensity ratios. In Figure 6, it does seem that there is a lot of high DP/BP intensity ratios in the NeCE distribution, but the NoCE distribution – although it is significantly different - also have it. Furthermore, NoCE distribution is not unimodal, therefore indicating that the substitutions with high intensities ratio are not SNPs that were missed.

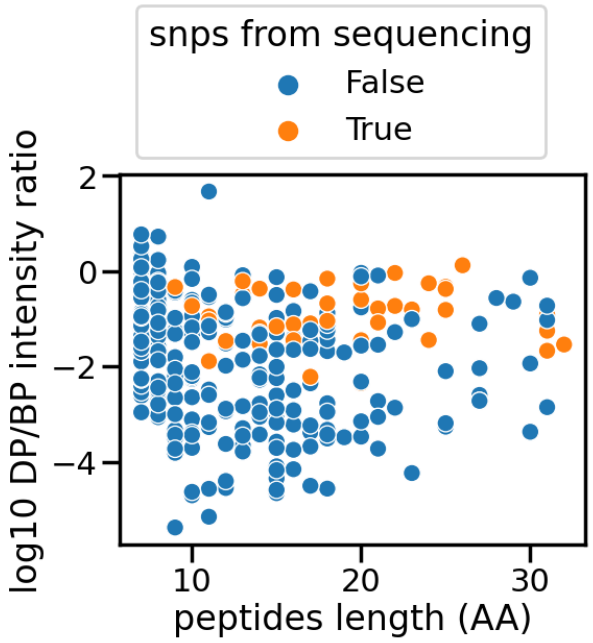

Figure 7: Shorter peptides tend to have larger DP/BP intensity ratio. The scatterplot shows DP/BP Intensity ratios of translation errors (Y axis) for each peptide length (X axis). Substitutions that overlapped with SNPs are marked in orange.

The results of plotting the intensity ratio of each translation error candidate as a function of peptide length (Figure 7) support the idea that there is some bias from ionization efficiency, as peptides of length 7 and 8 amino acids show the highest ratios and avoid low ratios. This may suggest that additional factors, such as differences in the chemical properties of the peptides, are also contributing to the observed intensity ratios. However, the observation of high intensity ratios for translation error candidates in all peptide lengths suggests that this bias does not fully explain the bimodal distribution of intensity ratios. Additionally, the absence of

SNPs in small peptides may indicate a measurement bias, further highlighting the need to consider potential sources of error when interpreting the results.

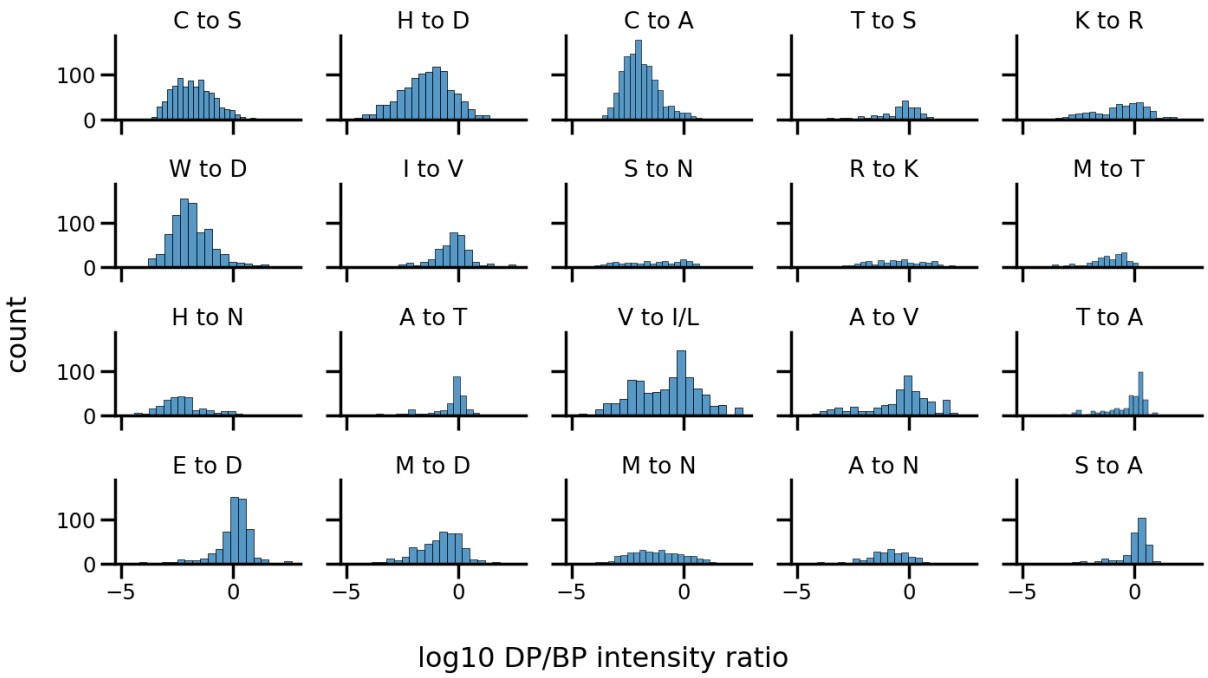

Figure 8: DP/BP intensity ratio distribution is different for different substitution types. DP/BP intensity ratio for each substitution, 20 most frequent substitutions are plotted. The graph displays the distribution of the DP/BP intensity ratios for each substitution.

Analysis of the DP/BP intensity ratio for each type of substitution revealed that some substitutions had a high error rate, centered around 0, while others had values centered on 10^{-2}. The most frequent substitutions, which accounted for 60% of the total substitutions observed, were plotted in Figure 8. The substitutions C to S, C to A, and W to D exhibited unimodal and low error rate, while T to S, A to T, K to R, and I to V were centered on a very high intensity ratio, that would correspond to non-logical error rate of close to 0.5. Notably, A>T, T>S, and I>V were the most frequent SNPs observed in the tonsil, accounting for 14 out of 44 SNPs that overlapped with the translation errors detected. This suggests a possible genetic basis for these substitutions.

Considering what I discovered here, I decided to do further analysis separately for all translation errors candidates and for TEL (Translation Error Low intensity ratio). I set the DP/BP intensities ratio threshold for TEL to be 1/100, for getting most of the lower peak (Figure 9). In

the research on *E. coli* the median intensity ratio was closer to 1/1000, but in Figure 12 it is already visible that that is not the case in all organisms, and particularly eukaryotes may have a higher rate of translation error.

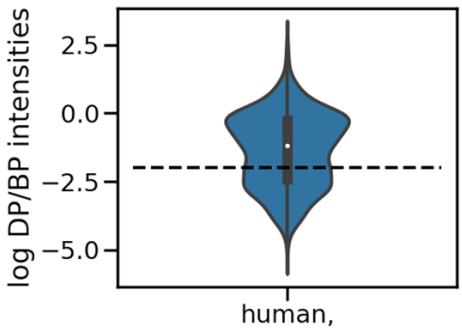

Figure 9: Violin plot of 3761 DP/BP intensity ratios found in human tissues. Dashed line separate between TEH – above the line – and TEL – below the line.

Correlation of expression to DP/BP intensities ratio

In previous work in E. coli it was shown that there is a negative correlation between DP/BP intensities ratios and protein or mRNA expression levels. I calculated and compared the mean DP/BP intensities and the sum of BP intensity for each protein in each tissue seperatly. The correlation between BP intensity to the ratio of DP/BP intensities is negative and high with Pearson coefficient of -0.67 (p-value: 10^{-230}, Figure 10A). As BP intensity is the denominator of DP/BP intensities, it is not very surprising.

To reduce this bias, I used the number of peptides assigned to each protein divided by the length of the protein as an estimator for protein expression and calculated the partial correlation between expression and intensity ratio with control for BP intensity. Most of the correlation got lost with the partial correlation, but it is still significant ($r_{\frac{DP}{BP}\text{intensities,peptide count}} = -0.22$ with p − value: 10^{-21}, $r_{\frac{DP}{BP}\text{intensities,peptide count|BP intensity}} = -0.06$ with p − value: 0.01 , Figure 10B).

The same was demonstrated with gene RNA expression data. Each of the tissue samples that were analyzed in LC-MS was also measured for RNA expression of genes. I used this data by comparing the RNA expression of each gene in each tissue with the mean DP/BP ratio in the

same protein in this tissue. Also here I got significant correlation and partial correlation. ($r_{\frac{DP}{BP}\text{intensities,RNA expression}} = -0.26$ with p − value: 10^{-29}, $r_{\frac{DP}{BP}\text{intensities,RNA expression|BP intensity}} = -0.07$ with p − value: 10^{-3} , Figure 10C)

I saw the same trend of a negative correlation between error rates per protein-to-protein expression level when I used the tRNA adaptation index (tAI) as a proxy for protein expression. tAI is a measure used to predict the efficiency of translation of a mRNA sequence into a protein, based on the abundance and isoacceptor identity of tRNA molecules in the cell[20]. tAI has been shown to correlate with the protein expression levels and with mRNA expression levels in various organisms and therefore it can also be used as an estimator for protein abundance[21,22]. I calculated tAI using the tRNA copy number of humans. The results were similar to what I got

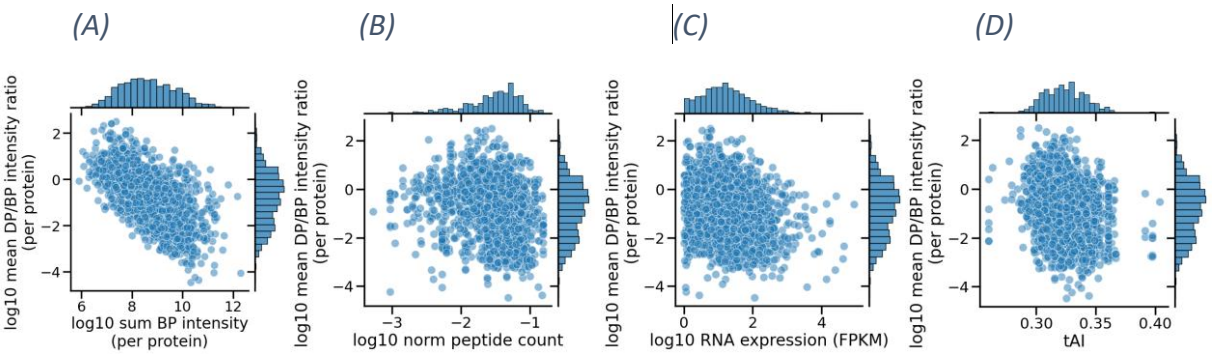

Figure 10: Significant correlation between estimators of protein expression to DP/BP intensity ratio. The scatterplots compares DP/BP intensity ratio (y-axis) to different estimators of protein expression (x-axis). Each dot represent the values of the peptide with the maximum intensity from protein that has at least one error. Number of proteins: 1,526. (A) Scatterplot of DP/BP intensity ratio compared to BP intensity. (B) Scatterplot of DP/BP intensity ratio compared to peptide count (C) Scatterplot of DP/BP intensity ratio compared to RNA expression of the protein in the same tissue. (D) Scatterplot of DP/BP intensity ratio compared to tAI calculated over copy number of tRNAs and RNA expression from each tissue.

for RNA expression ($r_{\frac{DP}{BP}intensities,tAI} = -0.27\ with\ p - value = 10^{-32}$, $r_{\frac{DP}{BP}intensities,tAI|BP\ intensity} = -0.06$ with p − value: 10^{-3}, Figure 10D).

This could indicate that proteins are more expressed less prone to have errors, this even though the pipeline is based on LC-MS results and therefore will be biased towards detecting substitutions in proteins that are expressed more – and therefore their base peptides would be detected more in LC-MS.

Comparing pattern of substitutions between tissues and organisms

Another way to quantify the amount of translation errors is to count how many different translation errors were detected – meaning translation errors in unique site in a protein that occurred towards distinct amino acid. As in Mordret et al. 2019, I made heat map matrix of substitution identifications – see Figure 11. The substitution identification matrix is comprised of 1,884 entries for E. coli, 249 for S. cerevisiae, and 5,510 entries for humans.

The analysis shows that the types of translation errors in humans are more diverse compared to *E. coli* and *S. cerevisiae*. This could be attributed to the larger and more diverse dataset of human tissues, which includes several cell types for each tissue and 29 different tissues, as

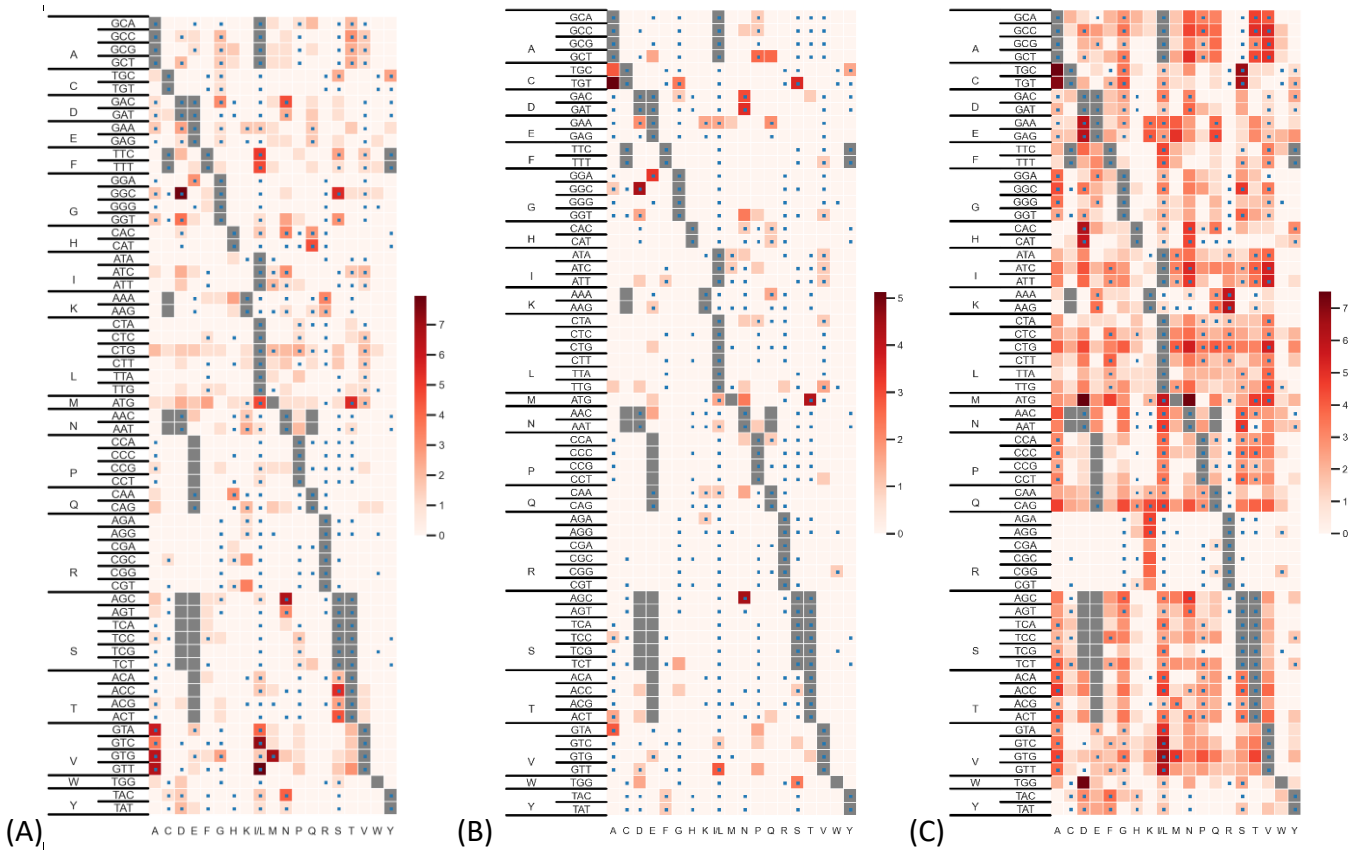

Figure 11: The substitution identification matrix for E. coli (A), S. cerevisiae (B), and 29 tissues from humans (C). Each entry represents a unique translation error, indicating the codon that was mistranslated into a specific amino acid in a particular site in the coding sequence. The color indicates log10 of how many times this translation error was found. Grey cells denote cognate amino acids or translation errors that have the same mass as known PTMs, while blue dots within the cells indicate near-cognate substitutions.

opposed to the smaller and more homogeneous dataset of E. coli.

As noted in Mordret et al. in 2019, one of the most prevalent translation errors in yeast

involves the conversion of cysteine to alanine, from both codons of cys. However, these substitutions are not observed in *E. coli*. This observation was explained by Sun et al. (2016) by demonstrating that the eukaryotic alanyl-tRNA synthetase (AlaRS) tends to mischarge $tRNA^{Cys}$ with alanine, while prokaryotic AlaRS does not exhibit this tendency. Consistently with that, we see those substitutions in human as well.

Due to the high diversity of translation errors in humans, it is noticeable that the codon rows for arginine and lysine have fewer entries compared to other amino acids. This trend is consistent across the three organisms studied. The relatively empty rows of arginine and lysine codons observed in both organisms are likely due to a "blind spot" in our detection pipeline. The pipeline assumes trypsin digestion during the MS preparation, which is an enzyme that cleaves after every arginine or lysine. The algorithm used to detect peptides in the MS data looks for peptides that end with K or R. If a substitution occurs in those amino acids, it would be difficult for the algorithm to detect it. The algorithm can tolerate some miscleavage events, but it would be challenging to find both base and dependent peptides in those cases.

High correlation error patterns in different tissues but not between the organisms

I created vectors of translation errors pattern, where each element represents the number of different positions that were observed with the same substitution type – meaning specific codon to specific amino acid. Then I computed Pearson, Spearman correlation and Jaccard score between each pair of samples (here only Spearman is shown, Figure 12).

Data from various sources, including publications [2,4,6,11], our own experiments, and collaborations with Yardena Samuels Lab (A375 cells), were compiled for analysis. By their statistical properties, Pearson correlation tends to emphasize the similarity of values, including the influence of extreme values, while Jaccard score emphasizes the agreement of patterns, with equal weight given to each entry. Spearman correlation falls somewhere in between these two measures, as it is less influenced by extreme values but also takes into account the magnitude of differences between elements in the vectors, not just their presence or absence.

The Pearson correlations are higher in general, with mean correlation of 0.43, higher than mean correlation of 0.30 for Spearman. This indicates that a lot of the similarity between the matrices is contributed from elements in the substitution matrix that have extreme values. Strong

correlations were observed between most tissues. Except for bone marrow, all the tissues clustered together in Spearman correlation and Jaccard score heat maps. Cell lines from human source and *S. cerevisiae* have not clustered with the human tissues. E. coli clustered with human tissues in the heat maps of Jaccard and Spearman but not in Pearson. High correlations were observed also between experiments using the same cell type, such as B cells with

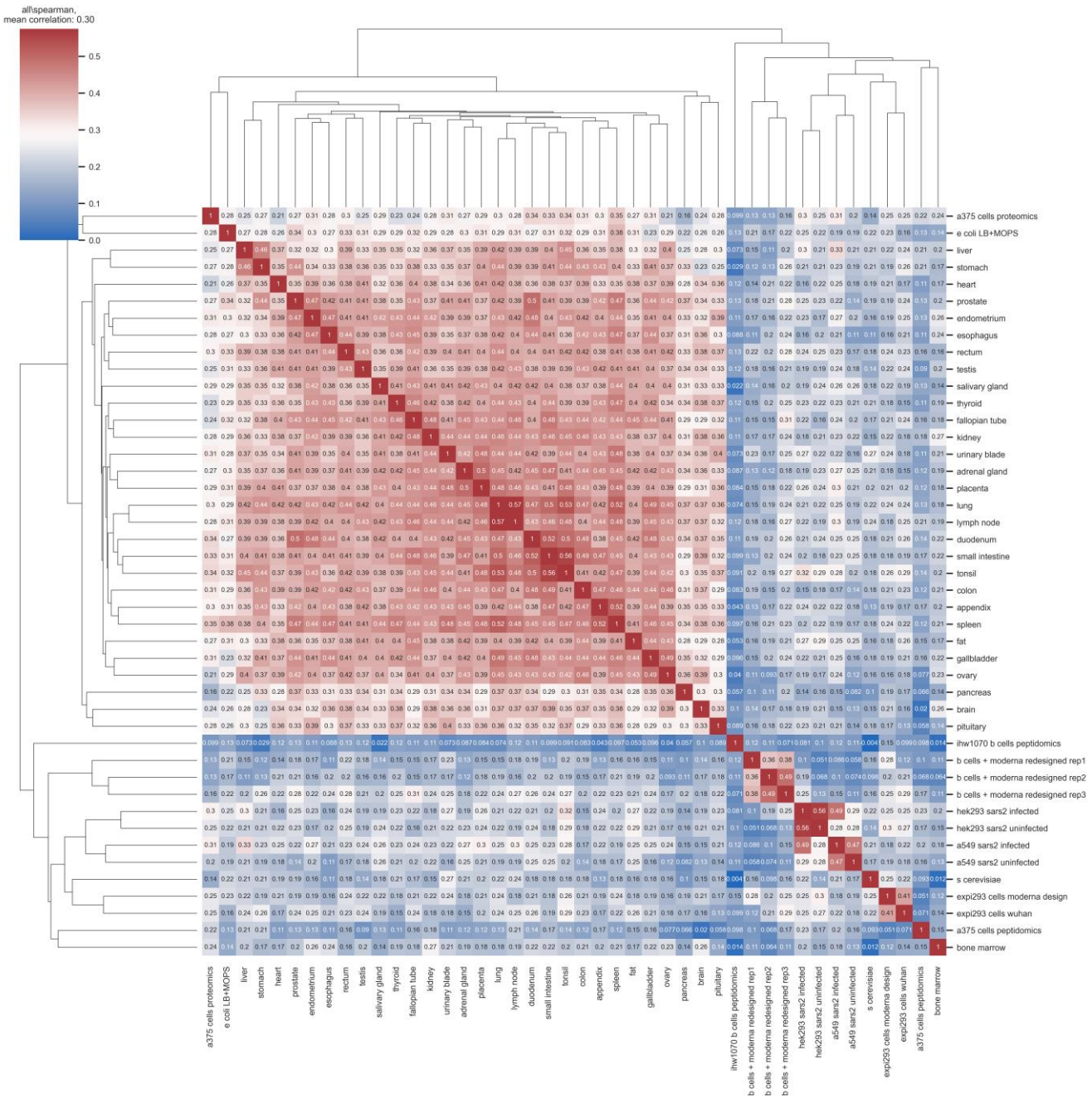

Figure 12: Correlation between tissues is high relatively to other organisms or cell lines. Each entry in the heat map is Spearman correlation between pair of datasets, with clustering (UPGMA method) with the Euclidean distances

Moderna-redesigned plasmid or Expi293F cells that was transfected either with Moderna-redesigned or Wuhan plasmid. Interestingly, the Pearson correlation of HEK293 and A549 cells infected with SARS-CoV-2 to each other (r = 0.63) was higher than the correlation of each of them to the same cell type uninfected – (r=0.61, r=0.47 for HEK293 and A549 cells, respectively). The Spearman and Jaccard score showed the opposite trend but still demonstrated higher correlations than those observed between HEK293 infected or uninfected cells and Expi293F cells, which are derived from HEK293 cells. This suggests that the virus infection has influenced the translation error pattern.

Spearman correlation between substitution is higher in near cognate compared to non-cognate codons

I then compared the Pearson, Spearman and Jaccard scores between NeCE and NoCE cells of substitution matrices of tissues only. I did the same for TEL only.

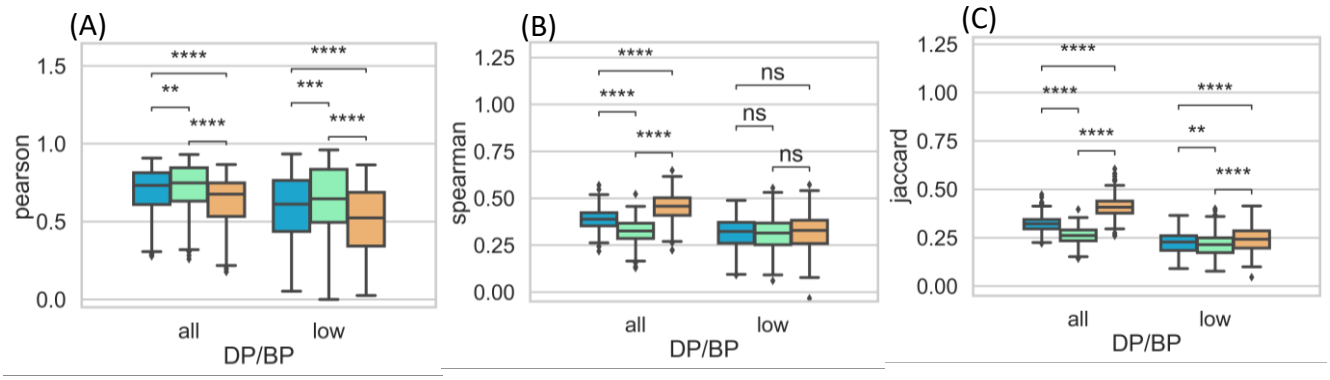

Figure 13: Higher similarity between NoCE substitutions than NeCE for Pearson coefficient, but the opposite for Spearman and Jaccard coefficients. Boxplots shows the distribution of Pearson (A), Spearman (B) and Jaccard coefficients separately for all substitutions (blue), NoCE (light turquoise), or NeCE (orange) .). P –value by Mann–Whitney U test, **: 1.00e-03 < p <= 1.00e-02, ***: 1.00e-04 < p <= 1.00e-03, ****: p <= 1.00e-04.

In all methods, TEL were less correlated than translation errors with all the range of DP/BP intensity ratios (Figure 13). This observation could imply that either translation errors that occur frequently are also common between tissues, or that the substitutions with high DP/BP intensity ratio are the outcome of a process that is more consistent in the cell than translation errors. It is important to note that since the tissues were obtained from different individuals, this consistency does not necessarily suggest genomic variation.

The Pearson correlation displays different trends compared to Spearman correlation and Jaccard score. Specifically, for NoCE, there are higher values for Pearson correlation than for NeCE, while the opposite is true in Spearman and Jaccard scores. This suggests that for NoCE, the cells with extreme values tend to repeat themselves between tissues, which may be caused by aaRS misloading errors that occur frequently and similarly across different tissues. One example of such a misloading event is the substitution of cysteine with alanine, which is a common translation error that occurs due to aaRS misloading. On the other hand, the opposite trend in Spearman and Jaccard scores indicates that the types of errors resulting from ribosome mismatching would repeat themselves to some extent across different tissues.

Identity of proteins errors overlaps between tissues

To investigate whether translation errors and the tendency to make them are a characteristic of a protein, I analyzed the Jaccard coefficient for proteins with errors in each tissue (Figure 14). Surprisingly, the TEL proteins had a higher Jaccard score, ranging from 0.03 to 0.39 with an average of 0.18, compared to all proteins with errors which had a range of 0.05 to 0.24 and an average of 0.15. This is unexpected because TEL had a smaller group size (mean group size = 50) compared to all proteins with errors (mean group size = 229). As sample size decreases, the probability of overlap decreases. This observation could indicate a difference between TEL and

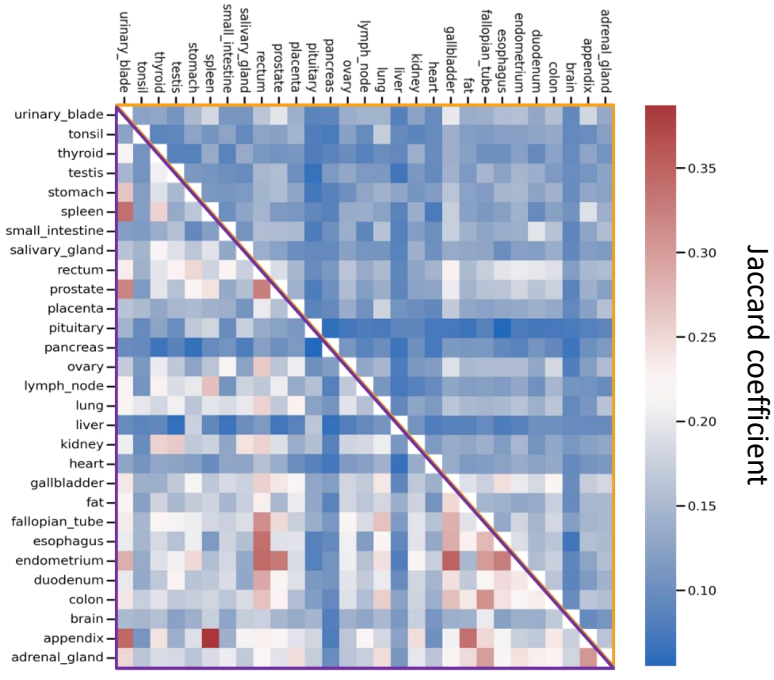

Figure 14: Jaccard coefficient for proteins with errors for each pair of tissues is higher for TEL. Upper orange triangle is for all errors, lower purple triangle is for TEL only.

TEH, aside from the obvious difference in DP/BP intensity ratio. Alternatively, it may be due to the high correlation between DP/BP intensity ratio and BP intensity. This suggests that peptides from the TEL group have high BP intensity, indicating strong expression of the protein. This may explain why groups of TEL protein have a higher Jaccard index between them, as many house-keeping genes are expressed highly and evenly between tissues[23].

To further ask if translation error tend to significantly reoccur in the same proteins rather than randomly across genes, I calculated hypergeometric cumulative distribution function (CDF) for each pair of tissues, as described in Methods (Figure 15). In short, I used the formula: $p(k, M, n, N) = \frac{\binom{n}{k}\binom{M-n}{N-k}}{\binom{M}{N}}$, with n and N is the sizes of the protein group in each tissue, k is the proteins that overlaps, and M is the population size. The entire population size was all the proteins detected in both tissues. All the results were significant: 10^{-136}-10^{-8} with a mean of 10^{-53} for TEL and 10^{-240}-10^{-28} with a mean of 10^{-97} for all translation errors. This indicates that indeed, errors tend to happen in the same proteins. However, one challenge of this analysis is selecting the entire population size (M in the formula). Using all the proteins detected assumes that translation errors would be detected in each protein if they exist. In reality, there are properties of the protein that will affect the rate of detection of translation error. To address this issue, I calculated the hypergeometric CDF between each pair of tissues with population size that includes only the proteins that were detected with errors across all tissues. This

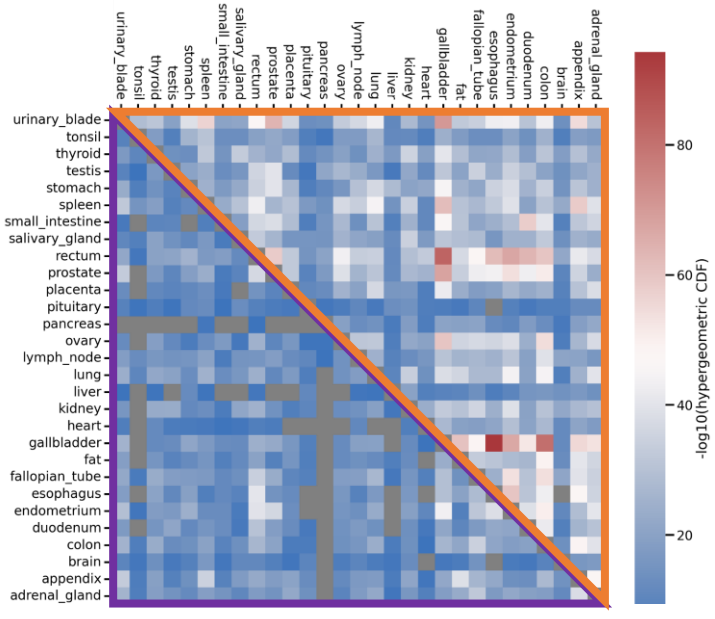

Figure 15: hypergeometric CDF of overlap between tissues is mostly significant. Each entry color represents -log10 of hypergeometric CDF between each pair of tissues, with population size of all proteins with errors for the group. Gray entries are for non-significant hypergeometric CDF. In orange upper triangle is for group of all translation errors, and in purple lower triangle it's only for TEL.

ensures that only proteins that could have been detected with errors are considered for the p-value. This also shows the significance of similarity between proteins with errors between tissues.

Out of 406 comparisons, one pair (pituitary and esophagus) did not pass the α family-wise error rate (FWER) threshold that was calculated using Bonferroni correction (see Methods). This occurred in all translation errors data, and there were more pairs that did not pass in TEL data, mainly in pancreas and tonsil. In general, the significance of the overlaps of TEL were lower than in all translation errors, as opposed to the Jaccard score. Most of the pairs had significant overlaps between themselves, emphasizing that errors tend to occur at the same proteins.

Different tissues have different rates of translation errors

Counting translation errors by site and destination, rather than intensity enables a comparison between the amount of peptides that were detected with errors and those without. The findings revealed that the rate of translation errors varied by almost an order of magnitude across tissues (Figure 16A, C). It is noteworthy that the ranking of the tissues was different for the ratio (peptides with errors) / (peptides detected) than for the median of DP/BP intensity ratio that the peptides with errors received in each tissue. The ranking of the tissues was not kept. Also, the ranking of tissues is preserved to some low extent between all substitutions and TEL for the ratio of peptides, but not for the ratio of intensities.

The equation: $\frac{peptides\ with\ errors}{peptides} \times Median(\frac{DP}{BP}\ intensity\ ratio)$ can be calculated to estimate the error rate for cells (Figure 16B, D). For all translation errors, this value ranges between 10^{-4} for colon to 10^{-3} for appendix, liver and stomach. For TEL only, it ranges between 1×10^{-6} for pituitary gland, to almost 8×10^{-6} for fat and pancreas. Overall, it means that the range detected is between 10^{-6}-10^{-3}. This is in line with other estimations of translation error rate in the literature[3]. Reanalysis of E. coli translation errors data shows that ratio of peptides with errors compared to peptides detected in LC-MS is 7×10^{-3}, this number is similar to the ratios from TEL.

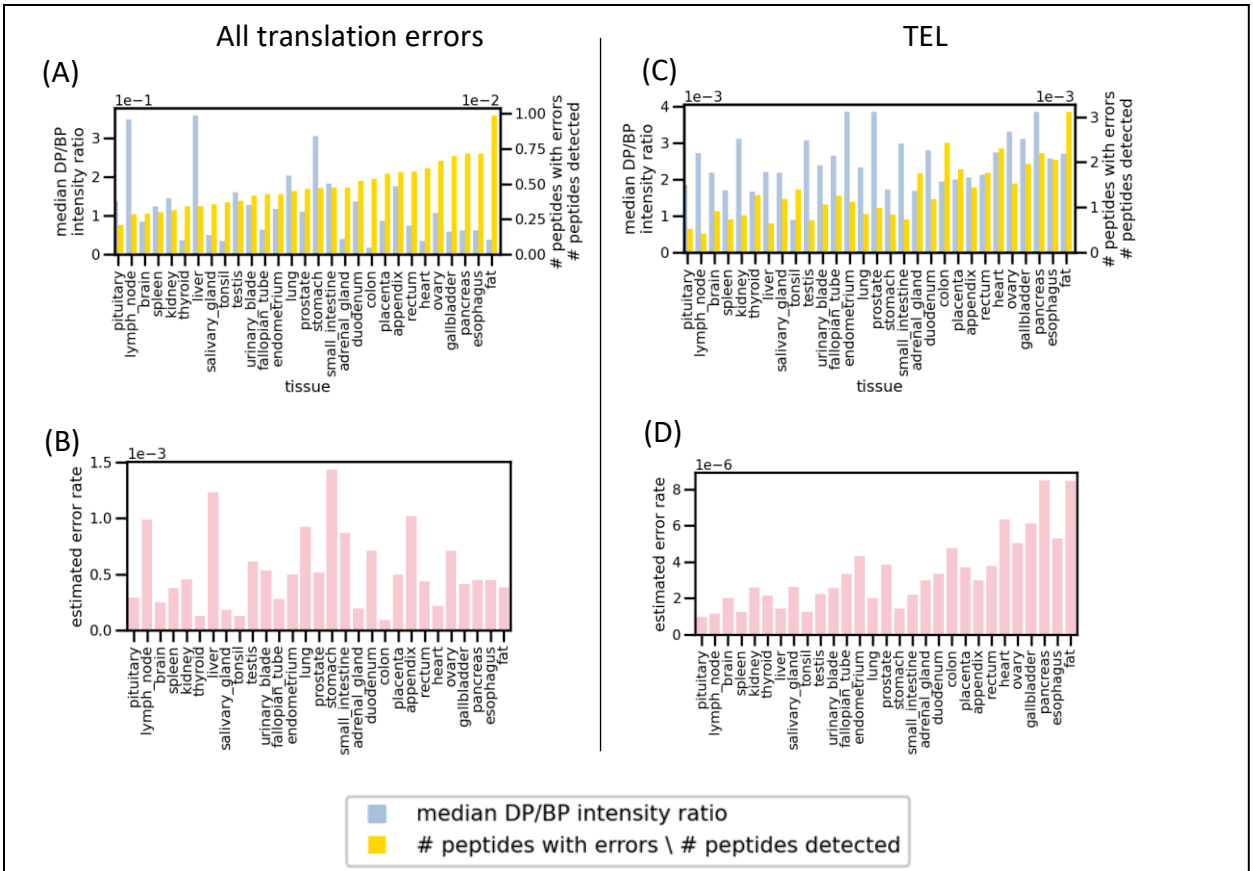

Figure 16: proportion of peptides with translation errors compared to peptides without translation errors. (A) Left y-axis is median DP/BP intensity ratios, and right y-axis is ratio of peptides with errors to peptides without errors, for each tissue on x-axis. (B) Same as C but only for TEL. (C) Y-axis is the ratio of peptides with errors to peptides without errors times the ratio of DP/BP intensity ratio for each tissue on x-axis. (D) Same as A, but only for TEL.

Comparing phenotype to genotype variation

I next turned to study the relationship between translation errors and mutations that fixate at the DNA level of orthologous genes. The main objective of this part of the study was to compare the extent of translation errors as "phenotypic mutation" at various proteins and sites to the rate at which the genotype evolves at such sites. I have done this by calculating the evolution rate score of each position in human proteins in comparison to their orthologues across vertebrate species. The rate of evolution score of each position in each protein was calculated using the rate4site score (in short R4S)[24] on the multiple sequence alignment (MSA) of each protein in the human proteome, each aligned to its orthologues in 100 vertebrate species (see Methods). The R4S score is inversely related to conservation of a position in the alignment, i.e. if conservation is low the position is more rapidly evolving.

To begin the analysis, I compared the R4S scores for all human proteins with MSA alignment, all proteins detected in LC-MS of tissues, and all peptides detected in LC-MS of tissues. If a specific position in a protein was included in more than one peptide, it was counted only once. Proteins from the entire human proteome that had MSA had the highest mean, which is reasonable given that proteins that are detected are typically highly expressed. Generally, highly expressed proteins evolve more slowly[20]. The gap between MS-detected proteins and MS-detected peptides was smaller but could represent the same phenomena. Proteins that exhibit higher expression levels are more likely to be detected through LC-MS, and as a result, have a greater representation in the group of detected peptides. To correct for this bias, I randomly chose one peptide per protein in the peptide groups. This correction resulted in the proteins and peptides detected being equal, confirming that the difference was due to this bias.

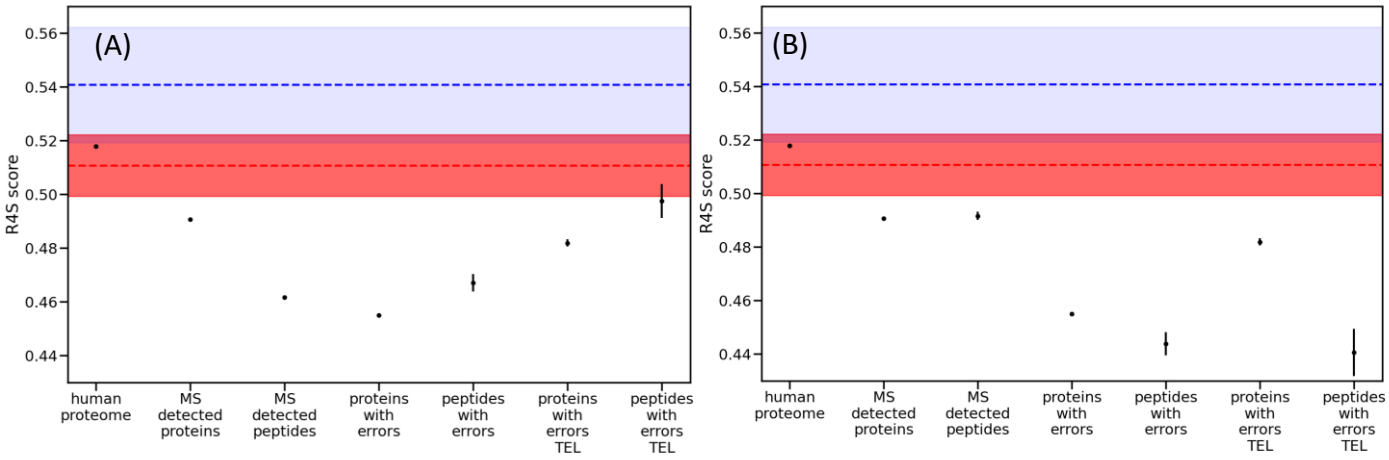

Figure 17: sites of errors are less conserved than the protein it is coming from or from the group of peptides that it is found at. Each dot represents on the y-axis the mean R4S for each group on the x axis. The black line around each dot is the standard error of the mean (SEM). Red dashed line represents the mean of all translation errors, with red area around it marking the SEM of it. Blue dashed line and blue area around it is the same but only for TEL.

The correlation between expression and conservation of protein could also explain why proteins that were detected with errors have R4S score that is similar to MS detected peptides and even lower. Although there is negative correlation between DP/BP intensity ratios to expression, the pipeline used to detect errors may be more likely to identify errors in highly expressed proteins due to the higher representation of their peptides in MS spectra. Peptides

with errors have higher R4S scores than proteins with (Figure 17A), both for all translation errors and for TEL. However, when only one peptide per protein is considered, the trend is reversed (Figure 17B), indicating that proteins with multiple errors are less conserved. It is difficult to determine from this whether the immediate environment of the translation error is more or less conserved than the rest of the protein.

Finally, I observed that the average of all translation error sites, both for all translation errors and for TEL, was more evolutionarily diverse than the peptides or the proteins that carry errors, with and without correction. This finding further supports the notion that translation errors tend to occur at sites that have a high rate of evolution, which is consistent with previous findings[4], and also indicate that translation errors occurs in specific sites and not at random.

Then I compared the mean and median R4S scores of positions with translation error with the positions around it (Figure 18, Figure 19).

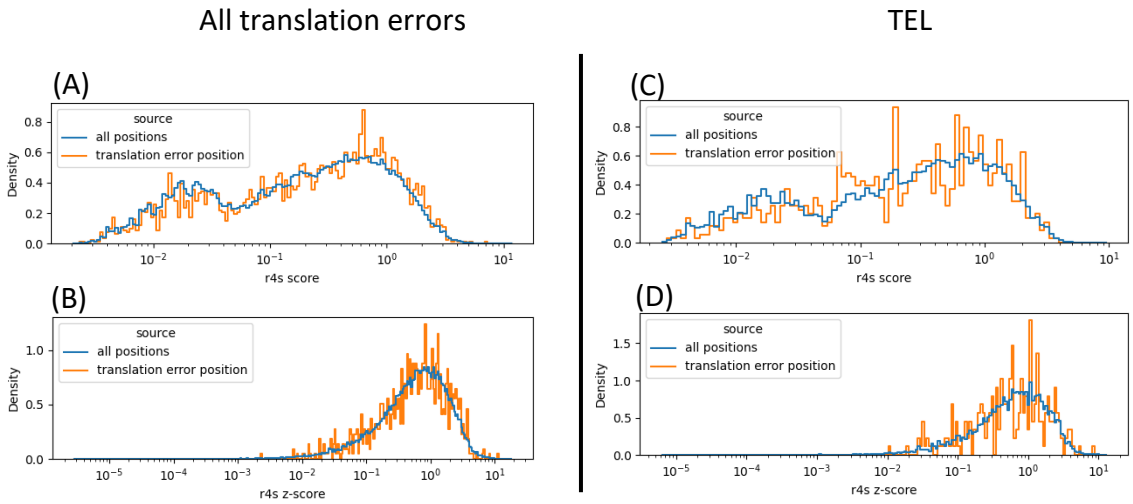

Figure 18: R4S distribution of translation errors sites (in orange) and positions around it (in blue). Results for all errors are represented in panels A, B while C, D show the results for TEL exclusively. (A), (C) shows the distribution of R4S scores. (B), (D) shows the distribution of z-score of R4S score calculated for each gene.

It was found that both mean and median of the positions with translation error were significantly higher than the positions around it (two sides t-test p-value: 9.0e-7, two sides MWU test p-value: 8.7e-8, Figure 19A, C). The results were statistically significant even when the scores were normalized for each gene by calculating the z-score of R4S for all gene positions (two sides t-test p-value: 1.4e-3, two sides MWU test p-value: 7.3e-3, Figure 19B, D). When

analyzed separately, only brain, kidney, pituitary gland, and spleen had significantly higher R4S score for evolutionary rate for translation error position compared to the other positions in the proteins that harbor them.

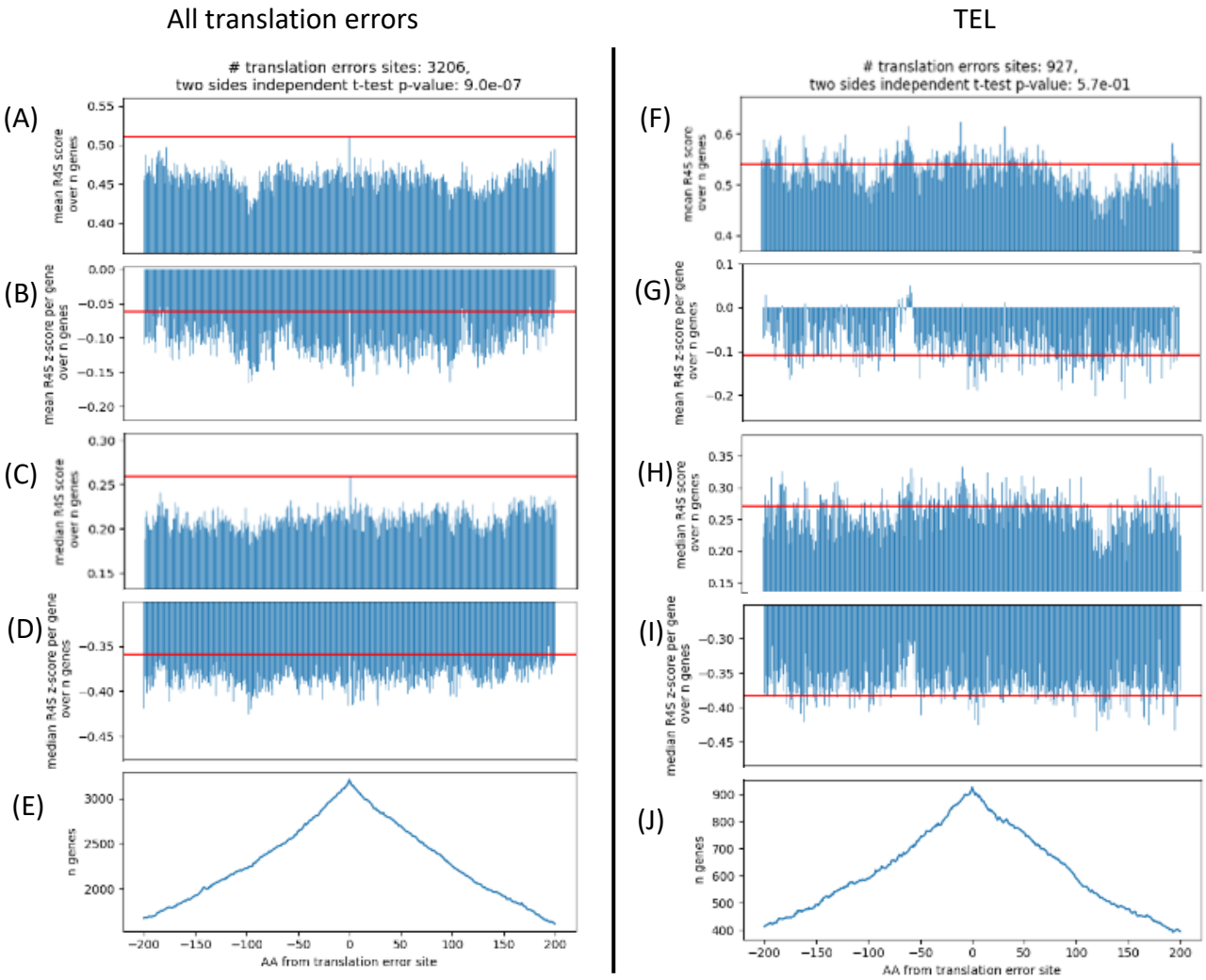

Figure 19: Significantly higher divergence across species for AA sites with errors compared to sites around it. Barplots shows R4S scores surrounding the site of translation errors, with all errors being aligned. The x-axis represents the distance of each amino acid from the site of the translation error, with position 0 being the error site itself. The R4S scores were calculated solely for amino acids in the CDS, with no consideration given to positions before the start codon or after the end codon. Results for all errors are represented in panels A, B, C, D, and E, while panels F, G, H, I, and J show the results for TEL exclusively. (A), (F) y-axis is mean of R4S. (B), (G) y-axis is mean of z-score of R4S score calculated for each gene. (C), (H) y-axis is median of R4S. (D), (I) y-axis is median of z-score of R4S score calculated for each gene. (E, J) y-axis is the number of positions over which the y-axis values were calculated.

Interestingly, there was no significance for the same analyses for TEL only (Figure 19F-J). Positions with errors in TEL had lower R4S score. This indicates that the TEL group is different

from the rest of the translation errors identifications not only in DP/BP intensity ratio. Previous research conducted in our lab on E. coli[4] also could not find a significant difference between sites of translation error to the protein it came from. However, it was significant only when compared to positions with the same codon or amino acid as the translation error site. It seems that most of the positions in E. coli had a lower R4S score than positions with errors. This could imply that TEL is not necessarily more reliable translation errors than TEH, or that the mechanism, including the genotype, of translation errors in E. coli and human is very different.

Design multipotential RNA vaccine

To create an RNA vaccine that targets multiple strains by leveraging translation errors, I obtained peptidomics data from human B cells infected with SARS-CoV-2 or virus proteins[2]. B cells, also known as B lymphocytes, play a critical role in the adaptive immune response by producing antibodies that specifically recognize and neutralize pathogens. When a virus infects a human, the virus proteins can be processed into peptides and presented on the surface of infected B cells, bound to MHC molecules. Because B cells are central to the antibody response, they are a useful model for studying the presentation of viral peptides on MHC molecules. Translation errors can also alter the level of presentation of a peptide on MHC, which is important for determining its potential immunogenicity. Therefore, it is crucial to understand which translation errors result in presentation on MHC to assess the resulting peptide's ability to elicit an immune response. By applying translation error detection pipeline (as described in the Methods section) on the peptidomics data that was either from infected with SARS-CoV-2 or virus proteins or from control group, I gained insight into the pattern of viral peptide presentation on MHC molecules in the context of an actual viral infection and without infection (Figure 20).

While peptidomics provides valuable information on the presentation of viral peptides on MHC molecules, it is a relatively low-throughput method compared to proteomics. This, coupled with the fact that translation errors make up a small fraction of total protein products and are difficult to detect in open searches, results in limited detection of translation errors in peptidomics data from B cells. I was able to identify 83 unique translation errors, representing sites in proteins that were altered to produce a different amino acid (Figure 20). To supplement

the limited data obtained from human B cells, I utilized tables of translation errors from previous research on *E. coli* and *S. cerevisiae* (Figure 9).

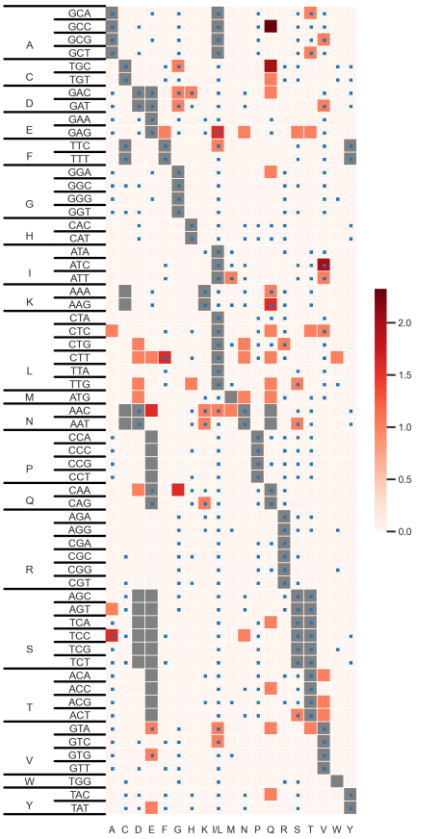

Figure 20: The substitution identification matrix for combined peptidomics results from B cells that were infected with SARS-CoV-2 or virus proteins or without infection. Each entry represents a unique translation error, indicating the codon that was mistranslated into a specific amino acid in a particular site in the coding sequence. The color indicates log10 of how many times this translation error was found. Grey cells denote cognate amino acids or translation errors that have the same mass as known PTMs, while blue dots within the cells indicate near-cognate substitutions.

I employed Pfizer-BioNTech and Moderna's published vaccine sequences as the foundation for the vaccine design, and redesigned it with synonymous mutations, such that the mutated codons will have different translation error patterns and will result in different variants of the Spike protein (See Methods). The goal was to introduce diversity into the vaccine sequence, such that the vaccine will provide protection against multiple variants of the virus. To make informed decisions on which codons to use in the vaccine design, I considered peptidomics data from human B cells infected with SARS-CoV-2 or virus proteins, as well as previous research on translation errors in *E. coli* and *S. cerevisiae*. The selection of codons was also influenced by the existing codon usage in the published vaccine sequences from Pfizer-BioNTech, Moderna and original sequence of Wuhan strain. Using an algorithm described in method, I chose synonymous codons in 37 specific sites along the sequence and designed and ordered the synthesis of 6 plasmid sequences: Moderna, Moderna-redesigned, Pfizer, Pfizer-redesigned,

Wuhan, and construct with GFP. All the sequences ended with Twin-Strep-tag® (TST)[25] for immunoprecipitation after.

To verify the expression of the GFP plasmid and the Moderna-redesigned construct, we performed FACS analysis using Streptactin fused to GFP. Figure 21 depicts the distribution of fluorescence intensity in cells. Interestingly, a small right indentation was observed in cells with the plasmid, indicating the expression of the protein. However, the indentation was barely visible, raising concerns about the reliability of this observation. One possible explanation for the low detection sensitivity could be that the Spike protein is a membrane protein, and the strep tag is located in a small cytoplasmic portion. This could contribute to the difficulty in detecting its expression.

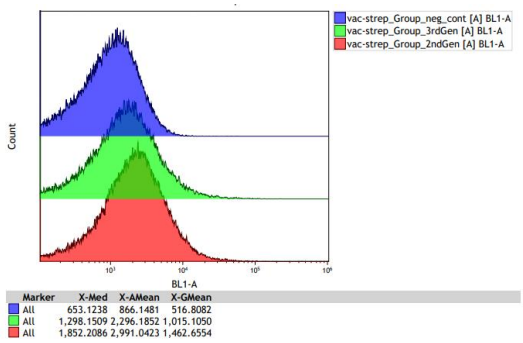

Figure 21: Distribution of fluorescence intensity in cells that were transfected with Moderna-redesigned construct and tagged with Streptactin (anti-strep antibody) fused to GFP, measured by FACS, contains higher values than in control group. The blue color represents cells without infection, while the red and green colors represent two repetitions of cells transfected with the Moderna-redesigned plasmid, using 2nd generation lentiviral plasmids and 3rd generation, respectively. The cells with the plasmid exhibit a small right indentation, indicating the expression of the protein.

Considering the low expression observed in B cells, we utilized Expi293F cells, which are derived from HEK 293 cells and are well-known for their suitability as a model for protein expression, in collaboration with the Life Sciences Core Facilities. The results of the western blot analysis performed using Streptactin showed the presence of bands corresponding to the Spike protein and its cleaved form in all plasmids except for the Moderna-original plasmid.

Analysis was carried out on both B cells infected with the Moderna-redesigned vaccine and Expi293F cells infected with the Moderna-redesigned and Wuhan plasmids, using LC-MS at the Life Sciences Core Facilities.

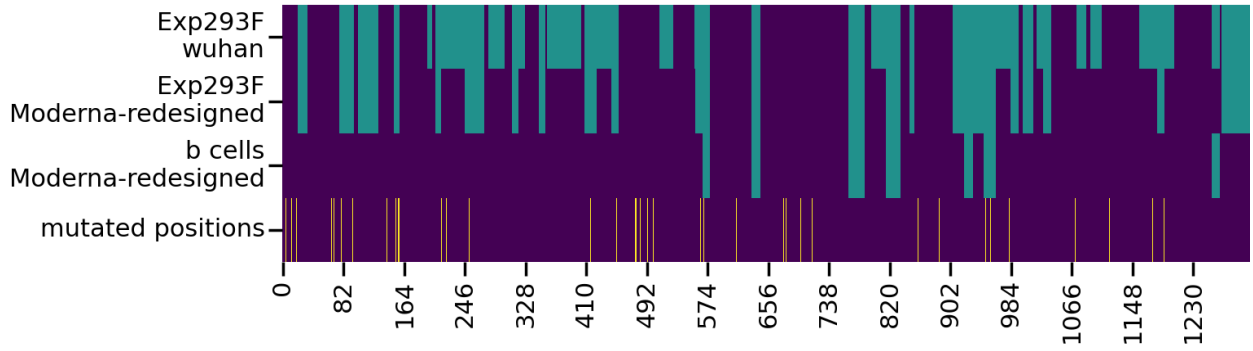

Figure 22: Expi293F cells expressed better the Spike protein, and the Wuhan sequence were expressed better than the redesigned vaccine of Moderna. Peptide coverage of the Spike protein by LC-MS. Y axis describes each sample, and x axis is AA positions in the Spike protein.

More peptides of the Spike were detected in Expi293F cells than in B cells, indicating that the expression of the Spike protein in Expi293F cells was better (Figure 22). One peptide in Wuhan strain covers positions 988-995. The only non-synonymous mutations in the vaccines compared to the Wuhan sequence are found at positions 986-987, where the amino acids K986 and V987 were replaced with prolines. There was no coverage of those positions in Expi293F cells transfected with Moderna-redesigned and also no peptide indicating mixed population in Wuhan transfected cells, so it is not clear if somehow the cells that were supposed to be transfected with Wuhan strain got transfected with Moderna-redesign, if Wuhan sample got contaminated with moderna-redesign cell population, or if the names of the samples were mixed along the way. It is safe to say at least that there are cells infected with Moderna-redesigned or other vaccine-based plasmid in the sample of Expi293F Wuhan.

To specifically detect translation errors, our translation error detection pipeline was applied (see Methods), and for the Expi293F cells a basic proteomics analysis was also conducted. The reference proteome was expanded by incorporating sequences that could result from translation errors in any of the 37 sites that were mutated. No translation errors were found using the general translation error detection pipeline. However, one substitution was found in a position that was mutated – Q957 has substituted with E. The modified peptide – 'LQDVVNQNAEALNTLVK' - was only found in Wuhan sample, while the unmodified peptide – 'LQDVVNQNAQALNTLVK' was detected in both samples. The intensity ratio of the

modified/unmodified in Wuhan sample is 0.03. This site was targeted by us because of a Q957R mutation in the Spike that occurs in some of Iota (B.1.526) strains. While it is possible that this change is due to spontaneous deamination of glutamine, which is a well-known phenomenon[26,27], it is unlikely because the rate of deamination for glutamine in pentapeptides ranges from 500 to 17,000 days (about 46 and a half years), and the rate of deamidation in proteins is estimated to be even slower[26], while the duration of our experiment with the Spike protein, which involved expressing the protein and conducting LC-MS analysis, was only a few months. months.

Both RNA vaccines (Moderna's and Pfizer's) carry the RNA modification of N1-Methylpseudouridine (m1Ψ) in each uracil (U) across the sequence. There are hints that this modification could lead to different patterns of translation error – in terms of frequency and destination amino acid. To check that, and to make our experiment more like the current vaccines, I also analyzed RNA spike sequence that were expressed by Noam Stern-Ginossar Lab and were processed in LC-MS by Tamar Gaiger Lab. They expressed the sequence directly from RNA in HEK293 cells. For investigating the influence of modifications on translation errors, they had the sequence with m1Ψ modification, its related modification pseudouridine (Ψ), and without modifications as control. Out of 48 peptides found, none of them were found with errors in the dependent peptides pipeline. To further investigate the presence of errors, I utilized a targeted approach. I introduced potential errors into the reference sequence of the Spike by converting U bases to A, G, or C, based on the genetic code, in every position of the Spike that were detected by MS. This resulted in the detection of seven modified peptides, with two representing simple amino acid replacements and five resulting in deduced substitutions into K or R and a shortened peptide length. The errors found were: L84 to R, L368 to R, L387 to V, N394 to K, F541 to Y, L560 to R, and I770 to K\R (Table 1). Out of it, some overlapped between the samples and some has not. In total, 5 were found in m1Ψ, 5 in Ψ, and 4 in sample with unmodified U. Those numbers can not indicate whether there is difference in the tendency of each of those modified uridines to have errors. By the nature of this search, all substitutions found were NeCE. The near cognate codons that could explain this mismatch were with a change of uracil to either adenine (A) or guanine (G), meaning that there was a

misincorporation of tRNA that matched in 2 positions to the codon, but carried uracil or cytosine (C) instead of adenine in the position that couples with uracil in the RNA codon. The generation of multiple hypotheses through this targeted approach may have implications for the accuracy of our findings. To address this, negative control experiment should be conducted where translation errors that will caused by mismatch of adenine, cytosine and guanine should be looked for.

Sequence	substituion position	substitution	original codon	nucleotide substitution	Sum Intensity	Intensity N1M	Intensity PseudoTP	Intensity UTP
AVEQDKNTQEVFAQVK	770	I to K\R	AUA	U -> A/G	2.04E+06	0	1.17E+06	8.69E+05
CVNYNFNGLTGTGVLTESNKK	541	F to Y	UUC	U -> A	5.2E+07	3E+07	1.5E+07	1E+07
PFNDGVYFASTEK	84	L to R	CUA	U -> G	1.41E+06	1E+06	0	0
PFQQFGR	560	L to R	CUG	U -> G	9.9E+07	4E+07	3.5E+07	2.1E+07
VNDLCFTNVYADSFVIRGDEVR	387	L to V	UUA	U -> G	3.55E+06	0	3.55E+06	0
VYADSFVIRGDEVR	394	N to K	AAU	U -> A/G	1.2E+07	1E+07	0	0
YNSASFSTFK	368	L to R	CUA	U -> G	5.1E+07	3E+07	1.5E+07	6.77E+06

Table 1: substitutions found in samples with RNA of wuhan strain Spike.

Methods

Vaccines redesign algorithm

Utilizing information from multiple sources, 44 targeted mutations in the Spike protein that were identified in variants of concern were collected. These variants of SARS-CoV2 were particularly monitored due to their prevalence in widespread strains or due to their higher virulence. Of the 44 mutations, 39 occurred at distinct sites, while some of the mutations in different strains were found at the same location. Additionally, two of the mutations involved the amino acid tryptophan, which has a single codon and therefore could not be altered to synonymous codons. The mutations identified was:

L5F, S13I, T19R, A67V, V70F, D80A, D80G, T95I, G142D, E154K, F157S, R158G, D215G, A222V, D253G, K417T, K417N, L452R, S477N, T478K, E484K, E484Q, S494P, N501Y, F565L, A570D,

D614G, Q677H, P681R, P681H, A701V, T716I, T859N, F888L, D950N, D950H, Q957R, S982A, Q1071H, D1118H, V1176F, K1191N

Because this is a first-of-its-kind experiment done on humans, we wanted to create as wide a variety of sequences as possible. For each of the mutation's sites, if its cognate amino acid had only two codons, for the redesigned version of the vaccine the codon that is not in use in the original vaccine was chosen. All the mutations that appeared in common sites with other mutations (sites number: 80, 417, 484, 681, and 950) belonged to this group. If there were more than two codons but up to four, and Moderna\Pfizer vaccines had different codons at this site, we incorporated different codons for each of the plasmids: Pfizer, Moderna, Pfizer-redesigned and Moderna-redesigned. If Pfizer and Moderna had the same codon or if the cognate amino acid has 6 codons, the codons for the redesigned vaccines were chosen based on translation errors scores. The codon with the highest score that were not the original sequences of the vaccines were assigned randomly to either Moderna's or Pfizer's redesigned version, and then the second best were assigned to the other. The scores were calculated as follows:

Substitutions matrices from B cell peptidomics, S. cerevisiae and E. coli were taken into account. To normalize the amount of translation errors found in each organism, each cell in the matrix was divided by the sum of all cells in the matrix. Then, the score of a specific substitution was the sum of the values in the normalized substitution matrix for the specific substitution desired.

To avoid having restrictions sites of BamHI, 879G>A mutation was introduced to Moderna and Moderna-redesigned sequences, 1734T>C for Pfizer and Pfizer-redesigned, and 3753A>C, 3754T>A and 3755C>G in Wuhan strain. All those mutations are synonymous and were observed in the other sequences.

The stability and folding of the mRNA vaccines were evaluated as well as change in tRNA adaptation index (tAI) for several human cell lines to ensure that the changes made to the codons would not affect the overall expression or structure of the Spike protein.

Finally, the sequences were ordered from Twist inside pTwist+Lenti+SFFV+Puro+WPRE plasmid.

Translation errors detection

MaxQuant and custom Python code were employed to detect translation errors, as described in Mordret et al. 2019. Peptidomics data were analyzed with non-specific digestion parameter and a minor change in the python code to allow that, and with protein FDR of 1 instead of 0.01 in whole proteome, to allow more detections. Error rate quantification was performed by fetching the values from dependentPeptides.txt table. To compare error rates from human source to error rates of other organisms (Figure 3), the python script quantify.py from: https://github.com/ernestmordret/substitutions/ was used, because it was compared to tables from the Mordret et al. that were also analyzed with this code.

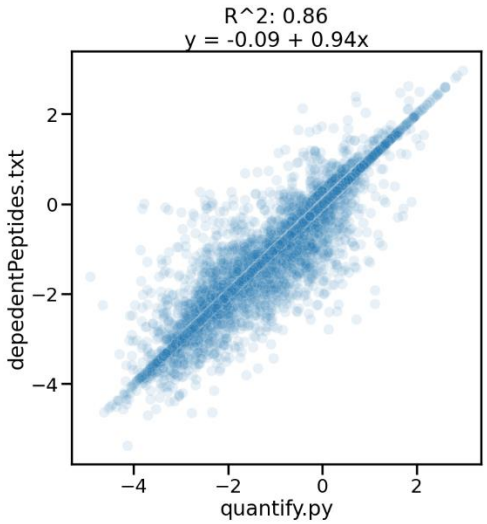

Figure 23: High correlation between different methods to quantify DP/BP intensity ratio for different peptides quantified by either quantify.py (x axis values) or depedentPeptides.txt (y axis)

Hypergeometric cumulative distribution function between pairs of tissues and Bonferroni correction

The probability of specific amount of proteins with errors to intersect between $tissue_i$ and $tissue_j$ was calculated using cumulative hypergeometric score using SciPy module for Python[28] with the following parameters:

P = set of proteins within tissue

PE = set of proteins with error within tissue

k = length($PE_i \cap PE_j$)

M = length($P_i \cup P_j$)

n = length(PE_i)

N = length(PE_j)

And running the function:

```
scipy.stats.hypergeom.logsf(k, M, n, N)
```

The same was calculated to explore the significance of identity of proteins with translation errors between tissues compared to other tissues, with M being the union of all proteins with errors across tissues:

$$M = length\left(\bigcup_{i=1}^{n} T_i\right)$$

p-value threshold were calculated using Bonferroni correction as following:

$$\mathrm{m} = \text{number of comparisons between tissues} = \frac{28 \times 29}{2} = 406$$

α = 0.05

$$\frac{\alpha}{m} = \frac{0.05}{406} = 3.1 \times 10^{-5}$$

Evolution score calculation

The evolution score for each position in the human proteome genes was calculated using rate4site. To do this, a proteome sequence alignment of human to 100 vertebrates was downloaded from UCSC, along with a tree file that contained data about the phylogenetic distances between all organisms in the alignment. Then each gene was separated to different fasta file using Python script. The command used to run rate4site was:

```
rate4site -s $GENENAME.fasta -o $GENENAME.res -t hg38.100way.nh
```

where $GENENAME represented the name of each gene in fasta format.

For analysis I only took r4s score for positions that had > 80 MSA.

Discussion

Investigating the misincorporation of amino acids during the translation process in human tissues is crucial for gaining insights into the underlying mechanisms of protein synthesis and identifying the causes of diseases such as neurodegenerative disorders[3,29]. Translation fidelity is also associated with life span[30], and understanding of it could improve protein engineering.

In this study, my main objective was to enhance the comprehension of translational errors and investigate their occurrence within the human context. To achieve this goal, I examined LC-MS data from 29 diverse human tissues that were obtained from the Human Proteome Atlas (HPA) using my laboratory's tools.

I encountered some difficulties when trying to apply the pipeline designed for E. coli to human data. The ratio of peptides that were translated accurately to peptides that carried mistakes reached many times close to 1/1, a phenomenon that was never reported before and made me suspect that those are not translation errors. Some of it I could attribute to the fact that human cells are diploid (except for germ line cells), and mutations can be detected as translational errors because reference proteomes do not account for these mutations. In those cases, heterozygosity could explain the 1/1 ratios. To overcome this challenge, I compared the translational errors I discovered to mutations identified in the same sample to estimate the proportion of results that were due to mutations. Indeed, SNPs that were different between alleles were detected sometimes as translation errors, but there were still many translation errors with 1/1 ratio that were not attributed to SNPs. There is a possibility that in the data I had not all the SNPs were mapped because of the high threshold that was used to detect it. Another possibility is that there are other events that could happen in the cell that could lead to those results. One such event is non-synonymous RNA editing[31]. Another possibility is that there are pseudogenes that are paralogs of genes in the human genome but with changes in some of the AA coded, that even if they are not completely translated, some peptides in it would be translated[32]. RNA modifications can also influence translation errors and cause wide differences between the rates of it in humans compared to *E. coli*. To further investigate the hypothesis of other processes that will result in detection of substitution of amino acid, there

are some approaches that could be applied, separately or together. One is to apply translation error detection pipeline on many organisms, species, and cells from many sources and to understand how the ratios change and what are the differences and the common for different samples. One such sample could be germline cells or studying specifically X chromosomes in samples from male, those two are examples of haploid human sources for translation errors and could be used as a control. Another approach to understand more about the factors that contribute to the high intensities ratio is to harness recent technologies in LC-MS analysis, as data independent acquisition (DIA)[33], retention time prediction[34,35], as well as MS/MS peak intensity prediction[36,37], to validate that the peptides detected are indeed with substitutions, and not the result of something else (e.g., PTMs). I also found out that spectra counts is the most appropriate way to quantify proteins in the data I analyzed, and perhaps a more accurate way of quantifying translation errors should be thought of, based in this knowledge.

Through my analysis, I estimated the frequency and characterized the pattern of translation errors and compared them to what is known in the literature[1,3,9]. Although I could not narrow down the wide range of rate of translation error, I found out that our results confirm with previous estimations. I also explored whether these errors occurred randomly or in specific proteins and sites by examining the identity of proteins with translation errors across different tissues and evaluating the conservation of positions with translation errors. My findings suggested that certain proteins and sites were more prone to errors than others. When I compared the proteins that have errors between tissues, I found out that there are proteins that tend to have errors in many tissues. As suggested in other researchs[4,11], this emphasizes that translation errors are dependent on their context and do not occur randomly. Furthermore, the analysis of the evolutionary conservation of the sites of translation errors suggests that these sites tend to be more variable across species. Taken together, this suggests that the tendency to have errors is genetically encoded. Moreover, the fact that these errors tend to occur at protein spots that are less sensitive to amino acid substitutions across evolution indicates that they may not necessarily be harmful to the overall protein structure or function. It suggests that the cell has mechanisms in place to prevent translation errors from occurring, but these mechanisms may not be perfect and may allow errors to occur at less

important sites without significant consequences. It also supports the idea that translation errors may have a role in evolution by introducing genetic diversity without harmful effects.

In more applicative approach, I used the results of the lab's translation errors detection pipeline to generate sequences of SARS-Cov2 Spike protein that will produce products with errors. Such sequence could be used for designing vaccine that will target multiple strains, and possibly enhance immune reaction. The current pipeline could not find translation errors in the redesigned sequence that was expressed as DNA in human cell line, nor in the original sequence of the Spike (Wuhan strain) that were expressed either as DNA or RNA. However, some errors were found using a more targeted approach rather than the pipeline's open search. The information that was achieved was not sufficient to indicate whether the redesigned sequence is more error prone than the original sequence, but it showed that errors can occur during the translation of the Spike protein.

Overall, my study contributes to our understanding of translational errors in the human context and highlights the importance of considering the diploid nature of human cells when analyzing data. The findings may have implications for future research on translational errors in human tissues and may inform strategies to minimize or utilize these errors.

Acknowledgements

I would like to express my gratitude to the individuals who have supported me throughout my research journey. First and foremost, I am deeply grateful to my mentor, Prof. Yitzhak Pilpel, for his invaluable guidance, encouragement, and stimulating discussions that have greatly enriched my work. I would also like to thank Dr. Orna Dahan for her endless help, patience, and invaluable advice, as well as Dr. Noa Hefetz-Aharon, who contributed significantly to the project. My sincere thanks go to all the members of the lab for their valuable insights. I also would like to acknowledge the collaborative efforts of the Yardena Samuels Lab, Tami Gaigers Lab, and Noam Stern-Ginossar Lab. Special thanks go to Shira Albeck and Tami Unger from the Life Sciences Core Facilities at Weizmann for their outstanding contributions to this work. Finally, I would like to express my heartfelt appreciation to my family, especially my wife Hadar Buium. I couldn't do it without you.

Literature

1. Parker, J. Errors and alternatives in reading the universal genetic code. *Microbiol. Rev.* **53**, 273–298 (1989).
2. Ogle, J. M. & Ramakrishnan, V. Structural Insights into Translational Fidelity. *Annu. Rev. Biochem.* **74**, 129–177 (2005).
3. Allan Drummond, D. & Wilke, C. O. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.* **10**, 715–724 (2009).
4. Mordret, E. *et al.* Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity. *Mol. Cell* **75**, 427-441.e5 (2019).
5. Bratulic, S., Toll-Riera, M. & Wagner, A. Mistranslation can enhance fitness through purging of deleterious mutations. *Nat. Commun.* **8**, 15410 (2017).
6. Miranda, I. *et al.* Candida albicans CUG mistranslation is a mechanism to create cell surface variation. *mBio* **4**, e00285-13 (2013).
7. Yanagida, H. *et al.* The Evolutionary Potential of Phenotypic Mutations. *PLOS Genet.* **11**, e1005445 (2015).
8. Whitehead, D. J., Wilke, C. O., Vernazobres, D. & Bornberg-Bauer, E. The look-ahead effect of phenotypic mutations. *Biol. Direct* **3**, 18 (2008).
9. Kramer, E. B. & Farabaugh, P. J. The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA* **13**, 87–96 (2007).
10. Romero Romero, M. L., Landerer, C., Poehls, J. & Toth-Petroczy, A. Phenotypic mutations contribute to protein diversity and shape protein evolution. *Protein Sci.* **31**, e4397 (2022).
11. Liu, Y. *et al.* Mistakes in translation: Reflections on mechanism. *PLOS ONE* **12**, e0180566 (2017).
12. Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, e8503 (2019).

13. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C. & Yates, J. R. I. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem. Rev.* **113**, 2343–2394 (2013).

14. Liigand, P., Kaupmees, K. & Kruve, A. Influence of the amino acid composition on the ionization efficiencies of small peptides. *J. Mass Spectrom.* **54**, 481–487 (2019).

15. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).

16. Becker, E. *et al.* The protein expression landscape of mitosis and meiosis in diploid budding yeast. *J. Proteomics* **156**, 5–19 (2017).

17. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. *Genome Res.* **9**, 677–679 (1999).

18. Carey, L. B. RNA polymerase errors cause splicing defects and can be regulated by differential expression of RNA polymerase subunits. *eLife* **4**, e09945 (2015).

19. Traverse, C. C. & Ochman, H. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proc. Natl. Acad. Sci.* **113**, 3311–3316 (2016).

20. dos Reis, M., Wernisch, L. & Savva, R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Res.* **31**, 6976–6985 (2003).

21. Sabi, R. & Tuller, T. Modelling the efficiency of codon-tRNA interactions based on codon usage bias. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* **21**, 511–526 (2014).

22. Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 3645–3650 (2010).

23. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet. TIG* **29**, 569–574 (2013).

24. Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.* **21**, 1781–1791 (2004).

25. Schmidt, T. G. M. *et al.* Development of the Twin-Strep-tag® and its application for purification of recombinant proteins from cell culture supernatants. *Protein Expr. Purif.* **92**, 54–61 (2013).
26. Robinson, N. e. *et al.* Structure-dependent nonenzymatic deamidation of glutaminyl and asparaginyl pentapeptides. *J. Pept. Res.* **63**, 426–436 (2004).
27. Robinson, A. B. & Rudd, C. J. Deamidation of Glutaminyl and Asparaginyl Residues in Peptides and Proteins. in *Current Topics in Cellular Regulation* (eds. Horecker, B. L. & Stadtman, E. R.) vol. 8 247–295 (Academic Press, 1974).
28. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
29. Kapur, M. & Ackerman, S. L. mRNA Translation Gone Awry: Translation Fidelity and Neurological Disease. *Trends Genet. TIG* **34**, 218–231 (2018).
30. Martinez-Miguel, V. E. *et al.* Increased fidelity of protein synthesis extends lifespan. *Cell Metab.* **33**, 2288-2300.e12 (2021).
31. Gabay, O. *et al.* Landscape of adenosine-to-inosine RNA recoding across human tissues. *Nat. Commun.* **13**, 1184 (2022).
32. Ruiz Cuevas, M. V. *et al.* Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep.* **34**, 108815 (2021).
33. Doerr, A. DIA mass spectrometry. *Nat. Methods* **12**, 35–35 (2015).
34. Bouwmeester, R., Gabriels, R., Hulstaert, N., Martens, L. & Degroeve, S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. *Nat. Methods* **18**, 1363–1369 (2021).
35. Moruz, L. & Käll, L. Peptide retention time prediction. *Mass Spectrom. Rev.* **36**, 615–623 (2017).
36. Degroeve, S. & Martens, L. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* **29**, 3199–3203 (2013).
37. Lin, Y.-M., Chen, C.-T. & Chang, J.-M. MS2CNN: predicting MS/MS spectrum based on protein sequence using deep convolutional neural networks. *BMC Genomics* **20**, 906 (2019).