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
Rehovot, Israel

עבודת גמר (תזה) לתואר
מוסמך למדעים

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

By
Avia Yehonadav

מאת
אביה יהונדב

זיהוי וניתוח של טעויות תרגום בחיידקים כנגד אינדוקציה של לחצים סביבתיים

Detection and Analysis of Translation Errors in *E. Coli* Upon Induction of Environmental Perturbations

Advisor:
Yitzhak Pilpel

מנחה:
יצחק פלפל

April 2nd 2018

י"ז בניסן ה'תשע"ח

1. Abstract

In this work we strived to establish how translational errors are affected by different environmental and genetic constraints. In particular we tested the cells' response to amino acid deprivation and growth in different cell culture conditions. The translation process is one of the few key mechanisms in which any living cell is highly dependent on. It requires a great amount of system components and heavily relies on energy consumption; once translation is initiated, maintaining it demands the breakage of three high energy bonds (GTP/ATP) to bind a single amino acid to the elongating peptide, making it the most costly process in bacteria cells (Russell & Cook, 1995). Nevertheless, the fidelity of the translation machinery is relatively low compared to other the fidelity of other stages in the central dogma.

Using a novel mistranslation detection tool, we show for the first time the effects of amino acid starvation on the translation accuracy throughout the proteome. Furthermore, we characterized the nature of detected mistranslation events, allowing us to determine whether the mistranslation was caused by a faulty codon-anticodon coupling on one hand, or by a misconducted charging of aminoacyl tRNA synthetase on the other. We have discovered that amino acid starvation takes a substantial toll on the translational fidelity. As expected, many of the translation errors that were detected in starved cells occur in positions in proteins in which they needed to encode the same amino acid for which they were starved. Interestingly, this was not always the case as some starvation conditions seem to have a compensation mechanism, allowing the cells to avoid excessive amount of mistranslation errors in the depleted amino acid positions. Remarkably, we discovered that the different environmental stresses can give rise to similar outcomes – in particular, in every stress examined we detect an abundance of Threonine to Serine amino acid mistakes, and error that perhaps might be caused by mischarging on the later amino acid on tRNAs of the former. Our results suggest that mistranslation is sensitive to growth perturbations of the cell and might have an important regulatory role in response to stress.

Contents

List of abbreviations

AA	Amino Acid
aaRS	aminoacyl tRNA Synthetase
HPLC-MS/MS	High-performance Liquid chromatography–mass spectrometry
RT	Retention time
PTM	Post translational modification
YFP	Yellow Fluorescent Protein
DP	Dependent peptides
BP	Base peptide
OD	Optical density
WT	Wild type
FDR	False discovery rate

2. Introduction

2.1 Error rates in central dogma

Genetic information propagation along the Central Dogma is subject to errors in DNA replication, RNA transcription and protein translation. DNA replication typically manifests the highest fidelity among these processes, featuring genetic mutation rate on the order of 10^{-10} per nucleotide per genome doubling (1,2). "Phenotypic mutations", i.e. errors in RNA and in protein translation, in which the wrong RNA nucleotide or amino acid are respectively incorporated, occur at considerably higher rate. A recent estimate made in bacteria, is that transcription error rate ranges between 10^{-5} to 10^{-6} per incorporated nucleotide (3). As for translation, the classical kinetic proofreading theory (4) suggested that the error rate per amino acid would have been extremely high (10^{-2}) at chemical equilibrium, and it is only due to the investment of energy in the form of hydrolysis of GTP that it can be reduced to about 10^{-4} on average.

Advances in the fields of DNA and RNA sequencing have allowed researchers to study in great detail the errors inherent to the processes of DNA replication and RNA transcription. The errors which characterize the third main process of the central dogma, translation, continued to remain difficult to observe. Understanding how these errors happen and the factors that govern them could bear critical insights about the biology of the translation machinery, and help scientists uncover the rules that govern protein plasticity and phenotypic adaptation.

2.2 Advances in protein errors detection

An early effort to detect protein mistranslation was presented by Edelmann and Gallant in 1977. In this study, the incorporation of a radioactively labeled cysteine that was translated from a cysteine-free protein was measured (Edelmann & Gallant, 1977).The results of this experiment lead to the first measurement of the translation error rate. Since then, molecular genetics strategies based on reporter constructs have permitted to monitor specific error rates at defined positions in the genome. However, these studies did not offer a large scale image of the errors in the natural proteome, and so the understanding of the effects of mistranslation in vivo remained vague.

Our lab recently developed an experimental/computational pipeline to detect amino acid substitutions from HPLC-MS/MS data in a high-throughput manner. With this tool we observed for the first time a substantial amount of mutated protein positions - the products of inaccuracies in translation in living cells of *E. coli* and *S. cerevisiae*. The newly obtained method offers a broad coverage of mistranslation detection, improving the number of mistranslated positions detected by orders of magnitude compared to previous detection attempts. Moreover, the method allows us to quantify the frequency of different mistranslated events in the same protein position, conveying an additional dimension to our data.

2.3 Computational pipeline for systematic detection of amino acid substitutions

The computational pipeline relies on whole proteome data sets obtained from HPLC-MS/MS instruments. Whole proteomes which were derived from cell lysates were trypsinized resulting in 8-10 amino acids long peptides. The output from the machine contains two important data sets about each peptide- retention time (RT) and mass-charge ratio (m/z). While the RT value is mostly affected by the hydrophobicity of the tested peptide, the m/z value is mostly affected by the mass of the peptide. For most peptides, both RT and m/z are sufficiently predictable, allowing to map each of them to a certain protein and thus determine the proteome pattern that characterizes a condition of interest. Nevertheless, some peptides do not agree with the standard prediction (determined by the genetic code or cDNA data), these are marked as dependent peptides (DPs). In contrast to DPs, base peptides (BPs) can be accurately mapped onto the proteome according to their RT and m/z values. If a certain amino acid substitution occurs in a given peptide, we can use the difference between the hydrophobicity of the original and new amino acid to predict a change in RT, and examine if the expected change was indeed observed. The m/z value would shift due to a difference in mass between the original and new amino acids. Our algorithm is designed to detect shifts in mass that can only be explained by amino acid substitution in a given peptide. Other mass shifts, like those who arise by post translational modifications (PTMs) are filtered out of the DPs pool and are regarded as negatives. In a case where a certain mass shift can be explained by both a PTM and a mistranslation event - we discard the observation in order to avoid false positives.

Once our mass spectrometry data is ran through the pipeline we attain a valuable data set of predicted amino acid substitutions. Analysis of the data set can be utilized to shed light on the relative contribution of aminoacyl tRNA Synthetase (aaRS) and ribosomes to translation errors. Furthermore, detecting and quantifying amino acid substitutions across the proteome may reveal evolutionary constraints imposed

by translation infidelity and allow us to ask how organisms locally modulate their error levels, and the strategies they employ to mitigate their deleterious effects.

2.4 Protein error causes and translation system components

The translation machinery is composed of a number of essential components for successful protein synthesis; the ribosome and its satellite proteins, charged tRNA molecules and an accurate transcript of the genetic code. Against this background, mistranslation may occur due to one of the following defects: 1. the ribosome failing to discriminate against imperfect codon-anticodon association in its A-site (ribosome error, or "mispairing") or 2. a faulty charging of an amino acid with an unfitting tRNA molecule (aminoacyl tRNA Synthetase error, or "mischarging").

The accuracy of both processes is amplified by kinetic proofreading, a general mechanism during which the addition of an irreversible, energy consuming step to a reaction permits the system to reach discrimination levels that are inaccessible at thermodynamic equilibrium. Theoretical models demonstrated that processes relying on kinetic proofreading increase their accuracy were necessarily, subject to a trade-off between incorporation speed, accuracy and energetic cost.

2.5 Perturbing the translational machinery

First attempts in testing the computational pipeline are based on comparing normal growth conditions of bacteria against drugged bacterial cells. These drugs target the ribosome and are expected to perturb its fidelity and thus create a stronger mistranslation signal compared to cells who grew in normal conditions. Once this was established, we pursued finding additional ways to provoke the translation machinery into creating an excessive amount of translation error. Choosing such conditions should be done with great attention, as some perturbations might yield a strong signal but are too stochastic in their nature to gather an adequate understanding of the mechanisms underlying the mistranslation events. For instance, cells which grow in extremely high temperatures are bound to have kinetic deformities that affect many enzymatic reactions, thus it would be impractical to study the causes of mistranslation events against this background. And so we carefully considered possible perturbations which would yield high protein error rate while remaining informative.

Finally we decided to starve bacteria for a subset of amino acids in 4 independent experiments. Our starvation assays rely on experiments conducted by Cluzel et al (Subramaniam, Pan, & Cluzel, 2013) which aimed to test the robustness of accurate tRNA-amino-acid charging under amino acid deprivation.

This was done by constructing a plasmid library in which each plasmid consisted the genetic sequence of Yellow Fluorescent Protein (YFP). Each variant contained a modified YFP nucleotide sequence that bears only one type of codon for a specific amino acid. The variant S^{UCU} , for example, has only UCU codons in serine positions throughout the YFP mRNA sequence. Cluzel et al show that the YFP fluorescence intensity differs between variants while they are starved for their corresponding amino acid, suggesting that certain tRNA-AA charging events are more robust to AA starvation than others when compared to isoacceptors for the same AA. These findings lead us to consider that we might obtain a comparable data set by using identical conditions to those demonstrated in the Cluzel et al paper. By doing so, we hoped to predict codons which are more likely to be mistranslated given a certain condition.

3. Goals of the study

The goals of this study were to create a translation perturbation system that will allow us to acquire new information about the factors that govern the translation accuracy. To address this goal, we planned to:

- ➢ Starve bacteria cells to three different amino acids in three independent experiments, according to an established starvation assay.
- ➢ Detect and characterize translation errors that arose upon amino acid starvation using a novel, high throughput HPLC-MS/MS method followed by a computational pipeline.
- ➢ Portray a particular mistranslation pattern to better define the complexion of different perturbation.
- ➢ Identify mistranslation-susceptible positions in the genome by measuring their individual translation error rate.

4. Methods

Strains and growth conditions

For the serine starvation dataset, BW25113 (WT) and JW2880-1 (ΔserA, obtained from the Keio deletion library) cells were plated on LB agar and incubated at 37°C overnight. 2 colonies of each strain were picked and grown in 3 ml of modified MOPS rich defined medium made according to Cluzel et al recipe (SI Appendix) and incubated at 37°C until stationary phase. BW25113 and JW2880-1 cell cultures were diluted 1/1000 and grown aerobically in

220 ml of modified MOPS rich defined medium and MOPS serine starvation medium accordingly in 500 ml Erlenmeyer flasks at 37°C (mediums were made according to Cluzel et al 2012 SI Appendix).

Proteome extraction

We adapted our proteome extraction protocol from Khan *et al.,* 2011[28]. Samples were each split into two 50 ml falcon tubes, centrifuged at 4000 rpm for 5 min, and washed twice with PBS (add 10 ml PBS, vortex, centrifuge for 5 min). Remaining PBS was vacuumed and the pellets were frozen in ethanol-dry ice. Pellets were re-suspended in 1 ml of B-PER bacterial protein extraction buffer (Thermo Fisher Scientific), pooled together, and vortexed vigorously for 1 min. The mixture was centrifuged at 13,000 rpm for 5 min. The supernatant (high solubility fraction) was collected and frozen in an ethanol-dry ice bath. The pellet was re-suspended in 2 ml of 1:10 diluted B-PER reagent. The suspension was centrifuged and washed one more time with 1:10 diluted B- PER reagent. The pellet was re-suspended in 1 ml of Inclusion Body Solubilization Reagent (Thermo Fisher Scientific). The suspension was vortexed for 1 min, shaken for 30 min, and placed in a sonic bath for 10 min at maximum intensity. Cellular debris was removed from the suspension by centrifugation at 13,000 rpm for 15 min. The supernatant was frozen in an ethanol-dry ice bath (low solubility fraction).

SCX fractionation, HPLC and Mass Spectrometry

400µg of protein was taken for in-solution digestion and processed by Filter aided sample preparation (FASP)[49] protocol using 30k Microcon filtration devices (Millipore). Proteins were subjected to on-filter tryptic digestion for overnight at 37°C and the peptides were fractionated using strong cation exchange (SCX) followed by desalting on C18 StageTips[50] (3M EmporeTM, St. Paul, MN, USA). Peptides were analyzed by liquid-chromatography using the EASY-nLC1000 HPLC coupled to high-resolution mass spectrometric analysis on the Q-Exactive Plus mass spectrometer (ThermoFisher Scientific, Waltham, MA, USA). Peptides were separated on 50 cm EASY-spray columns (ThermoFisher Scientific) with a 140 min gradient of water and acetonitrile. MS acquisition was performed in a data-dependent mode with selection of the top 10 peptides from each MS spectrum for fragmentation and analysis

Computational methods

Raw files were processed with MaxQuant v. 1.5.5.1. The list of parameters is available in the supplementary materials. High and Low solubility fractions were aligned separately. The amino acid substitutions identification procedure relies on the built-in dependent peptide algorithm of MaxQuant.

The Dependent Peptide search

Experimental spectra are first searched using a regular database search algorithm, without any variable modification, and the significance of identifications is controlled to a 1% FDR via a target decoy procedure. Identified spectra are then turned into a spectral library, and a decoy spectral library is created by reversing the sequences of the identified spectra. For each possible pair consisting of an identified spectrum in the concatenated spectral libraries and an unidentified experimental spectrum of the same charge, and recorded in the same raw file, we apply the following steps : first we compute the mass shift Δm by subtracting the mass of the identified (unmodified) spectrum to that of the unidentified (modified) spectrum, then we simulate modified versions of the theoretical spectrum by adding *in silico* this mass shift at every position along the peptide, and finally we evaluate the match between the theoretical spectrum and the experimental spectrum using a formula similar to Andromeda's binomial score.

For each unidentified peptide, the match with the best score is reported, the nature of the match (target or decoy) is recorded, and a target-decoy procedure[51] is applied to keep the FDR at 1%. Peptides identified using this procedure are called Dependent Peptides (DP), whereas their unmodified counterparts are named Base Peptides (BP).

Additionally, the confidence of the mass shift's localization is estimated using a method similar to MaxQuant/Andromeda's PTM Score strategy, which returns the probability that the modification is harbored by any of the peptide's amino acid.

DP identifications filtering

The list of all known modifications was downloaded from www.unimod.org, and those marked as AA substitution, Isotopic label or Chemical derivative were excluded. Entries in this list are characterized by a monoisotopic mass shift, and a site specificity (i.e. they can only occur on a specific amino acid or on peptides' and proteins' termini). We removed from our analysis any DP identification that could be explained by any of the remaining

modifications, using the following criteria: the recorded Δm and the known modification's mass shift must not differ by more than 0.01 Da, and the modification must be harbored by a site consistent with the uniprot entry with a probability $p \geq 0.05$. Conversely, we computed the list of all possible amino acid substitutions and their associated mass shifts. For every substitution, we only retained DP identifications such that the observed Δm and the AA substitution's mass shift did not differ by more than 0.005 Da, and the mass shift was localized on the substitution's original AA with $p \geq 0.95$.

Among the remaining DP identifications, those such that the peptide sequence after substitution was a substring of the proteome (allowing Ile-Leu ambiguities), were also removed, to prevent pairing of dependent peptides and base peptides between paralogs.

Finally, the FDR was controlled once again at 1% using the same procedure as above.

Error rate quantification

In order to assess the error rate we quantify and compare pairs of base and dependent peptides across many samples. For each independent substitution, we fetched the quantification profile of the base peptide from Max Quant's peptides.txt table, and similarly fetch the dependent peptide's quantification profile from the matchedFeatures.txt table. Whenever a peak has been detected and quantified for both the dependent and the base peptide, we estimate the translation error rate as the ratio of their MS1 intensities.

5. Results

5.1 Serine starvation assay as a system calibration experiment

We wanted to generate an informative data set obtained by harvesting proteomes from both WT and AA-auxotroph strains grown on either AA-limited or complete media (materials and methods). Hoping this would allow us to compare basal mistranslation patterns made in WT cells to starvation-driven mistranslation patterns made in starved auxotroph cells. Subsequently to our decision to follow Cluzel et al's experimental setup we tested which of the starvation assays that appeared on the paper will be most relevant for our system. The starvation assays demonstrated in the Cluzel et al paper were for the following amino acids: isoleucine, serine, proline, phenylalanine, leucine, arginine, and glutamine. We obtained all of the auxotroph strains used for these experiments (the Keio deletion library) and grew

them in identical conditions as those exhibited in the paper's supplementary material. In each growth experiment, optical density (OD) of the broth was measured until cell colonies have reached stationary phase. Growth curves for each strain were produced and used to determine the physiological response of the mutants to their corresponding starvation media and growing conditions. Simultaneously, WT and mutant strains were grown in complete media for comparison and used as negative controls. Out of the seven possible amino acid starvation assays we have considered, a few were eliminated due to technical incompatibilities with our methods while others were eliminated due to lack of expected physiological response. Finally, we decided on serine, proline and isoleucine as three different sets of starvation assays.

Since our method involves costly and stringent mass spectrometry and proteomics procedures, we decided to run a calibration test which consists only one of the three starvation assays. Furthermore, we wanted to assess which physiological time point would yield the maximum amount of mistranslation events. A plausible answer for this concern is stationary phase. However, the current understanding of protein degradation in bacteria is yet to be established (Dougan, Mogk, & Bukau, 2002) making it harder to predict whether mistranslated proteins that were synthesized throughout the bacteria's growth will remain stable at stationary phase. For this reason, we decided to harvest cells in three different time points: early logarithmic phase, middle logarithmic phase, and stationary phase (Figure 1).

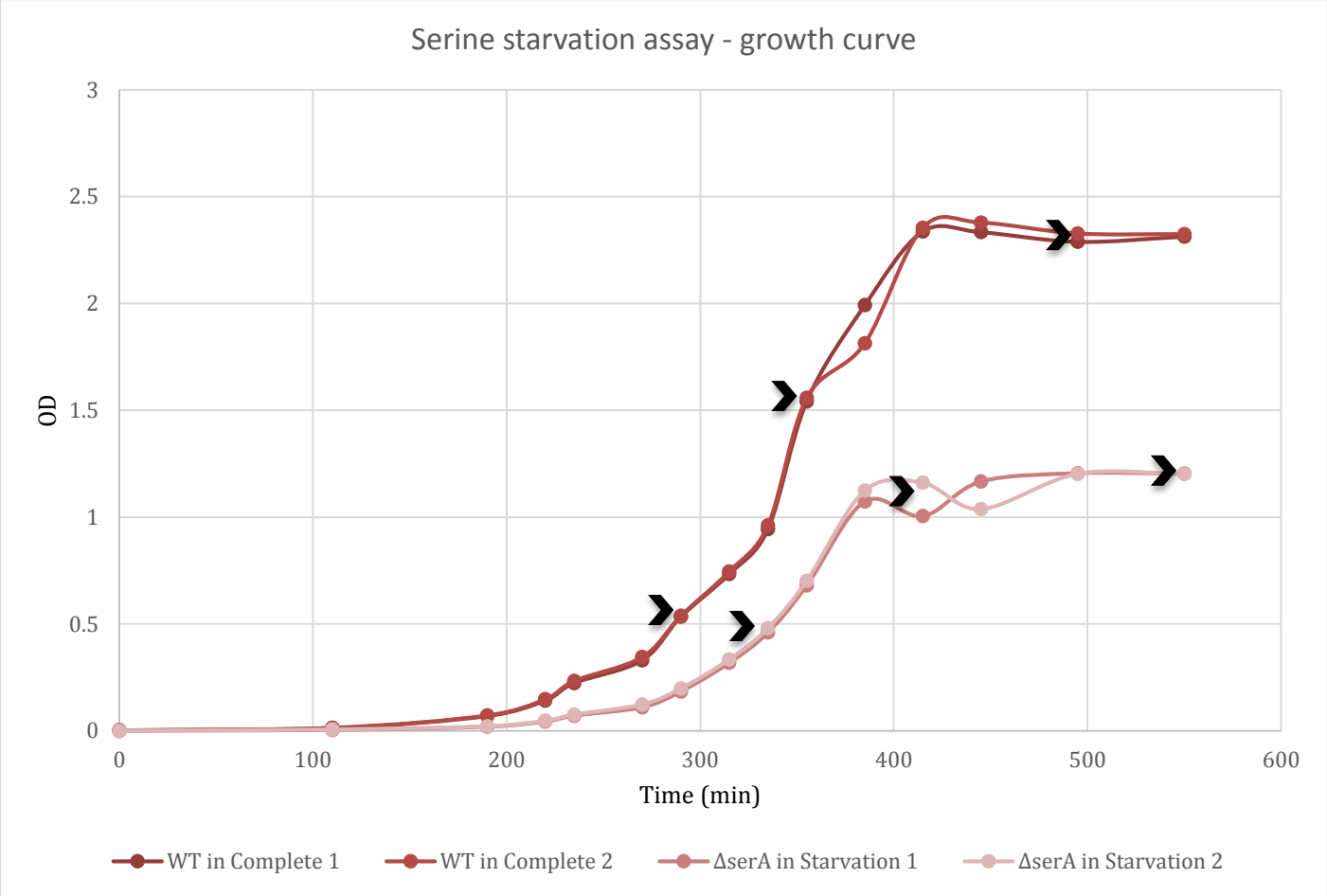


Figure 1 — WT strain grows more rapidly and reaches a higher OD than ΔserA grown in depleted Ser media. The mutant lags behind its WT strain in both final yield and growth rate. This is consistent in both biological repeats. The Black arrows represent cell harvesting points in each strain and condition.

By sampling from a few physiological stages we can conclude which time point will be best for cell harvesting in our future experiments in terms of mistranslation yield. Sampling in different time points was also used to test our speculation that some translation errors might accumulate over time.

Next, we chose serine to be the most suitable starvation assay for our calibration experiment. Serine has six different tRNA molecules that are utilized for its translation. Some of these tRNA molecules, in particular ACG and ACT isoacceptors, are more robust than others to *in-vivo* serine starvation and therefore manage to bind serine even though its free form abundance in the cell is relatively low. On the other hand, other serine-binding tRNA molecules TCT, TCA, TCG, TCC isoacceptors, display a significant depletion in their serine-binding ability (as suggested by Cluzel et al study). The redundancy of serine codons allows the above mentioned variability to persist and hints which codons are more prone to be mistranslated under starvation due to compatible and charged tRNA shortage.

5.2 Calibration experiment results

Two colonies of ΔserA (JW2880-1 obtained from the Keio deletion library at Yale Uni.(Baba et al., 2006)) were grown in starvation media while two colonies of its genetic background strain (referred to as WT, BW25113) were grown in complete media. OD measurements of all replicates were taken during their growth to determine the proper physiological stage for cell harvesting (Figure 1). Proteomes from all time points and replicates were extracted and prepared for HPLC-MS/MS proteomics assay (materials and methods). MS output was then processed in MaxQuant program and filtered in our computational pipeline to receive the results presented here.

The serine auxotroph lags behind its WT strain due to low growth rate. When stationary phase was reached, the auxotroph broth's had an OD of 1.2 (600nm) while WT had 2.3. Cells were harvested in three different physiological time points: early logarithmic phase, middle logarithmic phase, and stationary phase. When observing all mistranslation events which were derived from the same condition in different time points, we found that they did not vary significantly from one another. However, we learned that mistranslation events of type S to X, where X is a non-serine AA, become more abundant over time and with the exhaustion of the nutrients in the media.

Mistranslation events for samples from the same condition which were extracted in either high or low solubility buffers, didn't differ significantly from one another.

We have constructed a 64 by 19 matrix which we use as a mistranslation map. The rows represent the 64 codons of origin as they appear in the genome. The columns represent the 20 destination amino acids which were substituted instead of the AA encoded in the genome. The reason for having 19 instead of 20 columns is that isoleucine and leucine are pooled together as they are indistinguishable in this MS procedure. Cells in the matrix which cross with the codons to its AA are greyed out. This is also true for cells which cannot be distinguished with PTMs. The color scale of the matrix cells represents the number of events (i.e. unique positions within genes that showed a given replacement) measured for that particular substitution after a transformation to binary logarithmic scale (Figure 2).

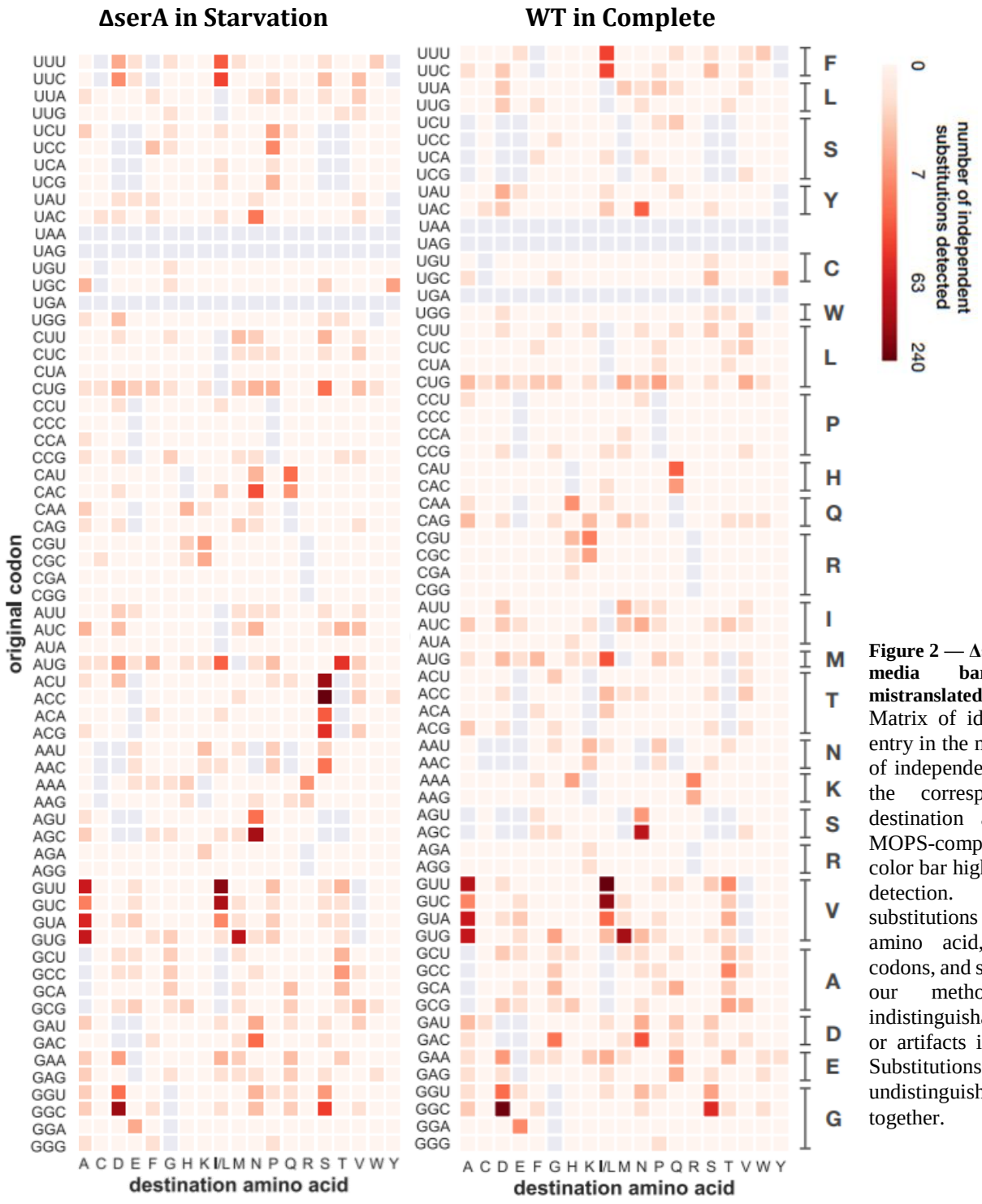


Figure 2 — ΔserA grown in serine depleted media bares more independent mistranslated peptides compared to WT. Matrix of identified substitutions. Each entry in the matrix represents the number of independent substitutions detected for the corresponding (original codon, destination amino acid) pair, in the MOPS-complete dataset. The logarithmic color bar highlights the dynamic range of detection. Grey squares indicate substitutions from a codon to its cognate amino acid, substitutions from stop codons, and substitutions undetectable via our method because they are indistinguishable from one of the PTMs or artifacts in the unimod.org database. Substitutions to Leu and Ile are a priori undistinguishable, and thus grouped together.

Overall, most of the substitutions detected in the starved auxotroph also appear in WT and in past mistranslation measurements of drugged bacteria cells (Bertels, Merker, & Kost, 2012). As expected, a prevalent distinction between starved auxotroph and WT is that substitutions of type S to X are highly abundant in starved auxotroph but not in WT cells. Moreover, as we predicted, S to X substitutions become more abundant over time in both number of substitutions detected (Figure 3) and error rate (Figure 4).

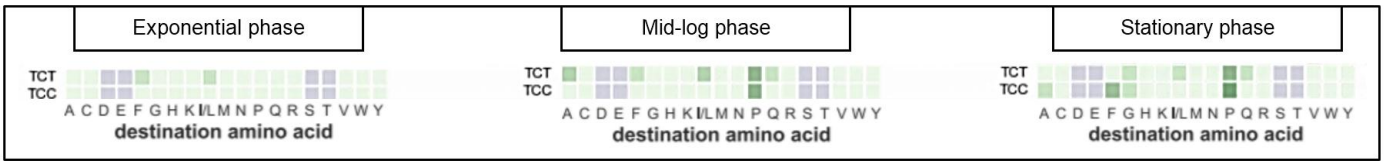
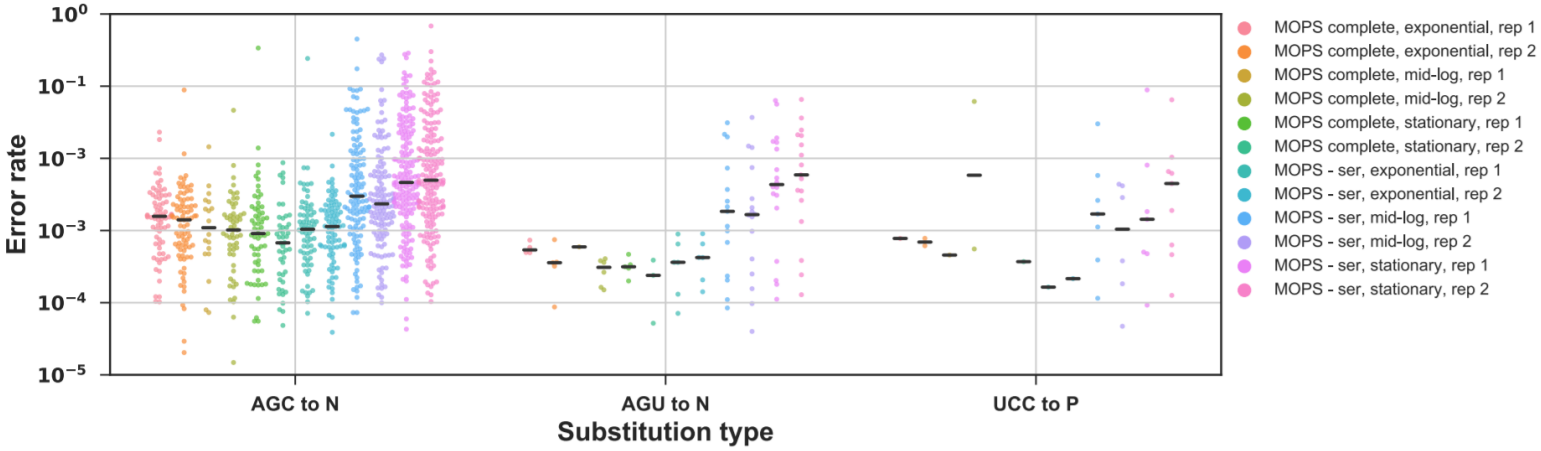


Figure 3 — number of independent peptides that bare S to X substitutions rise over time. Each matrix possesses the pooled detections from both biological repeats of the starved auxortroph. The serine codons TCT and TCC demonstrate an accumulative mistranslation effect in number of independent substitutions detected when comparing between time points. TCT and TCC are also two out of four codon that are prone to mistranslation upon starvation according to cluzel et al.

More prevalent substitution types which are relatively abundant in our data, do not demonstrate the accumulation affect (Figure 5). Additionally, the mistranslation map strikingly reveals that the most abundant substitution that appears in starved auxotroph but not in WT is of type T to S. This was an unexpected result and attempts to conform it fell short or were counterintuitive. Finally, we've hypothesized that the mild serine depletion in the starved auxotroph initiated a shift in the metabolic flux of the serine-glycine-threonine biosynthetic pathway, causing more glycine molecules to turn into serine and not threonine. If this was true, serine might have had a higher cellular abundance than threonine, which could trigger a faulty charging by threonine's aaRS that was shown to confuse S with



T in its editing site in certain conditions (Wu, Fan, & Ling, 2014).

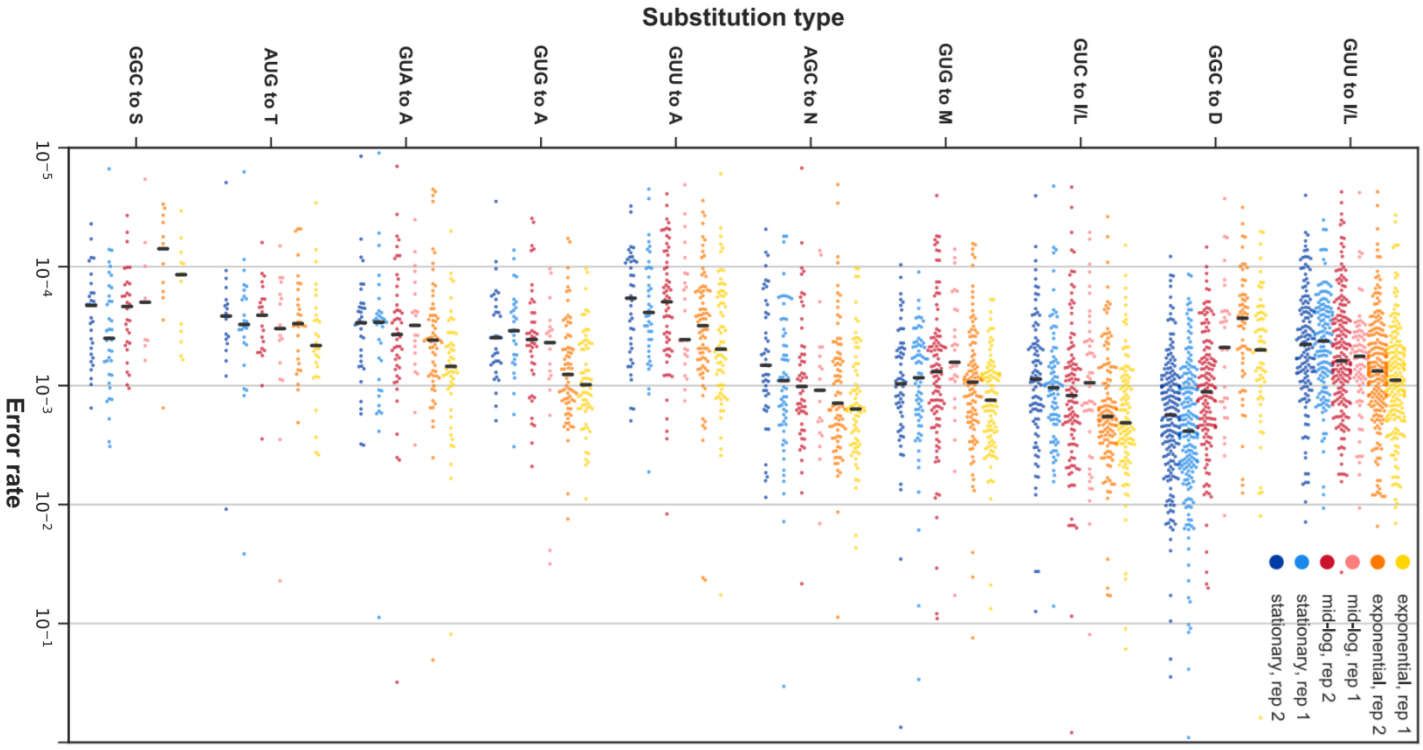


Figure 5 — The most abundant substitutions do not necessarily accumulate over time. For each of the top ten most frequently detected substitution types, we fetched the quantification profile of the dependent peptide and the base peptide. Each dot represents the ratio of intensities I_{DP}/I_{BP} for each of the samples, when both peaks have been detected and quantified. The black line indicates the medians of the distributions.

According to the results obtained by the calibration experiment, we have decided to design our future assays with a few modification. Firstly, as the mistranslation pattern is very similar between different time points in the same condition, we decided to harvest the cells only in the third time point. This will allow us to observe all the mistranslation events that occurred up to stationary phase. Secondly, we decided to add another negative control to our experiment design. As some of the results were curiously unexpected, it was difficult to pinpoint whether this phenotype is a product of a shift in the environment (complete versus starvation media) or a shift in genotype (WT versus auxotroph strains). Surely, some phenotypes can derive from the combination of shifts in both the environment and the genotype. In order to be able to discriminate possible phenotype casualties, we decided to run future experiments together with auxotroph grown in complete media. Finally, in order to test our metabolic flux

hypothesis, we decided to measure the internal abundances of the twenty amino acids in each of the conditions.

Upon construction of our experimental design according to our calibration experiment, we've launched our proline, isoleucine and serine experiments. As technical issues arose in the protein extraction and sample preparation process, causing low protein concentrations in high solubility buffer, we decided together with our collaborator to avert MS run for our high solubility samples and so the results shown are for low solubility samples only. This is unfortunate although judging from our calibration experiment, our proteome coverage was barely damaged.

5.3 Proline starvation assay

Growth measurements of the three different conditions, namely ΔproA in Pro starvation media, ΔproA in complete media and WT grown in complete media, were taken using OD reads in 600 nm wavelength. The starved proline auxotroph grew considerably slower than WT in complete. Additionally it reached only half of the yield when compared to WT and to ΔproA grown in complete media. The growth rate of starved auxotroph is similar to this found in the unstraved auxotroph (Figure 6).

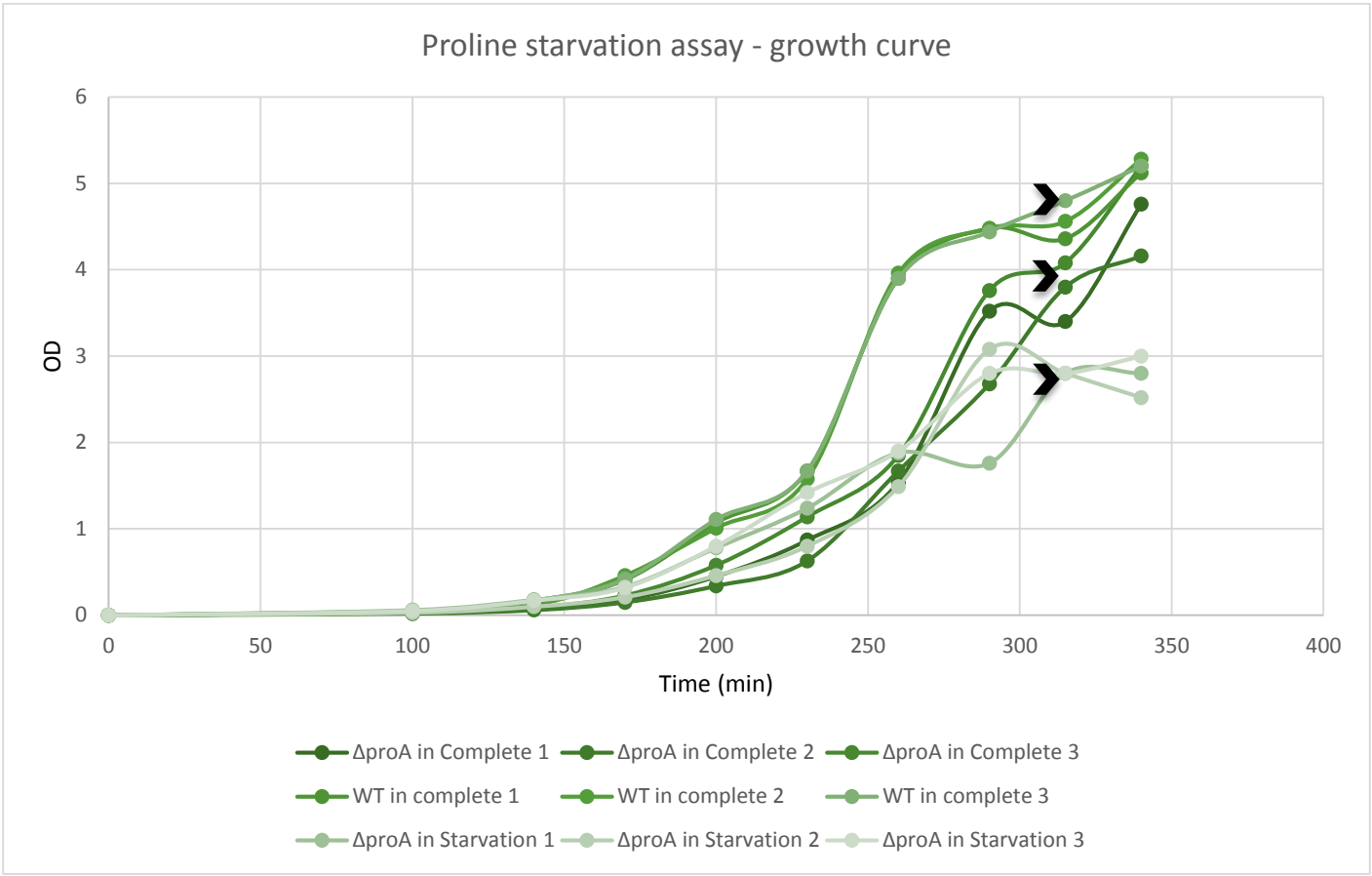


Figure 6 — WT strain grows more rapidly and reaches a higher OD than ΔproA grown in depleted Pro medium or in complete medium. The starved mutant lags behind its WT strain in both final yield and growth rate. Compared to its counterpart grown in complete media, the starved mutant lags behind inn yield but not in growth rate.

Proline has four codons and isoacceptors: CCG\A\T\C. The four codons vary in their relative frequency in the *E. coli* genome. CCG has a frequency of 2.67% per 1000 codons where the other three codons only reach up to 0.84%(Benson et al., 2013). According to the Cluzel et al study, the translation process of a recombinant YFP which all of its proline positions are encoded as CCG is the most robust to proline starvation. This suggests that there is a stringent regulation on the precise translation of the most abundant proline codon and thus insures high translation fidelity even in proline-depleted environment.When ΔproA is starved for proline it showed that CCA and CCC are the most error-prone codons while translating a recombinant YFP mRNA.

Proline auxotroph (ΔproA, JW0233-2) was grown simultaneously on complete and proline starvation media, together with its background strain (BW25113) grown in complete media. Unexpectedly, mistranslation events in which proline is substituted by a non-proline AA (P to X) were observed prominently in the data set derived from the unstraved auxotroph (Figure 7). In contrast, the starved

auxotroph hardly displayed P to X substitutions, making it harder to elucidate the source of such events. This is true for both error-rate and error-count values (Figure 8). Interestingly, CCG to X is the most dominant substitution of P to X in all condition. This is in contradiction to Cluzel et al results which show that CCG is the most robust isoacceptor out of all four.

Another striking result is the presence of T to S substitutions which were previously detected in serine starvation assay and were presumed to be associated with the biosynthetic pathway that threonine and serine share. The fact that proline does not partake in threonine and serine's biosynthesis (Bertels et al., 2012) hints that there might be a different, and more particular cause for T to S events.

Although T to S substitutions appear in all conditions, ΔproA grown in complete media has the highest T to S-error-rate in and mistranslated peptides count. Interestingly, this is in line with the fact that it's the only condition that shows P to X substitutions. From this we can conclude that proline auxotroph grown in complete media is more prone to mistranslation than when grown in proline starvation media. It seems that starved WT can sense the proline depletion and somehow compensate for its environmental proline depletion. Further experiments should be done to test the source of this bias.

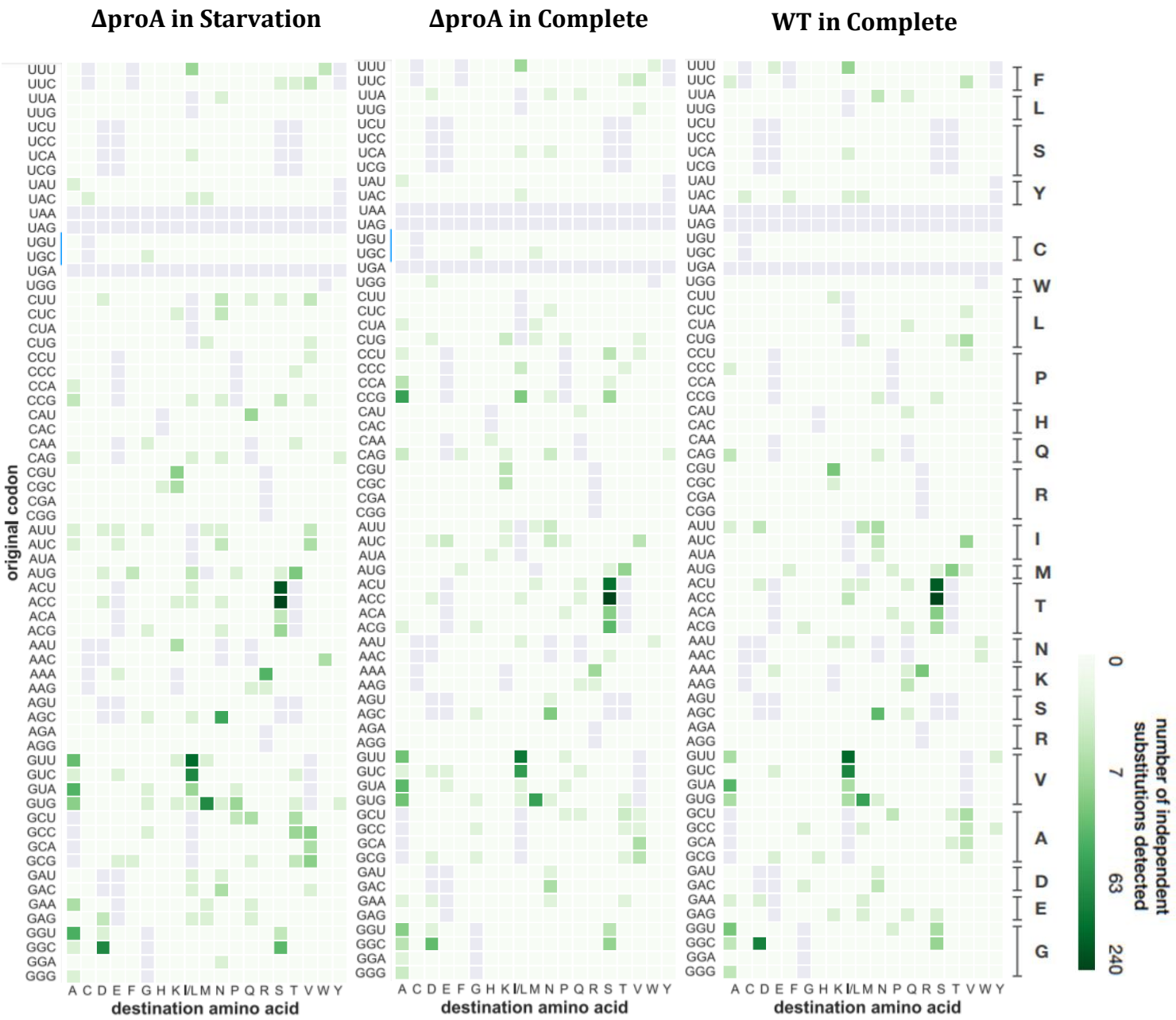


Figure 7 — Out of the three conditions in the proline assay, P to X substitutions are the most prevalent in ΔproA grown in starvation media. T to S substitutions appear in all conditions in a consistent manner. Abundant substitutions that appeared in our calibration experiment are persistent in to this conditions as well.

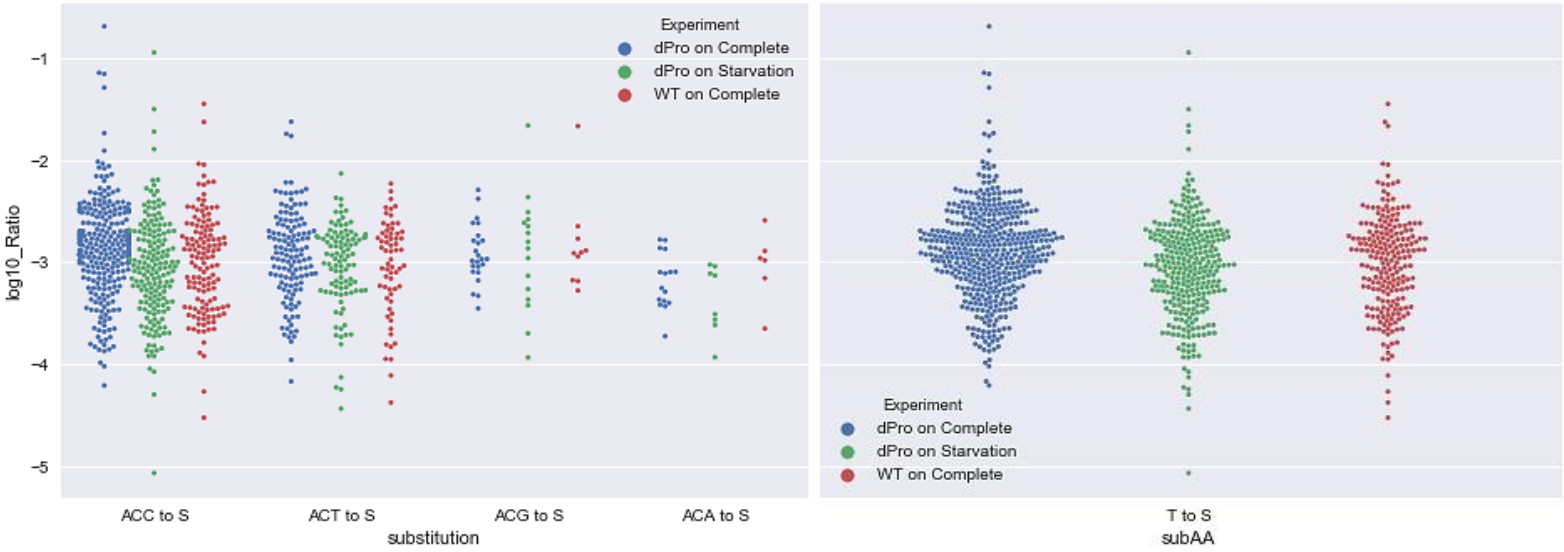


Figure 8 — S to X substitution is the most prevalent substitution in the proline starvation assay dataset. This makes it a suitable study case for comparison between conditions and experiments. Here we see that ΔproA in complete media has the largest amount of S to X substitutions, implying that this is the most erroneous condition out of the three presented. The left figures contains a breakdown of all codon to AA substitutions while the right figure contains pooled T to S substitutions. Note that ACC and ACT are the most common used codon of Thr in the genome.

5.4 Isoleucine starvation assay

Growth curves of the different conditions—ΔilvA (JW3745-2) in complete media, ΔilvA in Ile starvation media, and WT grown in complete media, demonstrate an apparent difference from one another. WT grown in complete media has, as expected, the highest growth rate and yield among the three conditions. The auxotroph ΔilvA grown in complete media lags behind the WT in both growth rate and yield. When ΔilvA grew in Ile starvation media, it demonstrated a very similar growth rate to ΔilvA grown in complete media but reaches about half of the yield (Figure 9).

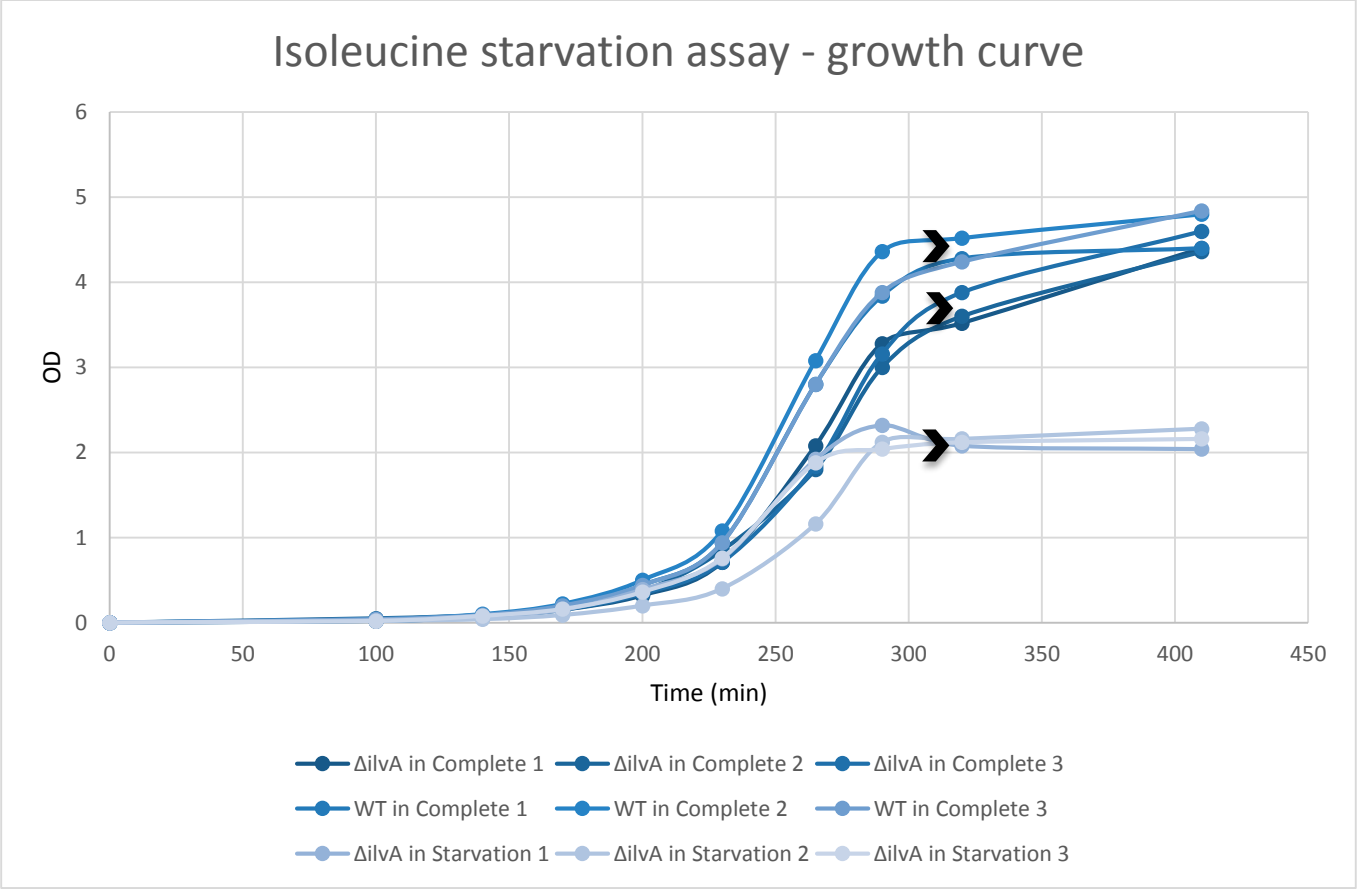


Figure 9 — S to X substitution is the most prevalent substitution in the proline starvation assay dataset. This makes it a suitable study case for comparison between conditions and experiments. Here we see that ΔproA in complete media has the largest amount of S to X substitutions, implying that this is the most erroneous condition out of the three presented. The left figures contains a breakdown of all codon to AA substitutions while the right figure contains pooled T to S substitutions. Note that ACC and ACT are the most common used codon of Thr in the genome.

There are three isoleucine codons in the genetic code - ATT\C\A. Their frequencies across the genome are 3, 1.8 and 0.7 percent, respectively. As ATT is the most prevalent codon out of the three, its isoacceptor's robustness to isoleucine starvation is correspondingly high (Cluzel et al).

As expected, starved ΔilvA is enriched with I to X substitutions and absent in the two other conditions. Starved ΔilvA also bares the highest error rate and error count in all mistranslation types. The mistranslation pattern (Figure 10) of WT and ΔilvA grown on complete media is similar to previous results obtained from WT samples.

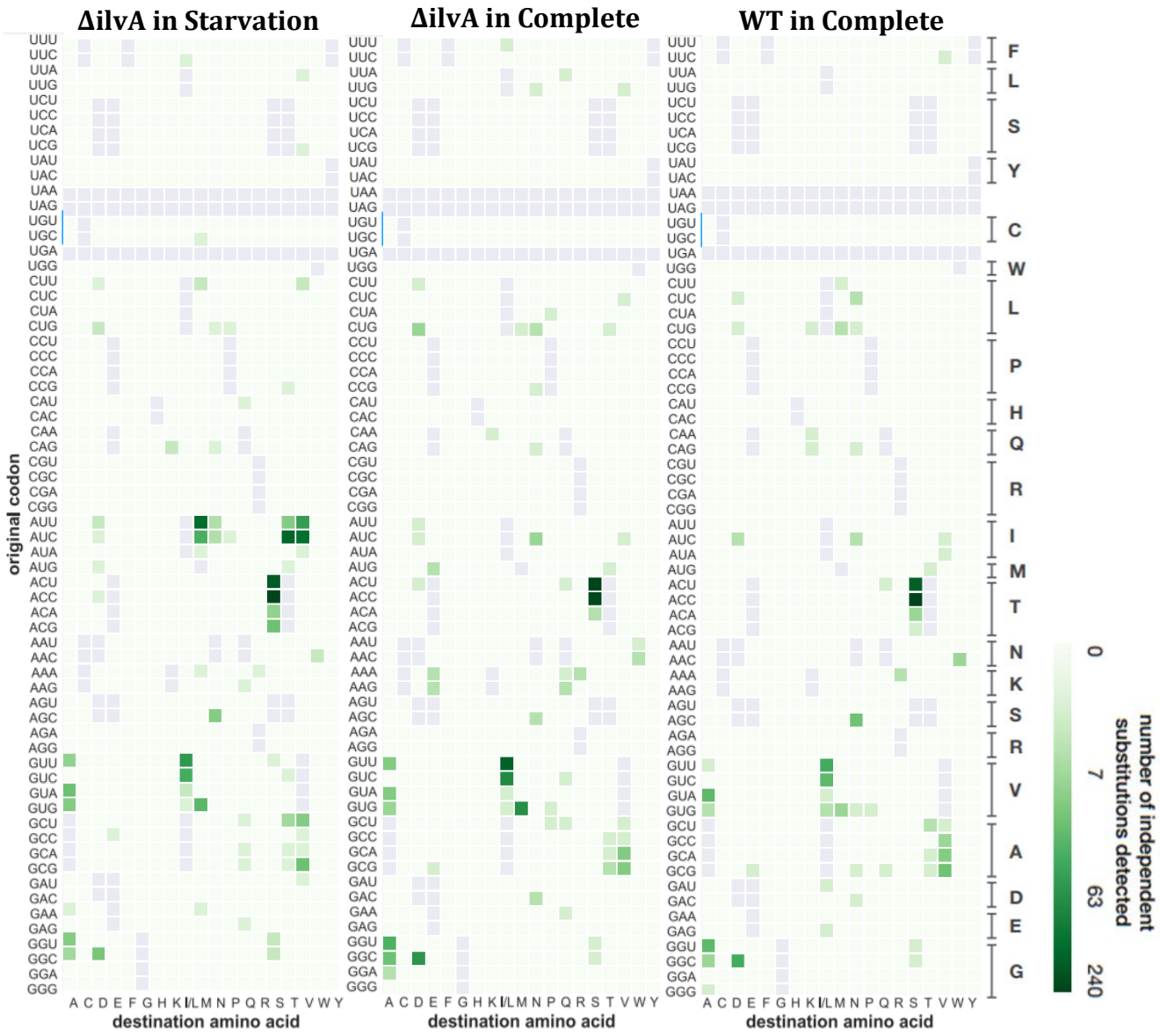


Figure 10 — S to X substitution is the most prevalent substitution in the proline starvation assay dataset. This makes it a suitable study case for comparison between conditions and experiments. Here we see that ΔproA in complete media has the largest amount of S to X substitutions, implying that this is the most erroneous condition out of the three presented. The left figures contains a breakdown of all codon to AA substitutions while the right figure contains pooled T to S substitutions. Note that ACC and ACT are the most common used codon of Thr in the genome.

According to our previous results near-cognate errors compose 88% percent out of all detectable mistranslation event. Consistently, all I to X substitutions are of type near-cognate. As T to S is the most abundant error in all conditions it can be used as a study case to compare between conditions. For example, starved ΔilvA has the highest T to S error-rate (Figure 11). Ile is mostly substituted by Thr, Met, and Val – near cognate substitutions. When codon-anticodon mispairing is on the line, the translation machinery is biased to mispair codons which follow a certain position-nucleotide relationship. For instance, uracil is significantly more likely to be mistaken with guanine in the third position rather than in the second position (Mordret et al., 2018). This bias only seems to be apparent in I to M substitution, as the Ile codons ATTile and ATCile are both near cognate to the Met codons ACTmet and ACCmet. Yet, we mostly observe ATTile to M as the translation machinery is confusing ATTile with ACTmet but less likely to do so with ATCile and ACCmet. In a closer look, this bias cannot be explained by the 4 by 4 matrices; more on this in the discussion section (Figure 12).

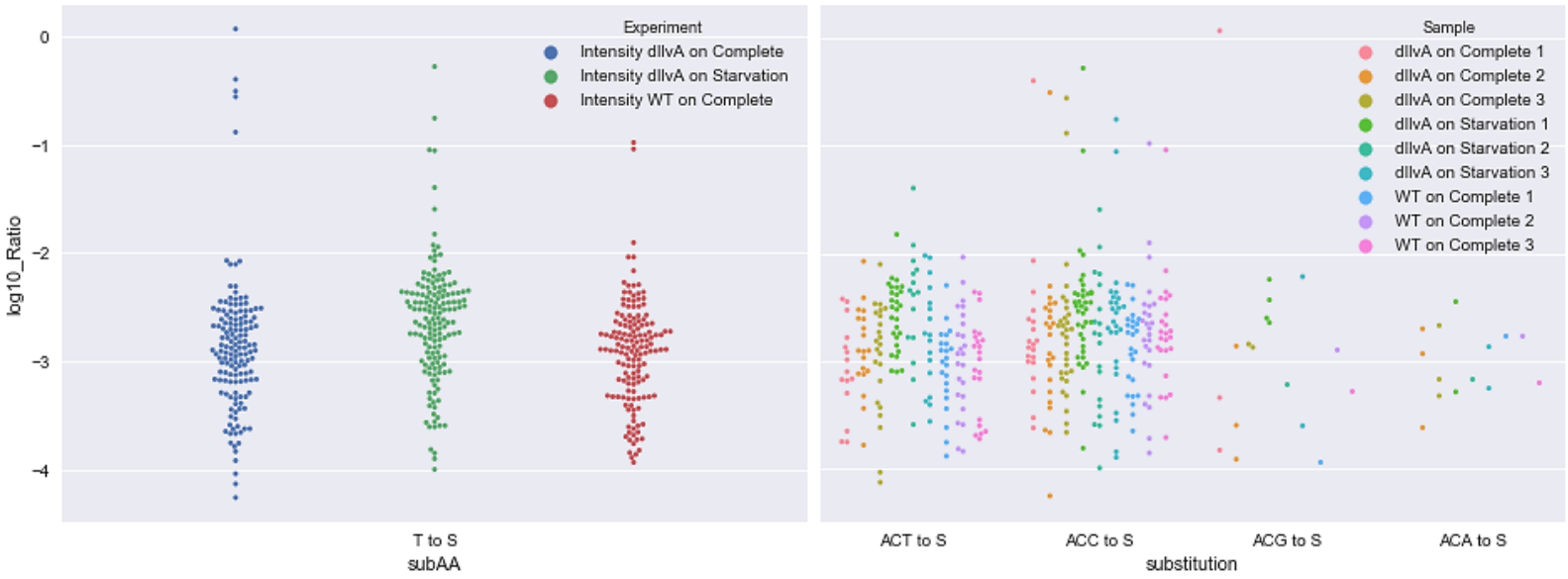


Figure 11 — S to X substitution is the most prevalent substitution in the isoleucine starvation assay dataset. This makes it a suitable study case for comparison between conditions and experiments. Here we see that ΔilvA in starvation media has a similar amount of S to X substitutions. Nevertheless, starved ΔilvA has the highest T to S error rate. The left figures contains a breakdown of all codon to AA substitutions while the right figure contains pooled T to S substitutions. Note that ACC and ACT are the most common used codon of Thr in the genome.

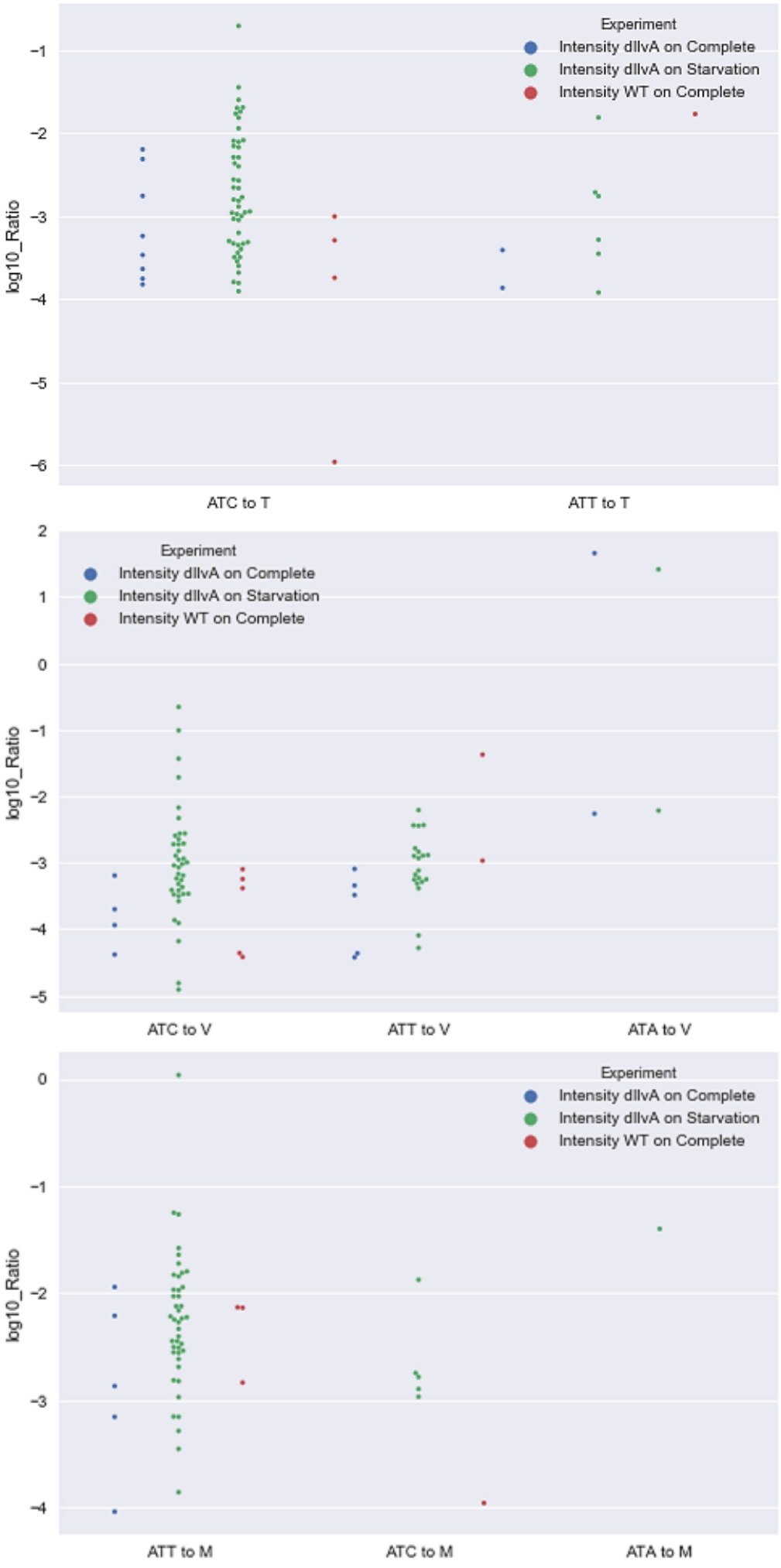


Figure 12 —ATC is mostly mistaken with Thr and Val codons while ATT is mostly mistaken with Meth.

Like in the proline experiment, T to S substitutions are enriched in all conditions and are apparently associated with the cells' struggle with excessive intracellular oxygen concentration (more on that in next results chapter). Putting I to X and P to X substitutions aside, the overall mistranslation patterns observed in isoleucine starvation assay are in agreement with previous patterns detected in proline starvation assay.

5.5 Further examination of T to S substitutions and there possible causes

We observed enrichment in T to S substitutions in all samples of proline and isoleucine experiments, whereas in the serine experiment ΔserA grown in starvation was the only condition that exhibited this result. In order to resolve this strange result, we re-examined all experiment setups to figure out whether there was an unintentional difference between the serine experiment and the two other experiments. Indeed, we realized that when conducting the serine calibration experiment we used a different type of flask to contain the bacteria broth. In the serine experiment we used 250ml Erlenmeyer flask which required repeated opening of the lid in order to measure the broth's OD, exposing the broth to possible contamination. To avoid this in future experiments, we started using a different Erlenmeyer flask which consists an attachment of a tube to its body, allowing to take OD measurements without opening the lid. This type of flask often has depressions at the bottom of it which are designed to stir the broth more rigorously and is named baffled flask. When the broth is stirred, baffled flask causes a substantial increase in the liquid-gas interface, allowing more air to diffuse into the liquid (Liu et al., 2016). Since more oxygen is available for the cells, this flask design is widely used in biology in order to allow cell grow more rapidly and reach a higher yield (Somerville & Proctor, 2013). This gave us the first clue as to why we do not observe T to S substitutions in WT samples derived from the serine (calibration) experiment dataset.

Furthermore, T to S substitutions are known to be associated with oxidative stress (Wu et al., 2014). It is evident that threonyl-tRNA synthetase (ThrRS), the enzyme which couples tRNAthr to thronine, mismatches tRNAthr to serine upon H_2O_2 treatment (Ling & Soll, 2010). This happens due to an oxidation reaction that occurs in the enzyme's editing site, preventing its self-proofreading when binding a Ser instead of a Thr and later charging it to the subsequent tRNAthr. Considering the possibility that using the baffled flask might have initiated oxidative stress in the proline and isoleucine experiments, we came up with an hypothesis; Bacteria grown in baffled flask is exposed to excessive amount of reactive oxygen species (ROS) which

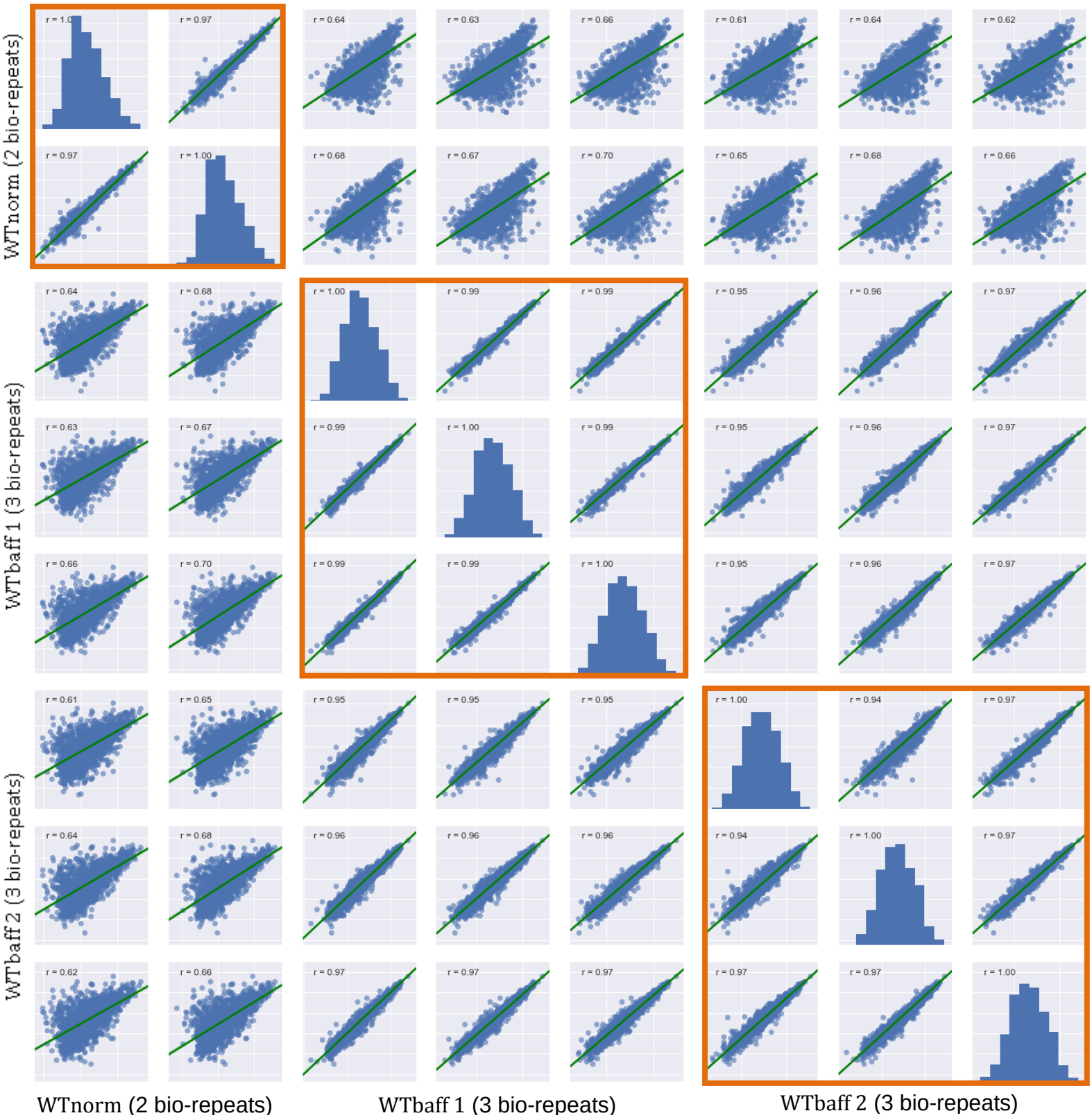


Figure 13 — WT grown in baffled flask correlates well with itself, even when grown in different days, but doesn't correlate well with WT grown in normal flask. Protein intensities comparison shows that WT grown in normal flask correlates well only with itself and seem to have depleted protein expression compared to the other WTs. Protein intensities were quantified and normalized using Max Quant program. Marked in orange squares are the biological repeats of the same experiment, well correlated with one another.

in turn causes the oxidation of ThrRS' editing cite (in the C182 position) and thus increasing the

mismatch between tRNAthr and serine.

To test this hypothesis, we compared our general proteome data which we've obtained from all different data sets. We compared three independent WT proteomes, from which one was grown in a normal flask (WTnorm) and the two others were grown in baffled flasks (WTbaff). We found that the protein expression level of the two proteomes which were derived from bacteria grown in baffled flasks were highly correlated with one another, while the WT grown in a normal flask didn't correlate with neither (Figure 13). We performed a two tailed t-test for WTnorm and WTbaff to test whether the two datasets are significantly different from one another. The two datasets were indeed different, with a p-value = 1.22e-19 (Figure 14). WTbaff also exhibits differential expression when compared WTnorm. According to Gene Ontology categories (Consortium, 2000), WTbaff is enriched with oxidative stress response proteins with FDR = 0.0114. Furthermore, seven additional genes which are known to be upregulated upon induction of oxidative stress response (Seo, Kim, Szubin, & Palsson, 2015) were upregulated in WTbaff. In contradiction to these results, starved ΔserA and its WT (both grown in normal flasks) have similar expression level distributions (p-value = 0.066) and don't show significant difference in expression of oxidative stress response proteins.

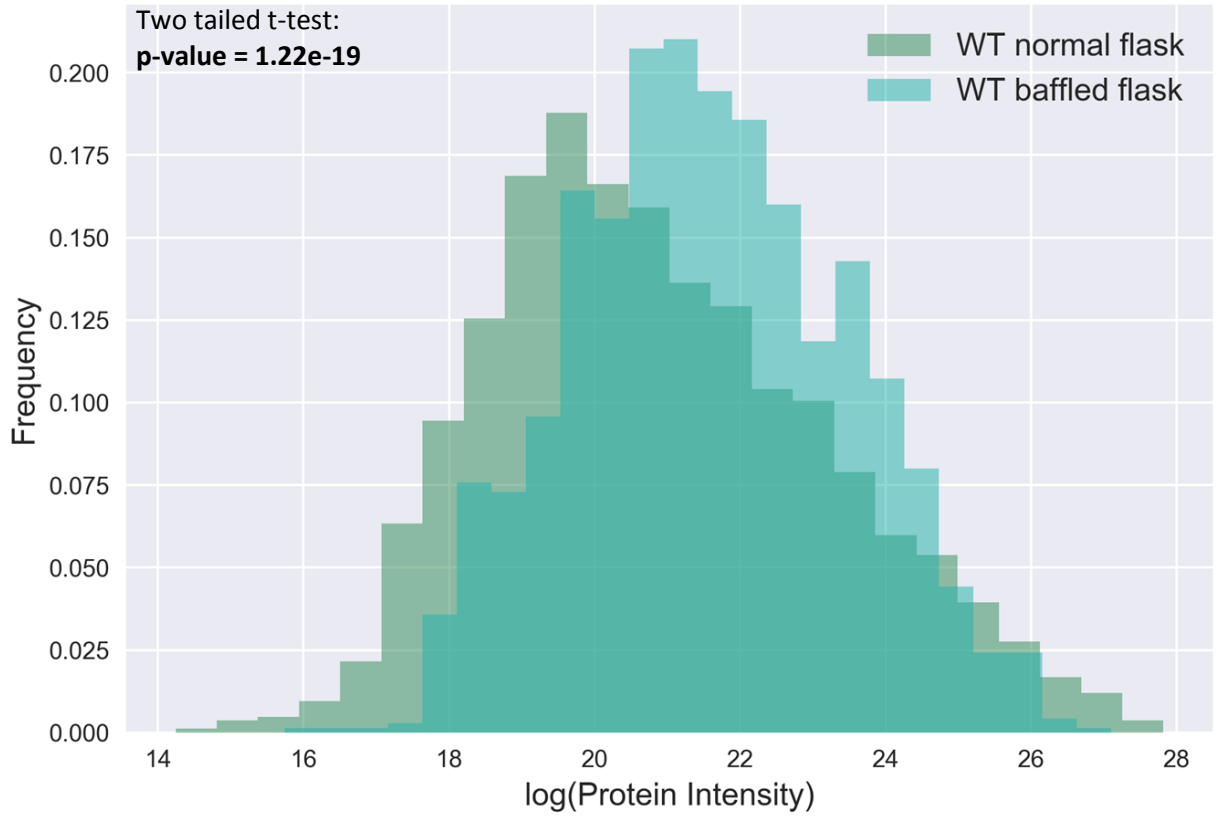


Figure 14 — WT grown in baffled flask has a different proteome distribution than WT grown in normal flask. Protein intensities in the two conditions differ from one another in a statistically significant manner

Next, we analyzed the very proteins which possessed T to S substitutions in order to test whether they share common biological or biochemical functions and properties. Proteins which were detected in all three biological repeats in a given condition and bare T to S substitutions, were tested for GO enrichments. We used all mistranslated proteins as a background for this analysis. Proteins with T to S substitutions which were found in mutants and WT grown in complete media were enriched with translation machinery proteins (p-value =). On the other hand, proteins with T to S substitutions which were found in mutants grown in starvation media were enriched for with both oxidative stress response proteins and translation machinery proteins.

6. Discussion

The objective if this study were to identify, characterize the translation errors that occurred in auxotroph bacteria cells which were starved for a single amino acid. The results for mistranslation detection and quantification support our hypothesis that starvation for a specific amino acid leads to a depleted internal abundance of that amino acid. This in turn, forces the translation machinery to lose its basal fidelity in general and even more so in the depleted AA's protein position. We show that the starved auxotrophs are more susceptible to mistranslation compared to WT and auxotroph grown in complete medium in all starvation assays. Nonetheless, not all experiments agree with the assumption that the starved auxotroph would be the most erroneous in the depleted AA position compared to WT and auxotroph grown in complete media; The proline auxotroph - ΔproA is more prone to make I to X substitutions than its counterpart grown in complete media, suggesting that there is a compensation mechanisms for proline depletion. It might be that the ΔproA cells can sense the relatively low proline abundance in its medium and act correspondingly to avoid excessive translation errors in its proteome. Unstraved Ile auxotroph also has the highest number of translation errors as well as error rate in most mistranslation events detected. The substitutions detected were of type P to A/S/L, where the first and second are of type near-cognate and the latter is of type non-cognate. We consider non-cognate mistakes to be results of faulty AA-tRNA charging by aaRS. Near-cognate substitutions are associated with either misidentification of the anticodon in the ribosome A site, or in some cases, misidentification by aaRS in its anticodon recognition site. This suggests that the starved Pro auxotroph might have been primed to

handle all types of mistranslation events and not necessarily the P to X type. Moreover, this result deviates from the notion that the overall prosperity of the starved auxotroph, judging from the growth curves, is in fact worse compared to the unstraved auxotroph. This suggests that the physiological condition of the bacteria is not necessarily indicative to its translation fidelity state.

Isoleucine auxotrophs on the other hand, demonstrate the expected results of elevated I to X frequency when grown in proline starvation media in comparison to auxotroph grown in complete media. WT grown on complete and the unstarved auxotroph have approximately the same amount of I to X substitutions. From this we can conclude that the Ile auxotrophs that were grown in Ile depleted media had difficulties incorporating Ile into elongating peptides. Isoleucine position were detected with four types of substitutions: I to M, N, V and T. All of these are of type near cognate, supporting our previous results which concluded that 88% of mistranslated positions detected could be explained by single codon-anticodon mismatch. When considering the translation error rate in each condition, it is apparent that starved Ile auxotroph is the most error-prone condition out of the three we have examined. Once again, we observe that when a certain condition is driven to make an exaggerated amount of translation error in specific position, Ile positions in this case, it is more prone to be generally inaccurate even though it should have all the needed components to make proper translation process. This could be explained by the genetic background of the auxotroph. The bacteria used for this work are mutant knockout strains for certain genes that are involved in the AA's biosynthetic pathway. We mustn't forget that the knocked out gene could interact with many components in the cell and might have biological roles which we are unaware of. Moreover, judging by biosynthetic pathways alone, it is plausible that the absence of the ilvA, serA or ProA genes cause a metabolic shift which is difficult to predict. Together with the specific AA shortage, the systems' complexity of the assay prevents us from concluding what are the possible outcomes of the overall cellular behavior when starving such auxotrophs.

When taking a closer look at I to X substitutions we come across some unexplained consequences in terms of codon-anticodon mismatch bias. For example, the Ile codon A<u>T</u>C is more prone to be mistranslated as G<u>T</u>C (Val), while A<u>T</u>T is only seldom misread as G<u>T</u>T (Val). As these two mistranslation types are identical in both nucleotide and nucleotide position (A is confused with G at the first codon position) and destination amino acid (Val), this bias cannot be explained by our previously described 4 by 4 matrices which aim to predict this type of biases. This bias also appears in V to T substitutions - where A<u>T</u>C (Ile) is commonly mistaken to be A<u>C</u>C (Thr), and yet A<u>T</u>T (Ile) to A<u>C</u>T (Thr) type of substitution is much rarer. Once more, we observe a case in which a certain near-cognate substitution is more prevalent to a certain codon-codon confusion rather than another, even though the problematic

position is of the same nucleotide and the same destination AA. Note that the Ile codons ATT and ATC share the same frequency in the genome and thus this bias cannot be explained by the fact that one of them is more abundant. This is an intriguing affect and might be related to the neighboring nucleotides of the problematic one. In both cases, the preferred mistranslation type (ATCile to GTCval and ATCile to ACCthr) has a cytosine in its third codon position where the unpreferred mistranslation type (ATTile to GTTval and ATTile to ACTthr) has a thiamine in its third position. As the preferred mistranslation types have three hydrogen bonds in the third anticodon position (C-G) rather than two (T-A), it might be more stable when binding to the codon and therefore harder to disassociate upon ribosome proofreading.

When we started investigating the proline and isoleucine assays' results, one of the first steps was to compare them to our calibration assay results. We immediately noticed that all of the different conditions in our Pro and Ile experiments bared a substantial amount of T to S substitutions, where in the serine calibration experiment only the starved Ser auxotroph demonstrated this effect while WT grown in complete media—didn't. This made us reconsider whether there was a significant difference between the experiment designs. Eventually we realized that the cell culture flasks used in the calibration experiment were not identical to the ones used in the Ile and Pro experiments. This caused a disproportionate amount of oxygen molecules to diffuse into the broth, leading the cells to deal with some potential oxidative stress. Through our proteome data, we showed that the WT cells grown on complete media that were derived from Ile and Pro experiments and grown in baffled flasks, were indeed over-expressing 14 different oxidative-stress related genes compared to WT grown in a normal flask. Although we've settled our oxidative stress hypothesis, this cannot explain why starved ΔserA which was grown in normal flask shows the T to S effect. Accordingly, we consider the prospect of T to S substitution as an outcome to various stresses, AA starvation including. As an evidence to this interpretation, previous mistranslation assays conducted in our lab which used antibiotics as the causative stress for translation errors—also exhibit T to S substitutions.

In conclusion, this study shows how translation errors are affected upon Ile, Pro and Ser starvation assays in a high throughput manner. We show that low translation fidelity can be induced by many environmental constraints and also support previous opinion that mistranslation can serve as a regulatory response to oxidative stress.

7. Literature

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., … Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. *Molecular Systems Biology*, *2*. https://doi.org/10.1038/msb4100050

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(D1), 36–42. https://doi.org/10.1093/nar/gks1195

Bertels, F., Merker, H., & Kost, C. (2012). Design and characterization of auxotrophy-based amino acid biosensors. *PLoS ONE*, *7*(7). https://doi.org/10.1371/journal.pone.0041349

Consortium, T. G. O. (2000). Gene ontologie: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556.Gene

Dougan, D. A., Mogk, A., & Bukau, B. (2002). Protein folding and degradation in bacteria:¶To degrade or not to degrade? That is the question. *Cellular and Molecular Life Sciences*, *59*(10), 1607–1616. https://doi.org/10.1007/PL00012487

Edelmann, P., & Gallant, J. (1977). Mistranslation in E. coli. *Cell*, *10*(1), 131–137. https://doi.org/10.1016/0092-8674(77)90147-7

Ling, J., & Soll, D. (2010). Severe oxidative stress induces protein mistranslation through impairment of an aminoacyl-tRNA synthetase editing site. *Proceedings of the National Academy of Sciences*, *107*(9), 4028–4033. https://doi.org/10.1073/pnas.1000315107

Liu, Y., Wang, Z. J., Zhang, J. W., Xia, J. ye, Chu, J., Zhang, S. L., & Zhuang, Y. P. (2016). Quantitative evaluation of the shear threshold on Carthamus tinctorius L. cell growth with computational fluid dynamics in shaken flask bioreactors. *Biochemical Engineering Journal*, *113*, 66–76. https://doi.org/10.1016/j.bej.2016.06.001

Mordret, E., Yehonadav, A., Barnabas, G. D., Cox, J., Dahan, O., Lindner, A. B., & Pilpel, Y. (2018). I ntroduction, 1–25.

Russell, J. B., & Cook, G. M. (1995). Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiological Reviews*, *59*(1), 48–62. https://doi.org/10.1.1.321.8181

Seo, S. W., Kim, D., Szubin, R., & Palsson, B. O. (2015). Genome-wide Reconstruction of OxyR and SoxRS Transcriptional Regulatory Networks under Oxidative Stress in Escherichia coli K-12 MG1655. *Cell Reports*, *12*(8), 1289–1299. https://doi.org/10.1016/j.celrep.2015.07.043

Somerville, G. A., & Proctor, R. A. (2013). Cultivation conditions and the diffusion of oxygen into culture

media: The rationale for the flask-to-medium ratio in microbiology. *BMC Microbiology*, *13*(1). https://doi.org/10.1186/1471-2180-13-9

Subramaniam, A. R., Pan, T., & Cluzel, P. (2013). Environmental perturbations lift the degeneracy of the genetic code to regulate protein levels in bacteria. *Proceedings of the National Academy of Sciences*, *110*(6), 2419–2424. https://doi.org/10.1073/pnas.1211077110

Wu, J., Fan, Y., & Ling, J. (2014). Mechanism of oxidant-induced mistranslation by threonyl-tRNA synthetase. *Nucleic Acids Research*, *42*(10), 6523–6531. https://doi.org/10.1093/nar/gku271