Identifying Gene Markers and Designing Transgenes in CHO Cells

Amrita Moyade
MSc Data Science
University of Bristol
Bristol, UK
tq24748@bristol.ac.uk

***Abstract*—Biotherapeutic proteins are a cornerstone of modern medicine, and Chinese Hamster Ovary (CHO) cells are the industry standard for their production. However, not all CHO clones perform equally well, and improving protein output often requires insights into which genetic or sequence-level factors contribute to higher expression. A key bottleneck in biopharmaceutical manufacturing is the lengthy and costly process of clone selection, where highly productive and stable CHO clones must be identified. This project addresses this challenge by mining RNA-seq expression data from CHO cell lines to achieve two objectives: (1) identify gene markers predictive of transgene (PRODUCT-TG) expression levels, and (2) characterize genes consistently expressed across conditions to inform future transgene design. We employed a combination of statistical correlation, Random Forest modeling, composite scoring, and deep sequence embedding analysis using DNABERT, alongside motif discovery with MEME and FIMO. Our findings reveal promising candidate markers and conserved sequence features that could streamline the cell line development pipeline and improve recombinant protein yields.**

The full codebase is available at: https://github.com/EMATM0050-2024/dsmp-2024-groupt27.git.

I. Introduction

Chinese Hamster Ovary (CHO) cells have become the predominant mammalian expression system in the biopharmaceutical industry for the production of monoclonal antibodies and recombinant therapeutic proteins. Their popularity stems from their ability to perform human-compatible post-translational modifications, scalability in suspension cultures, and robustness in industrial bioprocessing environments [1], [2].

Despite these advantages, a major bottleneck in bioprocess development remains the efficient selection of high-producing clones and the rational design of transgenes for optimal gene expression. Traditionally, clone selection relies on laborious screening of large numbers of cell populations, which is both time-consuming and resource-intensive [3], [4].

Recent advances in RNA sequencing (RNA-seq) technology have enabled a comprehensive investigation of gene expression profiles in CHO cell lines under various experimental conditions. This opens new avenues for data-driven approaches to identify gene markers associated with product expression and to uncover genomic features of constitutively expressed genes [5]. RNA-seq provides a system-wide snapshot of transcriptional activity. By analyzing transcript-level profiles, we can identify markers that correlate with high productivity and examine sequence-level drivers such as motifs, codon usage, or GC bias that may enhance gene expression.

In particular, the sequence properties of mRNA regions — such as the 5UTR, coding sequences (CDS), and 3UTR — are known to play critical roles in regulating transcript stability, translation initiation, elongation, and degradation [6]–[8]. Understanding these features provides valuable insight not only for predicting expression levels but also for designing synthetic transgenes with enhanced performance.

CHO cells are widely used in industrial biomanufacturing due to their regulatory acceptance and scalability. The global biologics market relies heavily on efficient clone development processes. Streamlining clone selection not only reduces time-to-market but also enables greater flexibility in product iteration and platform consistency. This project supports AstraZeneca's long-term strategy to reduce bioproduction bottlenecks and inform rational construct design.

Objectives: In this project, we analyze RNA-seq data from CHO cell lines to address two key objectives:

- Identify genes whose expression levels are highly correlated with PRODUCT-TG (transgene) expression, and propose them as potential markers for clone selection.
- Characterise constitutive genes consistently expressed across diverse experiments, with a focus on sequence features that can guide future transgene design.

To achieve this, we implemented batch-effect correction to address experimental variability, conducted correlation and machine learning-based ranking for marker discovery, and performed sequence feature analysis of the 5UTR, CDS, and 3UTR regions. Our results provide actionable insights for accelerating CHO cell line development and improving synthetic biology approaches for biopharmaceutical manufacturing.

II. Literature Review

RNA-seq is an established method for quantifying gene expression across diverse biological contexts [9]. Previous studies have demonstrated the potential of using molecular markers to characterize productivity in CHO cell lines [10]. Stability of expression is known to be influenced by untranslated regions [11], [12], codon usage [13], and overall transcript features. Machine learning models, including Random Forests, have been successfully applied to uncover non-linear relationships between gene expression and phenotypic traits [14]. Language

models like DNABERT [15] have recently emerged as powerful tools for capturing sequence-level information in genomics tasks.

Conserved sequence motifs are known to regulate transcript stability, translation efficiency, and subcellular localization of mRNAs in both prokaryotic and eukaryotic systems [6]–[8]. Tools like MEME and FIMO have been widely used to identify such motifs in genome-wide studies, offering insights into post-transcriptional regulation [16]. In particular, motifs within the 5'UTR have been linked to cap-independent translation and ribosome binding efficiency, while 3'UTR motifs can influence mRNA half-life through miRNA or RBP (RNA-binding protein) interactions.

Conserved sequence motifs such as TTCCTG and TCTCCT have been implicated in regulating transcript stability and translation efficiency. These short, GC-rich patterns are often targets for RNA-binding proteins and may modulate transcript half-life or ribosome accessibility. Identifying such motifs in constitutive genes provides a blueprint for engineering synthetic UTRs with improved regulatory behavior.

III. Data Description / Preparation

The dataset used in this study consisted of RNA-seq expression profiles and genomic sequence annotations derived from publicly available CHO cell line experiments. All data were processed using a standardized bioinformatics pipeline and included the following components:

A. Expression Matrix

The primary dataset, `expression_counts.txt`, contained gene-level expression values. The matrix included samples as columns and genes as rows, along with metadata such as transcript IDs, gene symbols, and peptide IDs.

Prior to analysis, expression values were log-transformed as $\log_2(\text{expression}+1)$ to reduce skewness and stabilize variance. Genes with near-zero expression across most samples were removed. PRODUCT-TG (product transgene) expression was embedded as a final row for downstream ranking analyses.

B. Metadata

The `MANIFEST.txt` file provided sample-level metadata including experiment IDs (used as batch labels), cell line identity, and laboratory source. These annotations were essential for batch-effect correction and grouping in statistical visualizations.

C. FASTA Sequence Files

The genomic regions associated with each transcript were extracted from the following Ensembl-derived FASTA files:

- `5UTR_sequences.fasta`: 5'UTR regions
- `CDS_sequences.fasta`: Coding sequences (CDS)
- `3UTR_sequences.fasta`: 3'UTR regions

FASTA files were parsed using BioPython. Sequence-level features such as GC content and length were computed, and codon usage was extracted for exploratory purposes in the CDS regions.

D. Data Cleaning and Integration

The following preprocessing steps were applied:

- **Missing Value Checks:** Ensured FASTA files and metadata contained no null or undefined entries.
- **Index Matching:** Expression and metadata were aligned using sample names; sequence data were joined using Ensembl transcript IDs.
- **Batch Label Encoding:** Experiment IDs were mapped to categorical batch labels for ComBat correction and visualization (e.g., PCA).
- **Sequence Statistics:** Length distributions and higher-order moments (e.g., skewness, kurtosis) were calculated per region.

This cleaned, normalized, and integrated dataset served as the foundation for subsequent statistical modeling and sequence-level analyses.

Dataset Characteristic	Value
Total Genes (raw)	36,000+
Genes after filtering	25,594
Samples (experiments)	108
Unique Batches (labs/projects)	12
FASTA Sequences Processed	3 regions (5'UTR, CDS, 3'UTR)

TABLE I: Summary of dataset characteristics

IV. Methodology

Our methodology follows two primary pipelines aligned with project objectives: gene marker identification and constitutive gene characterization.

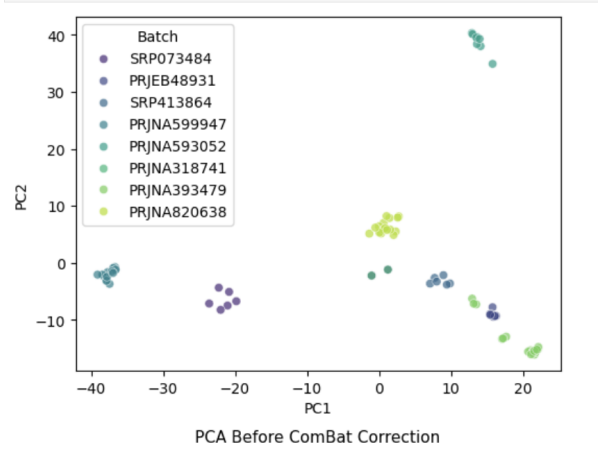
A. Data Preprocessing and Batch Effect Correction

The gene expression data was derived from RNA-seq experiments conducted across multiple laboratories. To correct for batch effect, the ComBat algorithm from the `sva` package was applied, which is widely used in high-throughput genomics studies [17].

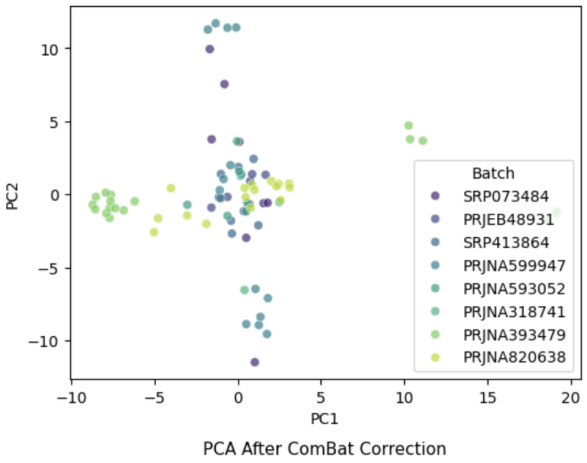
Prior to correction, expression values were log-transformed as $\log_2(\text{expression} + 1)$ to reduce right skew and stabilize variance. Genes with zero or near-zero expression across most samples were filtered out to remove noise.

To evaluate batch effects, we performed:

- **Principal Component Analysis (PCA)** was used to visualize batch effects in lower-dimensional space. By projecting high-dimensional expression data onto the first two principal components, we assessed whether samples clustered more by batch than by biological variation.



(a) Before ComBat Correction



(b) After ComBat Correction

Fig. 1: PCA Before and After Batch Correction

- **Hierarchical clustering** using Ward's linkage method was performed to evaluate global similarity patterns across samples. This allowed us to observe whether batches formed distinct subgroups based on expression profiles.
- **Levene's Test** was used to assess homogeneity of variance across batches, as it is robust to non-normal distributions. **Kruskal-Wallis Test**, a non-parametric alternative to ANOVA, was applied to detect median shifts between batches. Together, these tests helped validate the need for batch correction and evaluate the effectiveness of ComBat.
- **Mutual Information (MI)** analysis was employed to quantify the dependency between batch labels and gene expression profiles. A high MI value before correction and a lower value after would indicate successful reduction of batch-driven signals.

B. Marker Gene Identification and Ranking Strategies

To identify genes associated with PRODUCT-TG (transgene) expression, we began by computing **Spearman correlation coefficients** and corresponding p-values between gene expression and PRODUCT-TG levels. This provided an initial view of monotonic associations and informed later ranking strategies, although it was not used in isolation as a standalone ranking method.

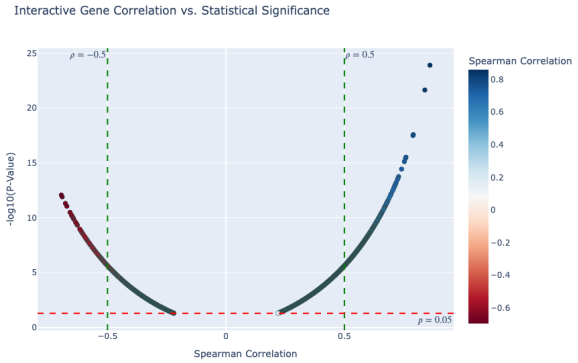


Fig. 2: Spearman correlation between gene and PRODUCT-TG expression

We then implemented four complementary ranking strategies, each capturing different aspects of gene–productivity relationships:

- **Ranking 1: Random Forest Regressor Importance.** A Random Forest Regressor was trained using gene expression data as input and PRODUCT-TG expression as the continuous target. Genes were ranked based on their model-derived feature importance scores.
- **Ranking 2: Regressor + Spearman Composite Score.** We computed Spearman correlation coefficients and p-values between each gene's expression and PRODUCT-TG levels. A composite score was calculated by combining: (i) the absolute Spearman correlation, (ii) statistical significance ($-\log_{10}(p)$), and (iii) regressor feature importance. All components were normalized using Min-Max scaling and weighted (0.4, 0.3, 0.1 respectively).
- **Ranking 3: Classifier Importance.** PRODUCT-TG values were binned into four equal-width quantiles ("Low", "Mid-Low", "Mid-High", "High") to reflect threshold biological behaviors and to capture the non-linear relationship of genes with Product-TG. A Random Forest Classifier was trained to classify samples into these productivity bands, and genes were ranked based on classification feature importance.
- **Ranking 4: Full Composite Score.** To account for both linear and non-linear gene-productivity associations, we computed a final composite score incorporating: (i) Random Forest Regressor importance, (ii) Spearman correlation, (iii) $-\log_{10}(p)$ significance, and (iv) Random Forest Classifier importance. Weights of (0.35, 0.30, 0.10,

0.25) were used after MinMax normalization of each metric.

This multi-view ranking approach ensures robust identification of marker genes that correlate linearly, non-linearly, or threshold-dependently with PRODUCT-TG expression. For results and overlaps across methods, see Table III and Table IV.

C. Constitutive Gene Characterization

Identification: Constitutive genes were defined based on consistent expression across all experimental conditions. We selected genes with:

- Expression > 1 across all samples
- Coefficient of Variation (CV) $< 10\%$

This yielded 25,594 genes that were then subjected to further sequence-level characterization.

Sequence Feature Analysis: For each constitutive gene, we extracted the **5UTR**, **CDS**, and **3UTR** sequences from ENSEMBL-derived FASTA files.

- **Length Distribution:** We computed sequence lengths for each region and plotted histograms. Statistical descriptors such as skewness, kurtosis, and Q–Q plots were used to assess distribution shape and normality.
- To statistically evaluate the distribution of sequence lengths, we conducted Anderson–Darling and D'Agostino–Pearson tests on all three regions (5'UTR, CDS, and 3'UTR). Results, summarized in Table II, indicate that none of the three gene regions follow a normal distribution (p-value < 0.001), justifying the use of non-parametric techniques in subsequent analysis.

Region	Anderson–Darling Statistic	D'Agostino–Pearson p-value
5' UTR	8158.17	0.00000
CDS	1879.68	0.00000
3' UTR	8570.71	0.00000

TABLE II: Normality test results for sequence lengths of constitutive gene regions. All regions deviate significantly from normal distribution.

- **GC Content:** Average GC content was calculated as 63.56% for 5UTR, 51.69% for CDS, and 44.29% for 3UTR. These values are consistent with literature indicating that elevated GC content contributes to transcript stability and translational efficiency.

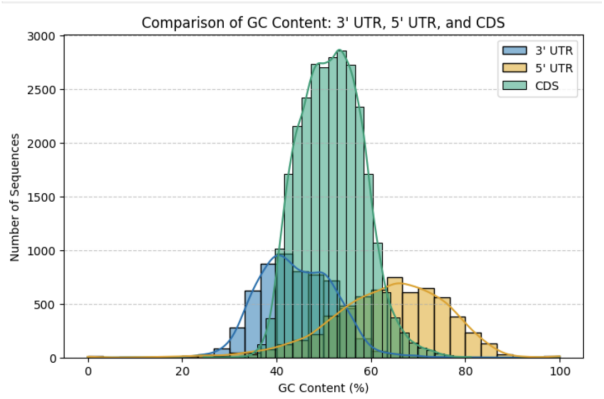


Fig. 3: GC content across 5UTR, CDS, and 3UTR regions of constitutive genes.

- **DNABERT Embeddings:** DNABERT was used to generate contextual embeddings for each region. These high-dimensional vectors allowed us to evaluate clustering patterns among constitutive genes, revealing latent sequence-level similarities.
- **Motif Discovery:** Using the MEME Suite, we discovered conserved sequence motifs across all regions. FIMO scanning showed that 91.3% of constitutive genes contained at least one of these motifs, suggesting functional regulatory roles.

We also performed distributional analysis of region-wise sequence lengths and GC content. The Anderson–Darling and D'Agostino–Pearson tests were used to assess normality of the length distributions. The 5'UTRs showed right-skewed GC content (60–70%), while CDS regions had tighter distributions around 55%, consistent with known compositional constraints in stable transcripts.

These findings indicate that constitutive genes share compositional and regulatory features that may underlie their stable expression profiles in CHO cells.

V. Results and Discussions

A. Batch Effect Correction Evaluation

To confirm the effectiveness of batch correction, we applied statistical tests and information-theoretic analysis:

- **Levene's Test**: Before correction, significant variance heterogeneity across batches was observed (Statistic = 259.43, $p < 0.0001$). After applying ComBat, the test statistic dropped to 4.14 ($p = 0.0001$), confirming improved homogeneity.
- **Mutual Information (MI)**: The dependency between batch labels and expression data was reduced from [1.564, 1.4548] to [0.6706, 0.3353], indicating that batch signal was substantially removed.

B. Gene Marker Ranking

Spearman correlation analysis identified 16,245 genes with significant associations to PRODUCT-TG levels (p-value $<$ 0.05). Random Forest Regressor modeling further highlighted

genes with high feature importance, many of which overlapped with the correlation-based findings.

To complement the regression and correlation approaches, a Random Forest Classifier was trained using discretized PRODUCT-TG expression levels. Although no formal classification accuracy was assessed, feature importances from the RFC provided an additional perspective and were integrated into the composite scoring framework.

The composite score combined scaled Spearman correlation coefficients, Random Forest Regressor feature importances, Random Forest Classifier feature importances, and $-\log_{10}$(p-value) significance transformations. Only statistically significant genes were included for final ranking.

Ensembl Transcript ID	RF-R	S+RF-R	RF-C	Final Composite
ENSCGRT00001019624	✓	✓	✓	✓
ENSCGRT00001024323	✓		✓	✓
ENSCGRT00001017013	✓	✓		✓
ENSCGRT00001005223	✓		✓	✓
ENSCGRT00001023166		✓	✓	✓
ENSCGRT00001031592	✓	✓		✓
ENSCGRT00001022794		✓	✓	✓
ENSCGRT00001029665		✓	✓	✓
ENSCGRT00001028824	✓			✓
ENSCGRT00001016259		✓		✓

TABLE III: Genes overlapping across multiple ranking strategies (Top-100 lists). RF-R = Random Forest Regressor Importance, S+RF-R = Spearman + Random Forest Regressor Composite, RF-C = Random Forest Classifier Importance.

Interpretation: Table III shows that several high-ranking genes, such as ENSCGRT00001019624 and ENSCGRT00001024323, appeared across multiple ranking methods. This overlap reinforces their relevance and suggests these genes exhibit both linear and non-linear relationships with PRODUCT-TG expression.

Ensembl Transcript ID	Total Hits
ENSCGRT00001019624	4
ENSCGRT00001024323	3
ENSCGRT00001017013	3
ENSCGRT00001005223	3
ENSCGRT00001023166	3
ENSCGRT00001031592	3
ENSCGRT00001022794	3
ENSCGRT00001029665	3
ENSCGRT00001028824	2
ENSCGRT00001016259	2

TABLE IV: Total number of ranking methods where each gene appeared in Top-100.

Interpretation: Table IV quantifies the number of ranking strategies in which each gene appeared among the top 100 candidates. The gene ENSCGRT00001019624 appeared in all four strategies, suggesting it is a robust marker gene candidate supported by multiple analytical perspectives.

To highlight the most promising candidates, Table V presents the top 5 genes identified through composite ranking. These genes include Glul, Actb, GAPDH, RPS13, and NSF — all of which are known to play important roles in cellular metabolism, structural integrity, translation, or vesicle trafficking, aligning well with biological expectations in high-producing CHO cell lines.

Rank	Transcript ID	Gene Symbol	Biological Relevance
1	ENSCGRT00001019624	Glul	Catalyzes glutamine synthesis; essential for CHO growth and recombinant protein production [18].
2	ENSCGRT00001023166	Actb	Structural cytoskeletal protein; maintains cell integrity critical for secretion [19].
3	ENSCGRT00001022794	GAPDH	Glycolytic enzyme; ensures energy supply for protein synthesis [20].
4	ENSCGRT00001024323	RPS13	Ribosomal protein; supports efficient translation and yield [21].
5	ENSCGRT00001031592	NSF	Mediates vesicle fusion; crucial for recombinant protein secretion [22].

TABLE V: Top 5 Genes Identified Through Composite Ranking

Interpretation: Table V highlights the top five candidate marker genes based on the final composite ranking. Each of these genes has a known role in protein production, secretion, or metabolic support, making them biologically relevant to high-yield CHO cell phenotypes.

C. Constitutive Gene Sequence Features

Constitutive genes were identified based on high mean expression and low standard deviation across all samples, reflecting stable transcription across diverse CHO cell line conditions. To investigate potential mechanisms behind their consistent expression, we analyzed several sequence-level features, including GC content, transformer embeddings, and motif enrichment.

GC Content Analysis: We computed the GC content of three gene regions—5'UTR, coding sequences (CDS), and 3'UTR. The average GC content was 63.56% for 5'UTR, 51.69% for CDS, and 44.29% for 3'UTR. The elevated GC content in untranslated regions supports literature findings that higher GC density can enhance transcript stability and secondary structure formation, contributing to sustained gene expression.

Transformer-based Embedding Patterns: We extracted DNABERT embeddings from each region of constitutive genes

and projected them into a lower-dimensional space. The embeddings exhibited clustering patterns, suggesting the presence of conserved regulatory elements or sequence features that are learnable by attention-based models.

D. Motif Discovery and Expression Impact

MEME Suite was used to identify statistically enriched motifs from each region, followed by FIMO scanning to map their occurrences. We found that 91.3% of constitutive genes contained at least one enriched motif, reinforcing their regulatory relevance.

Motif enrichment analysis revealed two complementary sets of patterns. The most frequently occurring motifs—`GGGGAGGGGG`, `CCCCCCCCCC`, and `TCTGCT.2`—tended to be GC-rich or repetitive (Figure 4). These motifs may generally contribute to transcript stabilization or RNA structural motifs that enhance processing efficiency.

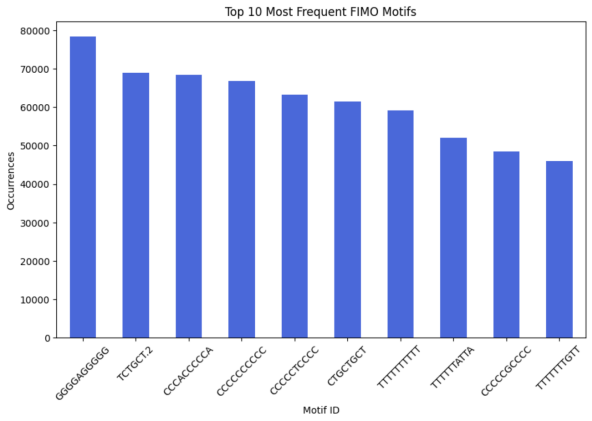


Fig. 4: Top 10 Frequent FIMO Motifs

We then stratified motifs by their statistical association with PRODUCT-TG expression using a fold-change approach. The top upregulating motifs—such as `TTTGGCTTCT`, `ACCACACCCG`, and `CGGGAGAGGT`—were enriched in high-expressing samples (Figure 5). These may correspond to elements enhancing translational efficiency or stability and are prime candidates for rational transgene design.

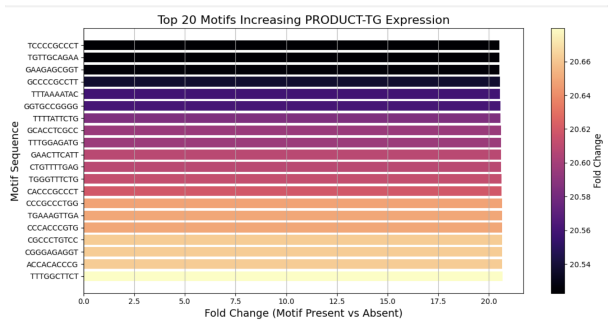


Fig. 5: Top 20 Motifs increasing Product-TG expression

In contrast, a separate analysis of downregulating motifs revealed sequences significantly associated with reduced PRODUCT-TG expression (Figure 6). These included homopolymeric and AU-rich motifs such as `GAAAAAAAGA`, `TAATAATAAT`, and `CTTTTTTATT`, as well as palindromic sequences like `TCCCCTCTCG`. These motifs may recruit destabilizing RNA-binding proteins or contribute to transcript degradation signals such as AU-rich elements (AREs).

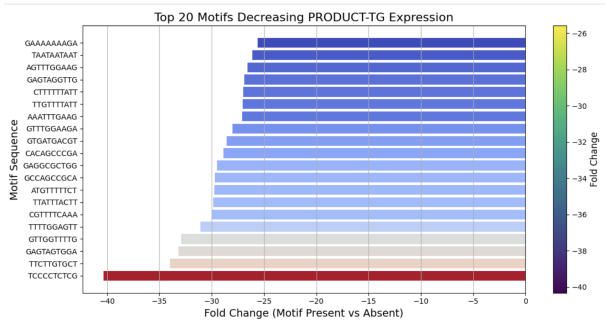


Fig. 6: Top 20 Motifs decreasing Product-TG expression

Their strong negative association (fold change $<$ -25 for all top 20, and $<$ -40 for the most repressive) suggests these should be explicitly avoided in synthetic 5'UTR or 3'UTR constructs to prevent unintended repression of transgene expression.

Global Motif Significance: Figure 7 provides a global overview of all motifs evaluated, showing both their $\log_2$ fold change and statistical significance. Red points denote motifs significantly associated with expression shifts ($p_{\text{adj}} < 0.05$). The upper-right quadrant contains high-confidence upregulating motifs, while the leftmost region captures potential repressors.

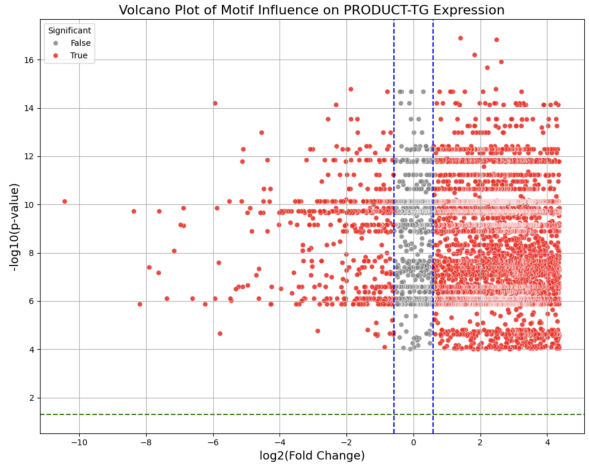


Fig. 7: Influence of Motifs on Product-TG expression

Taken together, this motif analysis highlights both promoting and suppressive elements that correlate with PRODUCT-TG expression, providing a data-driven basis for rational UTR engineering.

E. Technical Challenges

Several technical challenges were encountered during the project, spanning data quality, model design, and tool integration:

- **Batch Variability:** RNA-seq expression data originated from multiple laboratories with varying experimental protocols, leading to significant batch effects. These were corrected using the ComBat algorithm. PCA plots (Figure 1) before and after correction illustrated improved sample clustering and variance stabilization.
- **Data Imbalance:** PRODUCT-TG expression was highly skewed, with a small number of samples exhibiting very high expression. This imbalance posed difficulties for classification models and motivated the use of class weighting in the Random Forest Classifier.
- **Tied Expression Values:** The initial use of TPM-normalized expression data resulted in many tied values, particularly for genes with zero or low expression. This affected correlation-based methods and hence endall's Tau correlation was also computed to provide a more robust assessment of monotonic associations before switching to raw expression values.
- **Ranking Strategy Development:** Designing a robust gene-ranking framework required careful integration of outputs from correlation, regression, and classification models. Balancing their relative importance via MinMax normalization and weighting was done iteratively to preserve interpretability and biological relevance.
- **MEME Suite Setup:** Installing and configuring the MEME Suite locally proved non-trivial, requiring manual dependency resolution and shell-based integration. Additional complexity arose in coordinating the input/output format handling between Python and MEME/FIMO tools.
- **Working with DNABERT:** DNABERT required sequences to be pre-tokenized and encoded according to model expectations. Its inference step was both time- and memory-intensive due to the transformer architecture, necessitating efficient batching and memory management strategies.

Despite these challenges, the integration of multiple analytical techniques ultimately enabled robust marker gene identification and characterization of constitutive gene sequence features, with direct implications for synthetic transgene design in CHO cells.

VI. Future Work and Improvement

Future directions include:

- Experimental validation of top marker genes using independent CHO datasets or qPCR assays.
- Functional validation of identified 5UTR motifs via reporter assays.
- Integration of proteomic and metabolomic data to refine clone selection markers.
- Exploration of secondary structure features in UTRs and CDS to predict mRNA stability.
- Application of alternative embedding models to complement DNABERT findings.
- Comparative evaluation of alternative deep embedding models such as Nucleotide Transformer for sequence representation.
- Use of enriched 5'UTR motifs in synthetic constructs to test their influence on transcript stability and expression yield.

VII. Conclusion

This project successfully identified gene markers strongly correlated with PRODUCT-TG expression and characterized sequence-level features of constitutive genes in CHO cells. By integrating statistical correlation, random forest modeling, and composite scoring, we built a robust gene ranking framework to support predictive clone selection.

We further analyzed sequence properties such as GC content, DNABERT-based embeddings, and motif enrichment across 5'UTR, CDS, and 3'UTR regions. Our findings revealed that stable expression is associated with specific compositional and regulatory features, including conserved GC-rich motifs and repressive AU-rich sequences.

Importantly, we identified motifs significantly associated with both increased and decreased PRODUCT-TG expression. These insights provide dual guidance for synthetic design: upregulating motifs (e.g., `TTTGGCTTCT`, `CGGGAGAGGT`) may be leveraged to enhance expression, while downregulating motifs (e.g., `TCCCCTCTCG`, `GAAAAAAAGA`) should be excluded from synthetic constructs.

Altogether, our results contribute to AstraZeneca's long-term strategy for rational cell line engineering. Future applications could combine predictive gene markers with customized UTR designs to enable fully machine-guided transgene optimization workflows—reducing time-to-market and improving recombinant protein yields in biomanufacturing.

References

[1] K. e. a. Jayapal, "Recombinant protein therapeutics from cho cells—20 years and counting," *Chemical Engineering Progress*, 2007.

[2] G. Walsh and E. Walsh, "Biopharmaceutical benchmarks 2022," *Nature Biotechnology*, vol. 40, no. 12, pp. 1722–1760, 2022.

[3] S. e. a. Fischer, "Cho cell line development for therapeutic protein production: current state and future perspectives," *Biotechnology letters*, 2015.

[4] P. e. a. Gronemeyer, "Cho cell lines in biotechnology: Applications, advantages, and challenges," *Biotechnology journal*, 2014.

[5] X. e. a. Xu, "Rna-seq analysis of cho cell transcriptome and its relevance to biopharmaceutical production," *Biotechnology and bioengineering*, 2011.

[6] C. Mayr, "Regulation by 3'-untranslated regions," *Annual Review of Genetics*, vol. 51, pp. 171–194, 2017.

[7] N. Ryczek, A. Łyś, and I. Makałowska, "The functional meaning of 5'utr in protein-coding genes," *International Journal of Molecular Sciences*, vol. 24, no. 3, p. 2976, 2023.

[8] Y. Chu *et al.*, "A 5 utr language model for decoding untranslated regions of mrna and function predictions," *Nature Machine Intelligence*, vol. 6, pp. 449–460, 2024.

[9] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[10] R. Z. Edros, S. McDonnell, and M. Al-Rubeai, "Using molecular markers to characterize productivity in chinese hamster ovary cell lines," *PloS One*, vol. 8, no. 10, p. e75935, 2013.

[11] N. Ryczek, A. Łyś, and I. Makałowska, "The functional meaning of 5'utr in protein-coding genes," *International Journal of Molecular Sciences*, vol. 24, no. 3, p. 2976, 2023.
[12] C. Mayr, "Regulation by 3'-untranslated regions," *Annual Review of Genetics*, vol. 51, pp. 171–194, 2017.
[13] G. Hanson and J. Coller, "Codon optimality, bias and usage in translation and mrna decay," *Nature Reviews Molecular Cell Biology*, vol. 19, no. 1, pp. 20–30, 2018.
[14] M. Lovrić, K. Pavlović, M. Vuković, S. K. Grange, M. Haberl, and R. Kern, "Understanding the true effects of the covid-19 lockdown on air pollution by means of machine learning," *Environmental Pollution*, vol. 266, p. 115900, 2020.
[15] Y. Chu *et al.*, "A 5' utr language model for decoding untranslated regions of mrna and function predictions," *Nature Machine Intelligence*, vol. 6, no. 6, pp. 449–460, 2024.
[16] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "Meme suite: tools for motif discovery and searching," *Nucleic acids research*, vol. 37, no. suppl_2, pp. W202–W208, 2009.
[17] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.
[18] "Glul glutamate-ammonia ligase [cricetulus griseus]." https://www.ncbi.nlm.nih.gov/gene/100764163. Accessed: 2025-04-22.
[19] "Actb actin beta [cricetulus griseus]." https://www.ncbi.nlm.nih.gov/gene/100689477. Accessed: 2025-04-22.
[20] "Gapdh - uniprot p17244." https://www.uniprot.org/uniprotkb/P17244/entry. Accessed: 2025-04-22.
[21] I. Wool and Y. Chan, "The roles of ribosomal protein s13 in ribosome biogenesis and translation," *Trends in Biochemical Sciences*, vol. 24, no. 4, pp. 178–183, 1999.
[22] "Nsf - n-ethylmaleimide-sensitive factor - uniprot p18708." https://www.uniprot.org/uniprotkb/P18708/entry. Accessed: 2025-04-22.