Bharati Vidyapeeth
(Deemed to be University)

TOPIC - Differential Translation Control in Caulobacter crescentus in all Cell Cycle stages

AN INTERNSHIP REPORT SUBMITTED TO,

BHARATI VIDYAPEETH (DEEMED TO BE UNIVERSITY),
RAJIV GANDHI INSTITUTE OF IT AND BIOTECHNOLOGY, PUNE

FOR PARTIAL FULFILMENT OF

THE AWARD OF A DEGREE OF MASTER OF SCIENCE

IN BIOINFORMATICS

SUBMITTED BY

Kumari Karuna

PRN: 2416120014
SEAT NO: 2424850207

GUIDED BY

Dr Ajeet Kumar Sharma

PLACE OF INTERNSHIP

IIT JAMMU
25 MAY – 15 JULY 2025

जैव-विज्ञान और जैव-अभियांत्रिकी विभाग
Department of Biosciences and Bioengineering
भारतीय प्रौद्योगिकी संस्थान जम्मू
Indian Institute of Technology, Jammu
जगती, एन एच **44**, नगरोटा, जम्मू- **181221**, जे & के, भारत |
Jagti, NH 44, Nagrota , Jammu - 181221, J&K, India.
W: www.iitjammu.ac.in

To Whom It May Concern,

This is to certify that **Ms. Kumari Karuna**, a student in the M.Sc. Bioinformatics program at Rajiv Gandhi Institute of IT & Biotechnology under Bharati Vidyapeeth (deemed to be University), Pune has successfully completed Summer Internship at IIT Jammu from May 26, 2025 to July 15, 2025.

During this period, the intern actively participated in various research activities, projects, and discussions, gaining valuable insights and practical experience in the field of Computational Biology and Bioinformatics.

We wish Ms. Kumari Karuna continued success in their future academic and professional endeavors.

Dr. Ajeet K. Sharma
Assistant Professor
Department of Physics,
Department of Biosciences and Bioengineering
Indian Institute of Technology Jammu

Date: July 8, 2025

Objective

The primary aim of this study was to explore the mechanisms of translational regulation in *Caulobacter crescentus* across its cell cycle. Specifically, the objectives were:

1. **To investigate the dynamics of translational control across six distinct cell cycle stages**
   – By integrating RNA-seq and Ribo-seq data to quantify gene-specific translational efficiency (TE), and determine how translation output varies independent of transcript levels.
2. **To identify and evaluate molecular factors that influence translation efficiency, and analyse their stage-wise patterns**
   – Including codon-specific translation time (dwell time), tRNA Adaptation Index (tAI), and mRNA secondary structure (predicted via minimum free energy, MFE), and to correlate these features with TE at each stage.
3. **To assess functional category-specific regulation of translation**
   – By grouping genes based on biological roles (e.g., motility, membrane proteins, cell wall synthesis) and analysing how TE and its influencing factors vary across these categories, highlighting selective translation of genes according to cellular needs.

Abstract

This internship project explored the dynamics of differential translation control in *Caulobacter crescentus* across six defined stages of its cell cycle. The central objective was to examine how translational regulation changes over time and how it relates to key factors such as tRNA abundance and translation efficiency (TE). To capture the dynamics of codon translation, two computational strategies—the Offset Method and Scikit-ribo—were employed to estimate codon dwell times at each stage, allowing for comparative analysis between the two approaches.

Further, correlations were examined between TE and several influencing parameters, including the average and geometric mean codon translation rates and the tRNA adaptation index (tAI). While most correlations proved inconclusive, a Wilcoxon test revealed a statistically significant relationship for the first 10 codons, suggesting a possible early-stage regulatory effect.

To understand these dynamics more deeply, TE and tAI correlations were analysed not only across the entire gene set but also in subsets, such as genes with stable RNA expression across stages and those inactive in specific phases. Functional classification allowed a focused investigation into categories like motility, membrane proteins, and cell wall synthesis, tracking how TE, tAI, and RNA-seq signals varied throughout the cell cycle.

Despite comprehensive statistical analysis, consistent trends remained elusive, emphasising the complex and multilayered nature of translational regulation in bacteria. Nevertheless, this study provides valuable groundwork for future experimental validation and deeper insights into post-transcriptional gene regulation mechanisms.

Introduction

Gene expression in bacteria is a tightly controlled process that ensures proteins are made at the right time and in the right amount. While a lot is already known about how genes are regulated at the transcriptional level, there's still much to learn about how translation—the process of turning mRNA into protein—is controlled, especially across different stages of the cell cycle.

This internship project focused on the bacterium *Caulobacter crescentus*, which is known for its well-defined and asymmetric cell cycle. Because of its clear staging and natural biological significance, it provides a great model to study how gene expression changes over time, not just at the mRNA level, but also during translation.

We aimed to explore how translation efficiency (TE) changes during the six stages of the cell cycle and how it's influenced by factors like tRNA availability, codon usage, and mRNA expression. To do this, we used two computational approaches—the Offset Method and Scikit-ribo—to calculate codon dwell time, which gives us an idea of how fast or slow a ribosome translates each codon.

We also looked at how well TE correlates with other features like the tRNA adaptation index (tAI) and average translation rates, and explored how these patterns differ across functional gene groups such as those involved in motility, membrane transport, and cell wall synthesis. Genes were further classified based on whether they were consistently expressed or only active at specific stages.

By the end of this study, we hoped to gain a clearer picture of the hidden layers of translational control and how they might affect bacterial growth and function. This project not only helped deepen our understanding of gene regulation but also gave valuable experience in using bioinformatics tools to analyse complex biological systems.

Methods and Materials

- Processing of NGS (Next-generation sequencing) data.

(a) RNA Sequencing (RNA-Seq)

RNA-Seq is a high-throughput sequencing technique used to measure gene expression by sequencing the complete set of RNA transcripts in a cell or tissue. The process begins with RNA isolation from a biological sample, followed by mRNA enrichment (typically using oligo(dT) beads or rRNA depletion). The isolated mRNA is reverse-transcribed into cDNA, which is then fragmented, adapter-ligated, and amplified via PCR to create a cDNA library. This library is sequenced using next-generation sequencing platforms, such as Illumina, producing millions of short reads stored in FASTQ format. These reads are aligned to a reference genome using tools such as STAR or HISAT2, and read counts are quantified to estimate gene expression levels. The data is analysed using visual tools, such as heatmaps, PCA, or volcano plots, to identify differentially expressed genes. While RNA-Seq reveals transcriptional activity, it does not directly reflect protein levels or translational regulation, making it essential to integrate it with methods like Ribo-seq for deeper insights.

(b) Ribosome Profiling (Ribo-Seq)

Ribo-Seq is a high-resolution technique that captures active translation by sequencing ribosome-protected fragments (RPFs)—short mRNA segments (28–30 nucleotides) shielded by ribosomes during translation. Unlike RNA-Seq, which captures all transcripts, Ribo-Seq reveals which mRNAs are actively being translated and the precise position of ribosomes on them.
The workflow begins with cell lysis and enzymatic digestion to remove unprotected mRNA, isolating only ribosome-bound fragments. These fragments are then processed into cDNA libraries through adapter ligation, reverse transcription, and amplification. Using high-throughput sequencing platforms like Illumina, millions of RPF reads are generated for downstream analysis.
This analysis allows researchers to determine ribosome density, translation start sites, elongation rates, and detect events such as ribosome pausing and upstream open reading frames (uORFs). When integrated with RNA-Seq data, Ribo-Seq enables the estimation of translation efficiency (TE) and provides deeper insights into post-transcriptional gene regulation.
Thus, Ribo-Seq is a powerful tool for understanding the translational landscape of cells and uncovering fine-scale regulatory mechanisms of gene expression.

(c) Average codon translation time

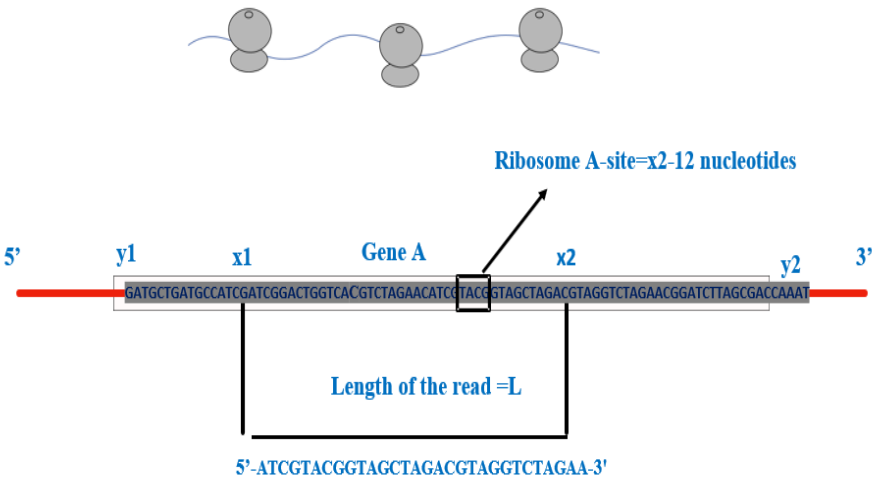
Average codon translation time reflects how long, on average, the ribosome takes to translate a specific codon across all genes. In this study, codon-level ribosome occupancy was first assigned using two approaches: the Fixed Offset Method, where a constant 12-nucleotide offset from the 3′ end of ribosome-protected fragments was applied to estimate A-site positions, and Scikit-ribo, a machine learning-based method that models A-site positions using read-specific features. For each method, ribosome density per codon was computed and averaged across all instances of a given codon to determine its average dwell time, providing insight into codon-specific translation dynamics and potential pausing sites.

$$\tau(i,j) = \frac{N(i,j)}{\sum_{j=2}^{L(i)} N(i,j)} \mathrm{T(i)}$$

Where,
N(i,j): the number of A sites assigned to the j^{th} codon position of the i^{th} gene,
L(i): the number of codons present in the i^{th} gene,
T(i): the time taken by the ribosome to translate the i^{th} gene,
⟨1⟩: the average time a ribosome spends translating a single codon (we are assuming it is uniform for all genes), T(i) = (L(i) - 1)·⟨1⟩

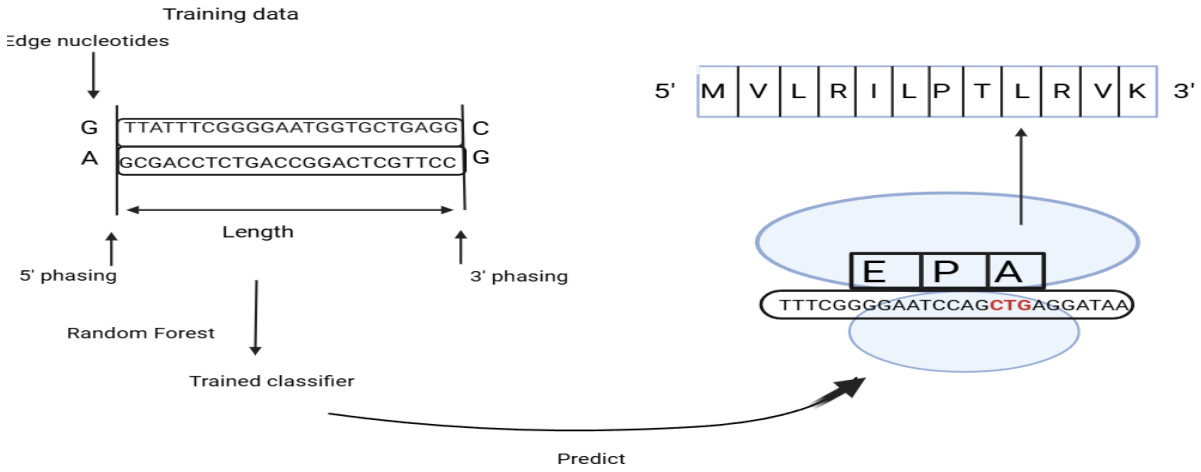
- **Fixed offset method to calculate the Ribosome A site**

To identify the ribosome A-site positions in ribosome profiling data, the Fixed Offset Method was applied. Since ribosomes protect ~28–30 nucleotides of mRNA during translation, the A-site position was estimated by subtracting 12 nucleotides from the 3′ end of each aligned ribosome-protected fragment (RPF).
A-site position=3′ end of RPF−12 nt\text{A-site position} = \text{3′ end of RPF} - 12 \text{ nt}A-site position=3′ end of RPF−12 nt
This approach allowed accurate assignment of codons being translated at a given time. The resulting A-site positions were then used to generate codon-level ribosome occupancy profiles for all genes, enabling downstream analyses such as codon translation rate calculation, ribosome pausing detection, and decoding efficiency estimation.



Fixed offset method

- **Machine learning based method (Sci-kit ribo)**
  Scikit-ribo is a machine learning-based tool that predicts the ribosome A-site using generalised linear models (GLMs) instead of relying on fixed offsets. It considers features such as read length, reading frame, and sequence context to improve A-site localisation accuracy. In this study, Scikit-ribo was used to estimate codon dwell times across six distinct cell cycle stages, enabling a detailed analysis of translation elongation rates and ribosome pausing. Additionally, TPM (Transcripts Per Million) values for genes were also calculated using Scikit-ribo, providing normalised expression levels that were further used to compute translation efficiency (TE).



Sci-kit ribo

(d) Translational Efficiency (TE) Calculation

Translational Efficiency (TE) was calculated by integrating Ribo-seq and RNA- seq data. TE represents the number of ribosomes translating an mRNA relative to its transcript abundance and was calculated using the formula:

$$TE = \frac{\text{TPM (Ribo-seq)}}{\text{TPM (RNA-seq)}}$$

(e) tRNA Adaptation Index Calculation

The tRNA Adaptation Index (tAI) quantifies how well the codons of a gene are adapted to the available tRNA pool, serving as a predictor of translation efficiency. In this study, tAI was calculated using tRNA TPM values with wobble base pairing taken into account.

Step 1: Codon Adaptiveness Score (s_i)

For each codon i, the adaptiveness score s_i is calculated by summing the TPM values of all tRNAs j that can pair with codon iii (including wobble):

$$s_i = \sum_{j \in tRNAs(i)} \text{TPM}_j \cdot \omega_{ij}$$

Where;

TPM_j is the abundance of tRNA j

w_{ij} is the **wobble pairing efficiency** between codon i and tRNA j (between 0 and 1)

Step 2: Relative Adaptiveness (w_i)

$$w_i = \frac{s_i}{\max(s)}$$

Step 3: Gene-Level tAI

$$tAI_{\text{gene}} = \left(\prod_{i=1}^{n} w_i \right)^{1/n}$$

This gives a geometric mean of codon adaptiveness, resulting in a single tAI value per gene.

(f) RNAFold

RNAfold is a computational tool from the ViennaRNA package that predicts the secondary structure of RNA sequences based on minimum free energy (MFE) folding. It takes an RNA sequence as input and calculates the most thermodynamically stable structure that the molecule is likely to form. In this study, RNAfold was used to predict the secondary structure of mRNAs, particularly near the start codon region, as secondary structure in this region can influence translation initiation and efficiency. The MFE values obtained were further used to analyse correlations with translation efficiency (TE) across genes and stages.

(g) Correlation Analyses

Various correlation analyses were performed to understand the factors influencing TE. These included:

- TE vs tAI (to assess codon adaptation's role)
- TE vs Average Translation Time per Codon

(h) Functional Annotation and Categorisation

Gene functions were extracted using the GFF annotation file of *Caulobacter crescentus* NA1000. These gene functions were grouped into broader functional categories (e.g., Translation, DNA Replication, Stress Response, Cell Cycle, Metabolism, etc.) to assess differential translational control at the functional level. This categorisation helped to determine whether certain biological processes were preferentially translated during specific cell cycle stages.

(i) Visualisation and Statistical Tools

- All statistical analyses and visualisations were performed using **R** and **Python**. Key libraries/tools included:
- **R**: ggplot2, dplyr, tidyr
- **Python**: pandas, seaborn, numpy
- **Plots used**: Heatmaps, Scatterplots with trendlines, Pie Chart (for function-level visualisation), Boxplots,

Workflow

This study explored how gene translation is regulated during the six cell cycle stages of *Caulobacter crescentus*. We started by processing RNA-seq and Ribo-seq data to measure mRNA levels and ribosome activity. Ribosome A-site positions—where the ribosome reads the mRNA—were identified using both a simple fixed-offset approach and a machine learning tool called Scikit-ribo. These positions helped calculate how long ribosomes spend on each codon, giving codon-specific translation times.

Next, we used tRNA expression data to calculate the tRNA Adaptation Index (tAI), which shows how well codons match the available tRNAs. We also predicted RNA secondary structures near the start of genes using RNAfold, since folding can affect how efficiently translation begins. Translational Efficiency (TE) for each gene was then calculated by comparing Ribo-seq and RNA-seq values. Finally, we examined how TE is influenced by tAI, codon translation speed, and RNA folding patterns through correlation analysis.

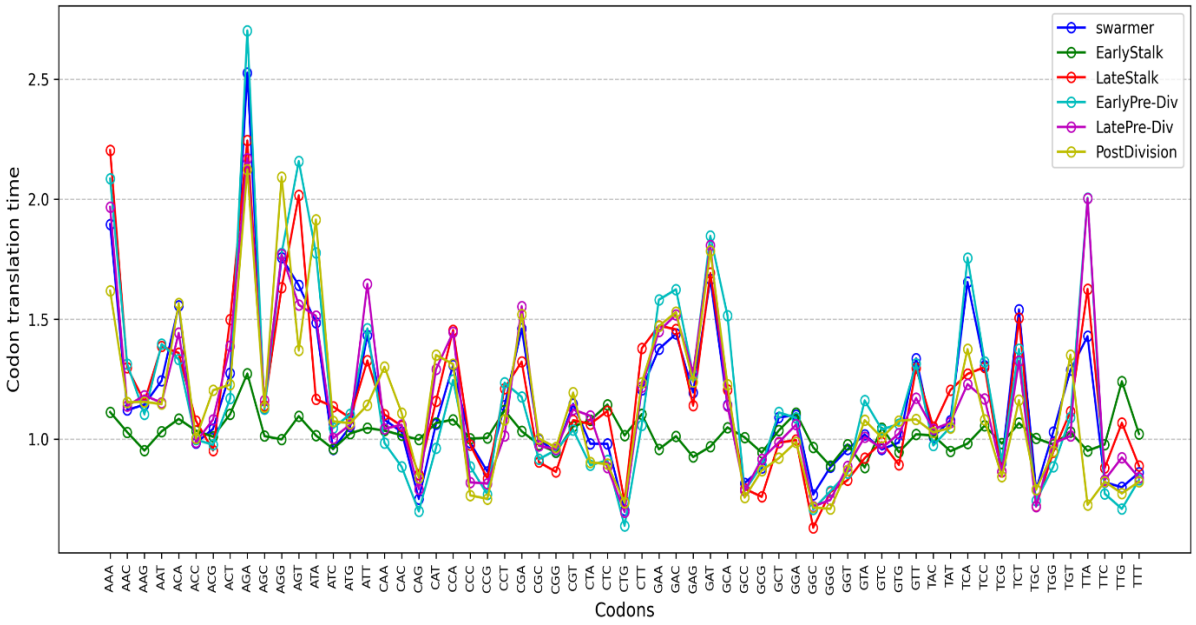
Summary

- RNA-seq & Ribo-seq: Processed to measure gene expression and translation activity.
- Ribosome A-site: Identified using fixed offset and Scikit-ribo.
- Codon Translation Time: Calculated from ribosome occupancy data.
- tAI Calculation: Based on tRNA expression and codon–tRNA matching.
- RNA Folding: Predicted minimum free energy (MFE) using RNAfold.
- TE Estimation: TE = Ribo-seq TPM / RNA-seq TPM.
- Correlation Analysis: Linked TE with tAI, codon speed, and RNA structure.

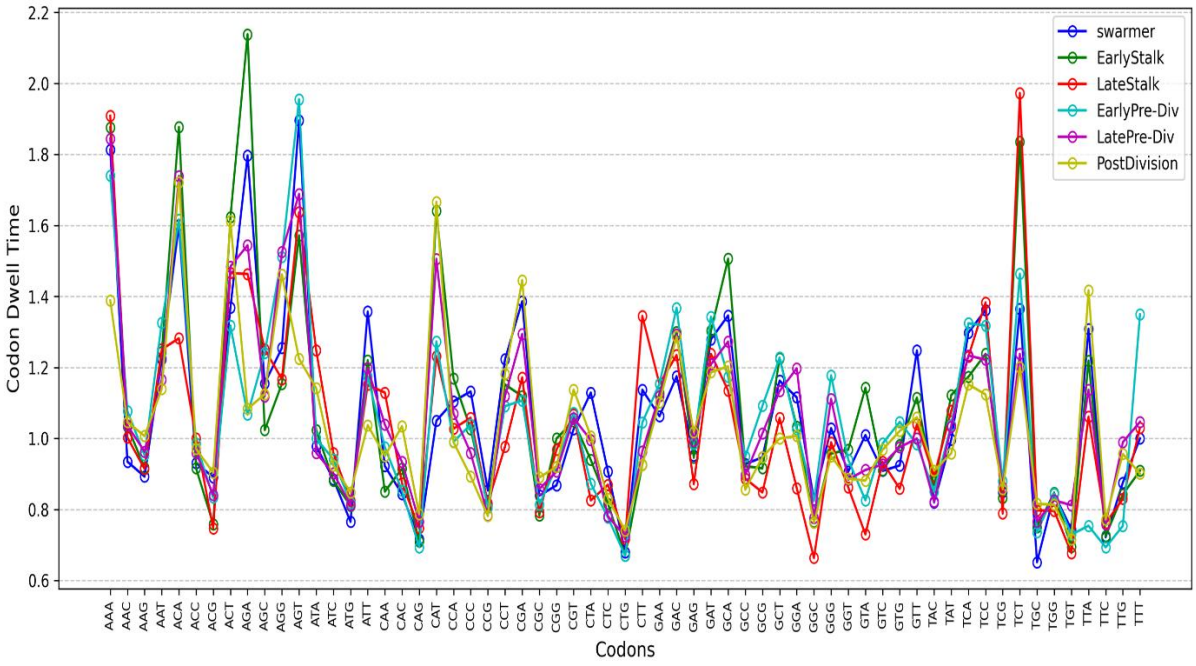
Result

- Codon Translation Time Varies Across Stages

  Codon translation (dwell) times were calculated for all six cell cycle stages using both the Fixed Offset Method and Scikit-ribo. The results revealed dynamic changes across stages, indicating that elongation rates are not constant during the cell cycle.



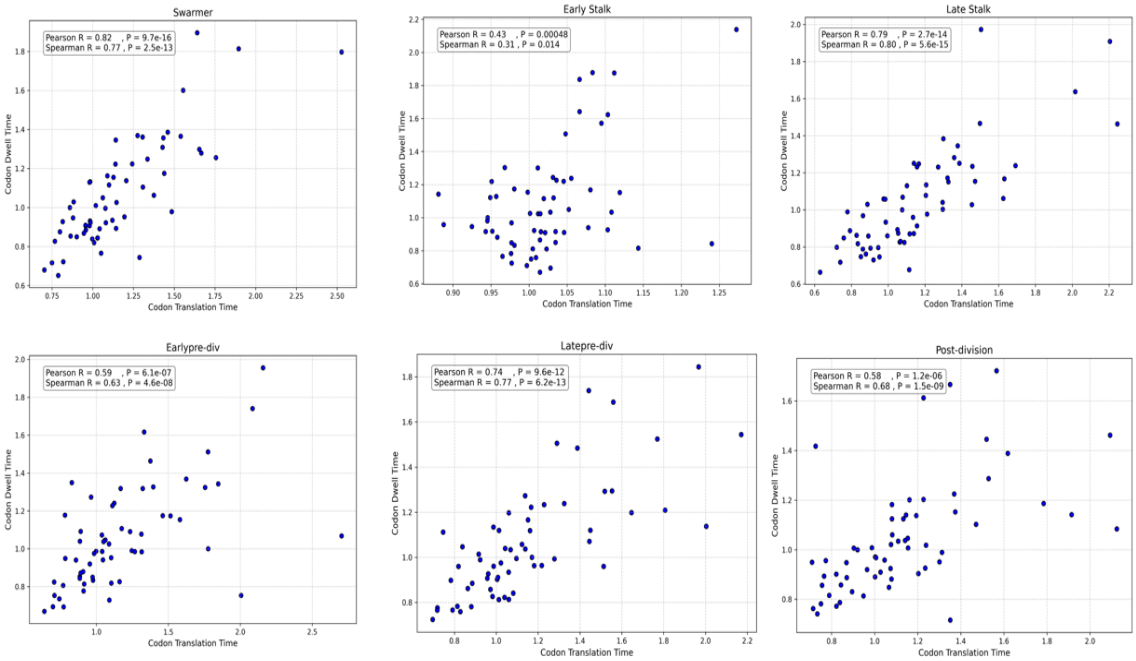
Using Offset Method



Using Sci-Kit Ribo

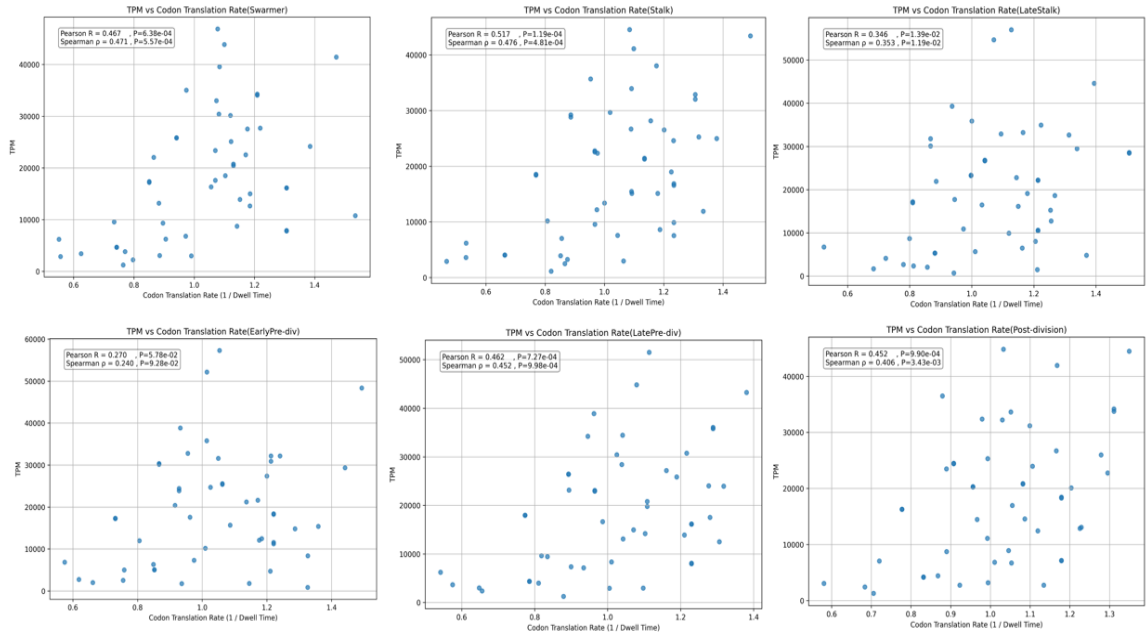
- Translation Time Correlation Between Methods

  A strong correlation was observed between codon translation times estimated using the offset method and Scikit-ribo, validating the robustness of both approaches.



- tRNA Abundance and Codon Translation

  Correlation between tRNA TPM values and codon translation rates showed that codons decoded by more abundant tRNAs tend to be translated faster, consistent with codon–tRNA adaptation.

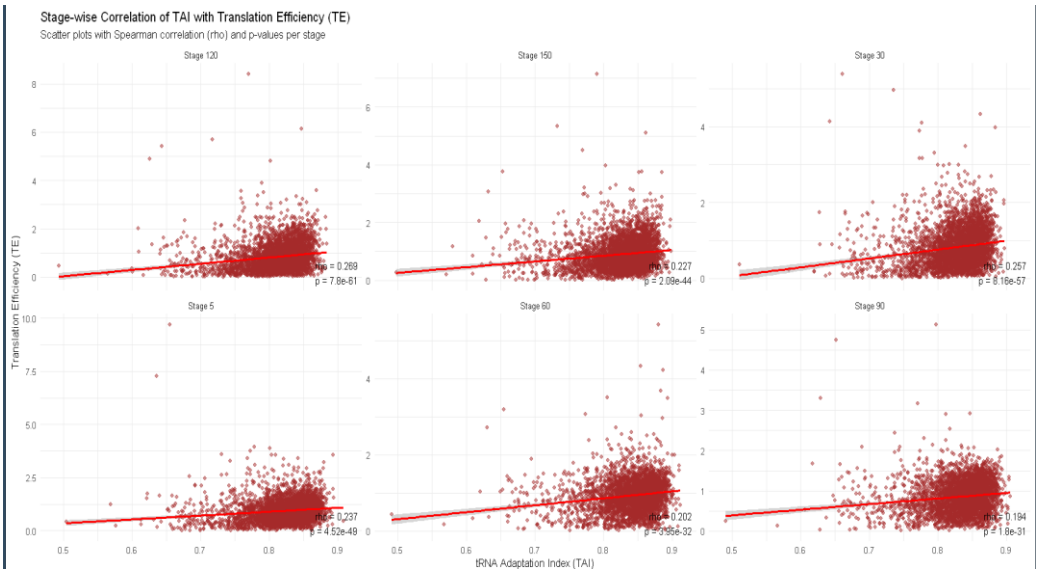


- TE vs Average Translation Rate (First 5–30 Codons)

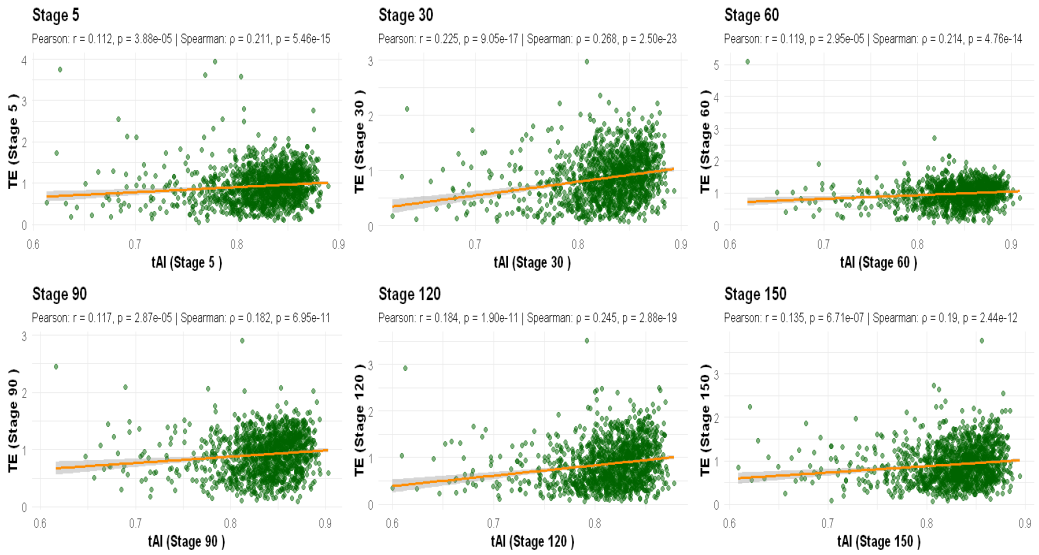
  Average translation rates and geometric means for the first 5, 10, 20, and 30 codons of each gene were correlated with TE. No significant relationship was found, except for a weak correlation in the 10-codon window, which showed significance in the Wilcoxon test.

- tAI and TE Correlation

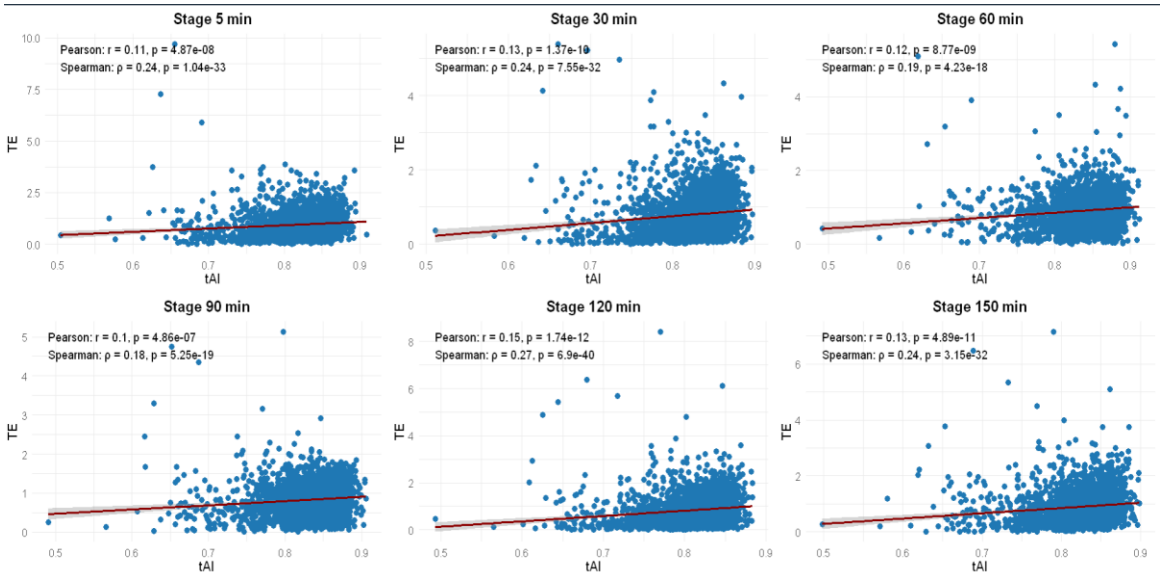
  tAI was calculated based on tRNA TPM values. A positive correlation was observed between tAI and TE across all genes. This correlation persisted even when genes were categorised based on RNA-seq stability across stages, but the correlation got weak when plotted for genes other than RNA-seq stable.



Plot for the whole gene set



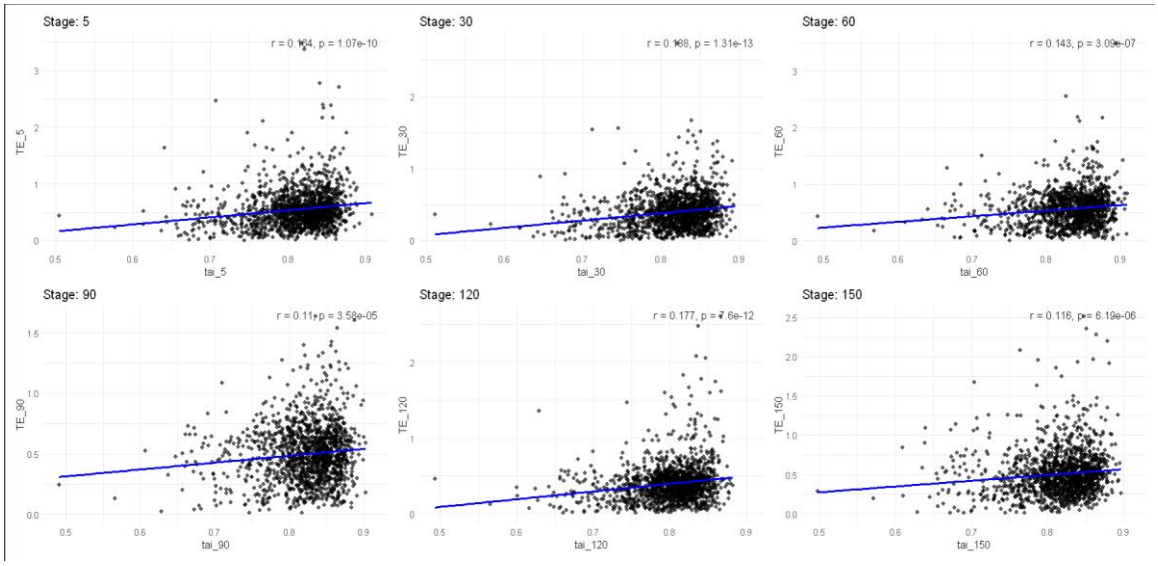
Plot for the genes that have RNA-seq stable throughout the stages



Plot for the genes other than RNA-seq stable genes

- Stage-Specific Gene Expression and TE–tAI Relationship

  Genes were grouped by expression pattern — those stably expressed in all stages and those turned off in certain stages. In the group that has genes which turn off in certain stages a consistent correlation between tAI and TE was observed, although the strength varied.

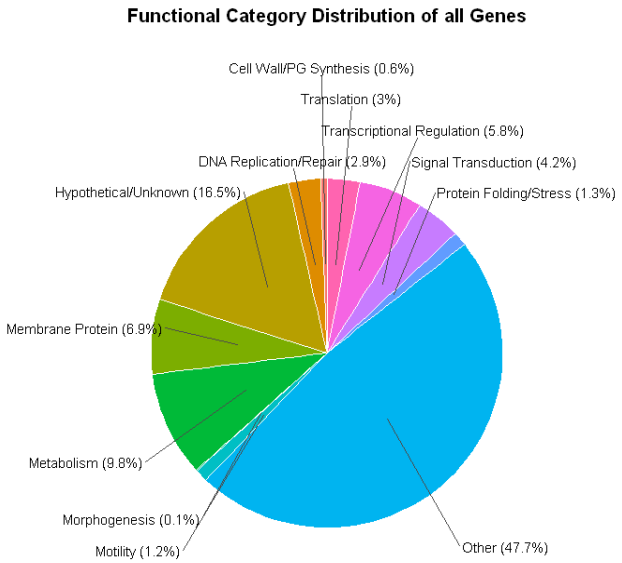


Plot for the genes that are off in some stages

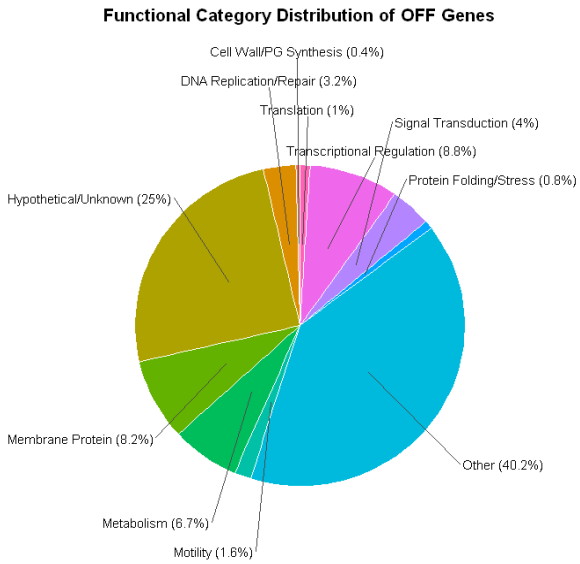
- Functional Category Analysis

  Three functional categories were taken from the set of regulatory genes— **Motility**, **Membrane Proteins**, and **Cell Wall Synthesis** — were examined in detail:

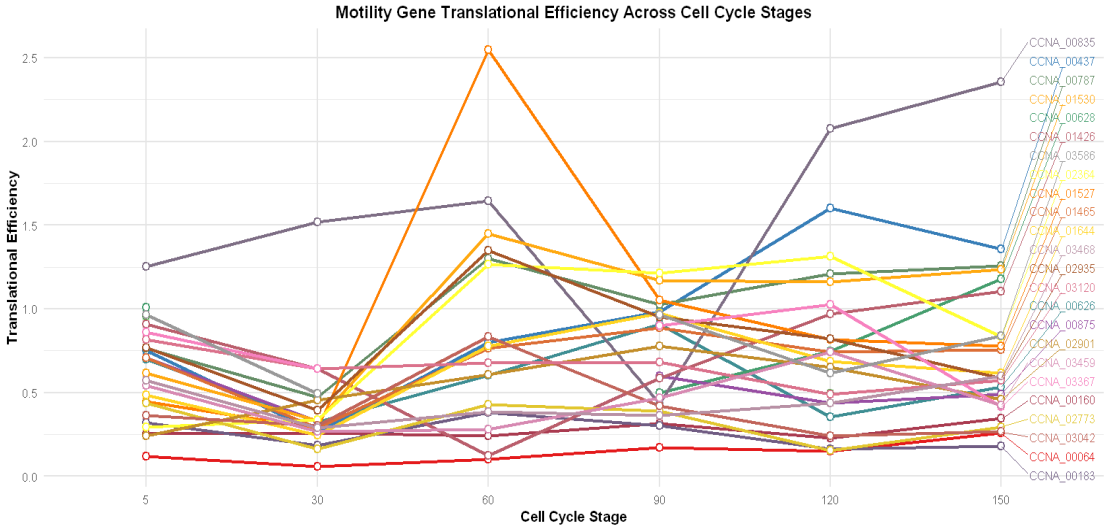
  - TE and tAI both fluctuated across stages.
  - A consistent correlation between TE and tAI was seen within each category.
  - No significant correlation was found between TE and average codon translation time in any category.



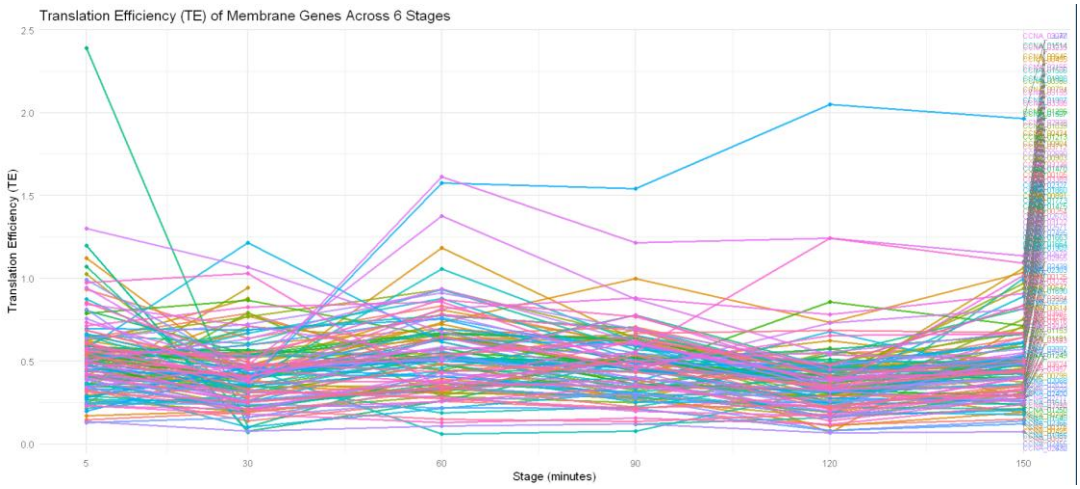
Functional Analysis of the Whole Gene Set



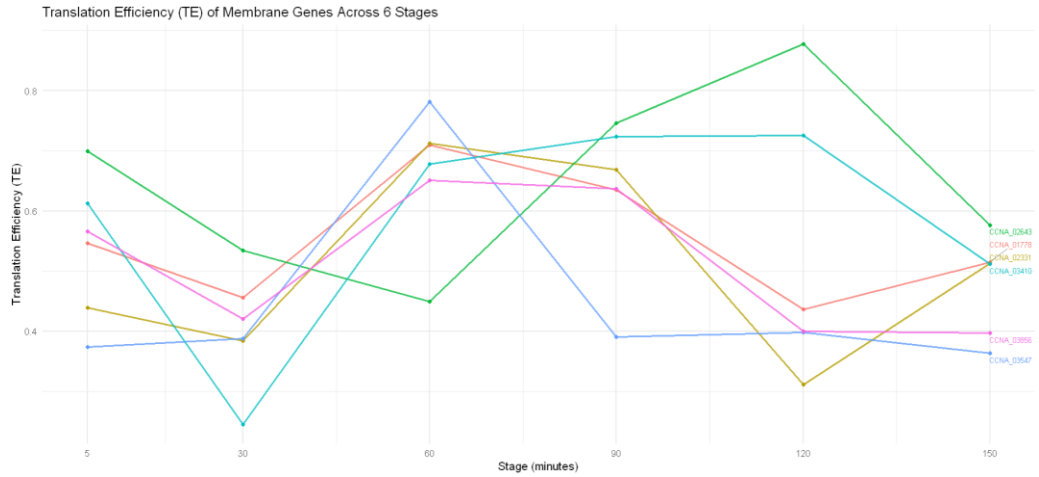
Functional Analysis of Regulatory Genes



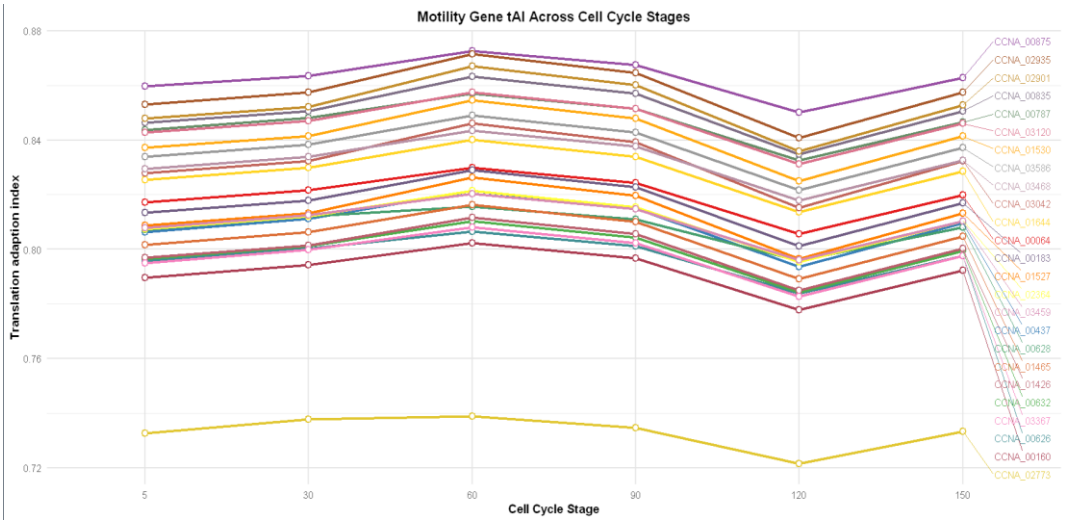
TE trend across stages for Motility Genes



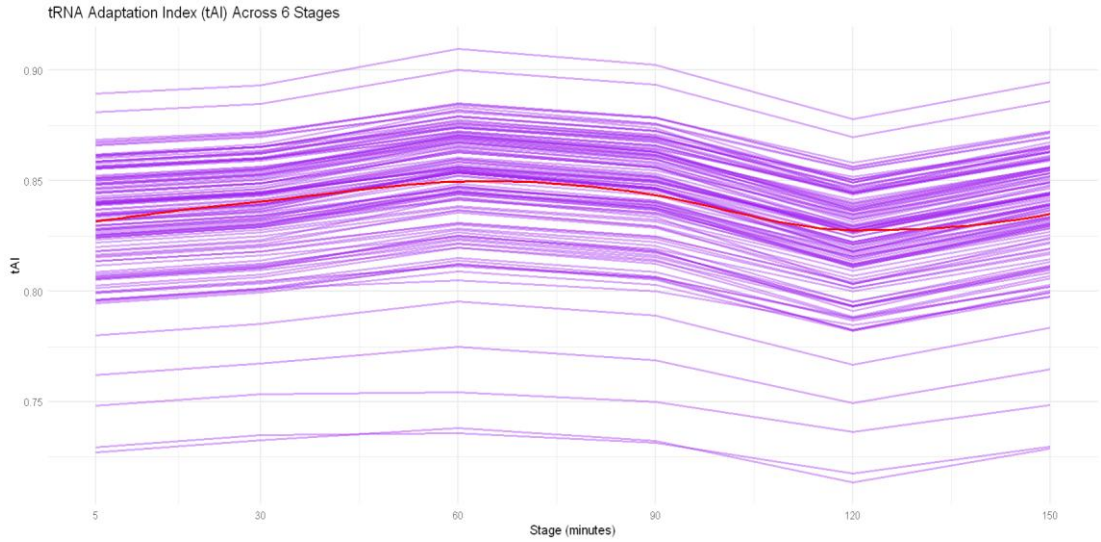
TE trend for the Membrane Protein Genes



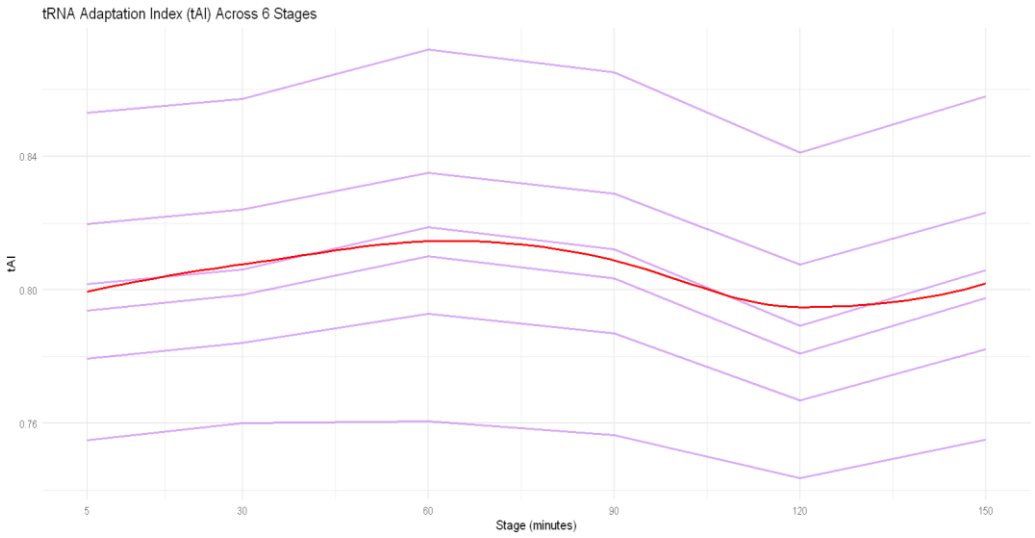
TE trend for the Cell Wall Synthesis Genes



tAI trend for Motility genes



tAI trend for Membrane Protein genes



tAI trend for Cell Wall Synthesis genes

Conclusion

This study provides insights into the differential translational control mechanisms operating across the six cell cycle stages of *Caulobacter crescentus*. By integrating RNA-seq, Ribo-seq, codon translation dynamics, tRNA availability, and RNA secondary structure analysis, we found that translational efficiency (TE) is not uniform but dynamically regulated.

A consistent positive correlation between the tRNA Adaptation Index (tAI) and TE suggests that codon usage and tRNA abundance significantly influence translation output. While codon-specific translation times varied across stages, they did not show a strong direct correlation with TE, highlighting the complex nature of elongation regulation. RNA secondary structure and other regulatory factors likely contribute to this complexity.

Functional category analysis further revealed that translational regulation is context-specific and may be coordinated with the cell's physiological needs at different stages. Overall, the findings underscore the importance of post-transcriptional mechanisms in shaping gene expression and contribute to a better understanding of bacterial cell cycle control at the translational level.

Acknowledgement

- I would like to express my sincere gratitude to Dr Ajeet Kumar Sharma, Department of Biosciences and Bioengineering, IIT Jammu, for providing me the opportunity to work on this research project and for their continuous guidance and support.
- I would like to thank my Phd Scholar Inayat Ullah Irshad, JRF Aneesa Mansoor, Senior Pinki Suthar and other lab members at IIT Jammu for their valuable inputs, resources, and constant encouragement.
- I also acknowledge the support of my home institute and faculty for allowing me to pursue this internship

References

- Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., & Weissman, J. S. (2009). *Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling*. Science, 324(5924), 218–223. https://doi.org/10.1126/science.1168978

- Wang, Z., Gerstein, M., & Snyder, M. (2009). *RNA-Seq: a revolutionary tool for transcriptomics*. Nature Reviews Genetics, 10(1), 57–63. https://doi.org/10.1038/nrg2484

- Li, G.-W., Oh, E., & Weissman, J. S. (2012). *The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria*. Nature, 484(7395), 538–541. https://doi.org/10.1038/nature10965

- Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). *ViennaRNA Package 2.0*. Algorithms for Molecular Biology, 6(1), 26. https://doi.org/10.1186/1748-7188-6-26

- Li, W., & Jiang, T. (2020). *Scikit-ribo: accurate estimation and robust modeling of translation dynamics at codon resolution*. Nucleic Acids Research, 48(4), e29. https://doi.org/10.1093/nar/gkz1217

- dos Reis, M., Savva, R., & Wernisch, L. (2004). *Solving the riddle of codon usage preferences: a test for translational selection*. Nucleic Acids Research, 32(17), 5036–5044. https://doi.org/10.1093/nar/gkh834

- Tuller, T., Waldman, Y. Y., Kupiec, M., & Ruppin, E. (2010). *Translation efficiency is determined by both codon bias and folding energy*. Proceedings of the National Academy of Sciences, 107(8), 3645–3650. https://doi.org/10.1073/pnas.0909910107

- Schrader, J. M., Zhou, B., Li, G.-W., Lasker, K., Childers, W. S., Williams, B., ... & Shapiro, L. (2014). *The coding and noncoding architecture of the Caulobacter crescentus genome*. PNAS, 111(52), E5533–E5541. https://doi.org/10.1073/pnas.1412975111