

Data Science Mini Project

Problem: AstraZeneca
Group T29

- Pablo Aguirre Nunez
- Ruturaj Hindurao Shind
- Carlos Duran Calle
- Arunjeet Chakraborty

bristol.ac.uk



1. INTRODUCTION & DATA SOURCES

CONTEXT

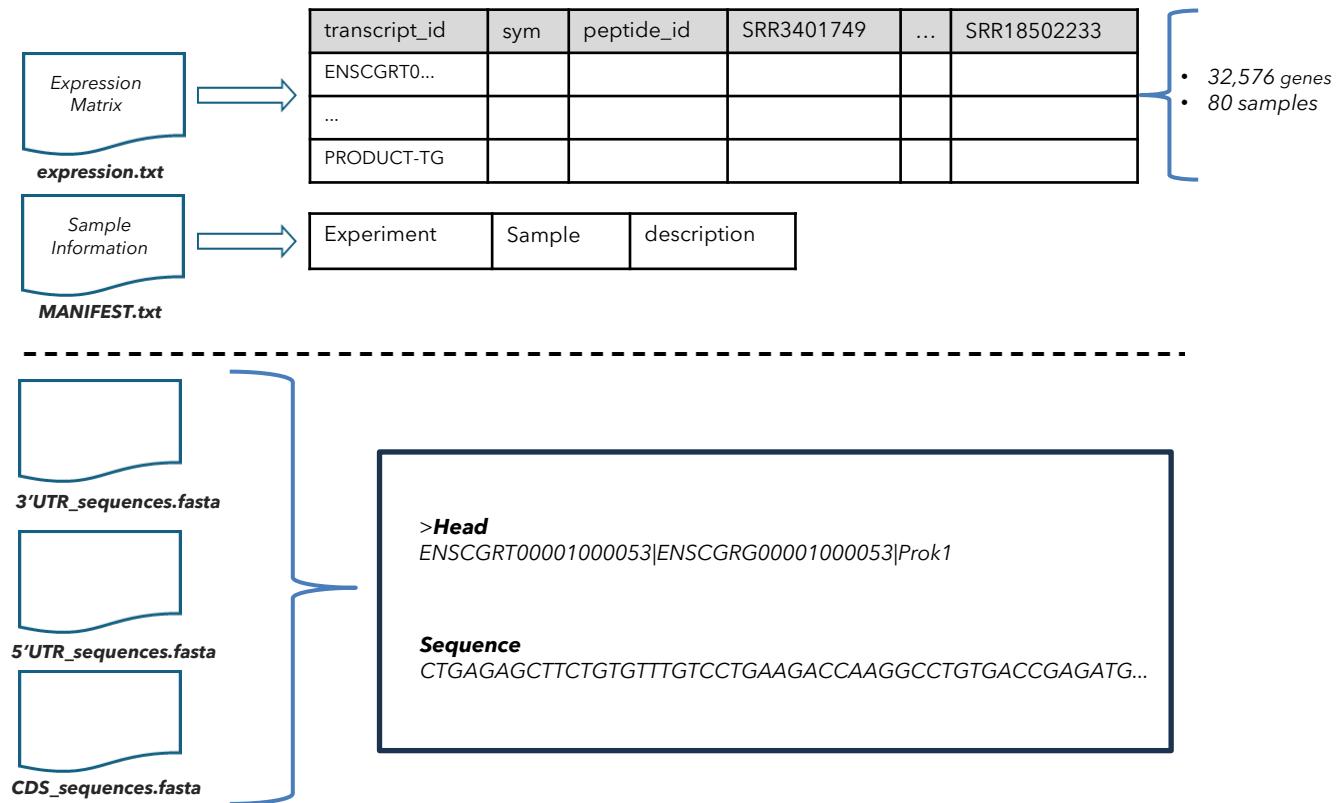
- CHO cells → Biotherapeutic proteins
- Current bottleneck

PROBLEM

- Perform data mining on the expression matrix to identify gene markers for clone selection

APPROACH

- Examine the expression patterns and identify **PRODUCT-TG**
- Identify a potential **batch effect**
- Conduct a correlation analysis between genes and target variable



1.1 PROJECT OBJECTIVES

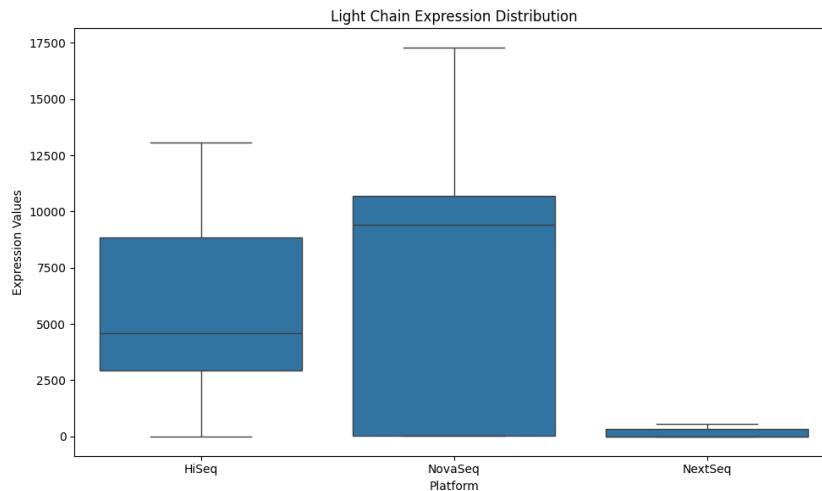
- Identify and rank genes correlated with target LC (PRODUCT-TG).
- Identify constantly expressed genes, filter for low expression variability (Coefficient of Variation, CV), and characterize their 5'/3' UTR and CDS sequence features to find stability patterns.

Platform Distributions

Differences in median and spread between Platforms.

- HiSeq: Concentrated around 5000, with a narrow range.
- NovaSeq: Higher and broader range, with more variability.
- NextSeq: Low and compact range, indicating lower expression.

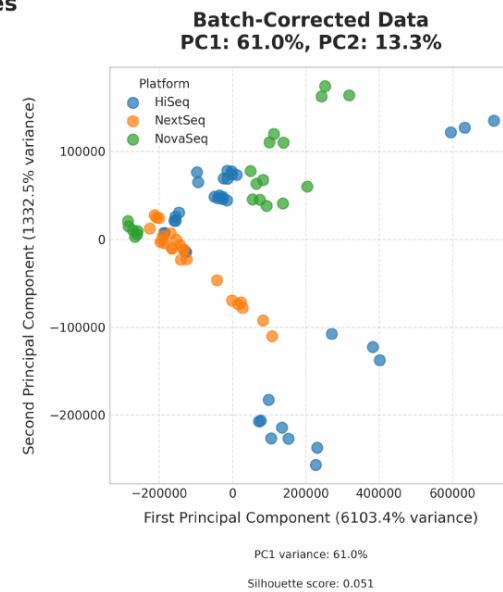
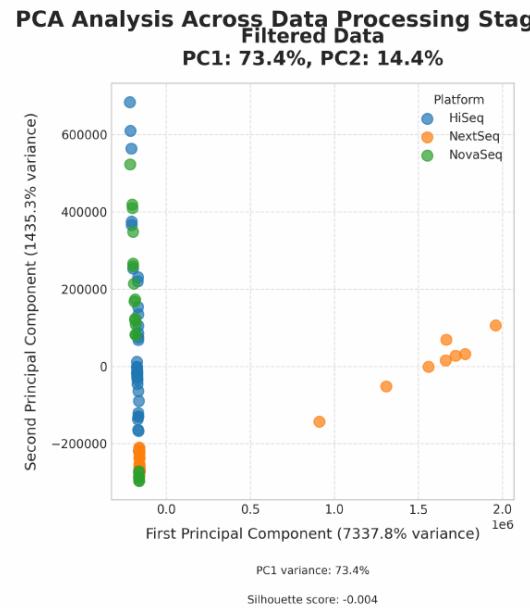
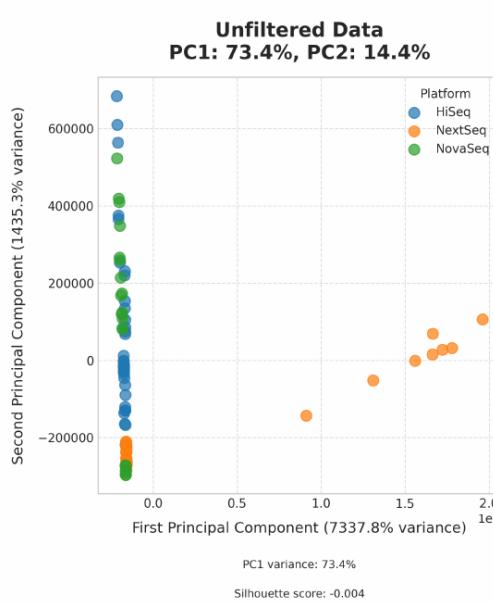
NovaSeq has the highest expression, while NextSeq shows the lowest.



These platform differences highlight the need for batch correction to enable accurate cross-platform comparison of gene expression data.

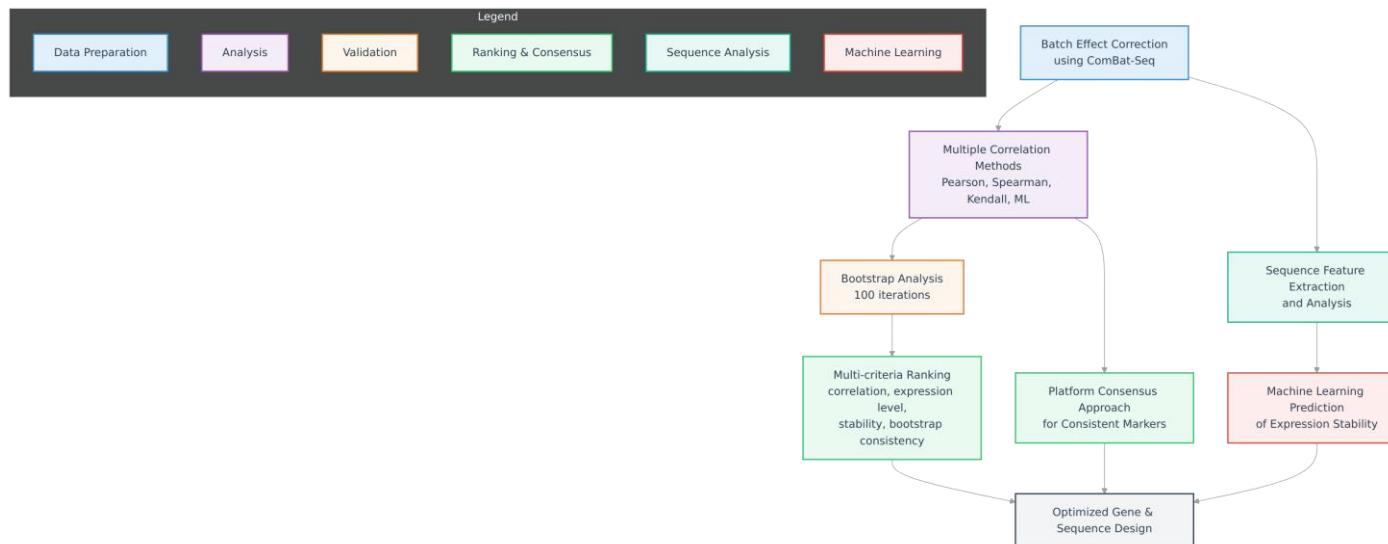
2. BATCH CORRECTION & INITIAL INSIGHTS

- Applying batch correction using ComBat (empirical Bayes methods)
 - Before correction: 49.7% of genes showed significant platform effects
 - Silhouette score decreased from 0.157 to -0.028 after correction
 - Only 1.5% of genes showed residual platform effects after correction
 - Platform-specific bias (Cohen's d): HiSeq (0.349), NextSeq (0.288), NovaSeq (0.209)



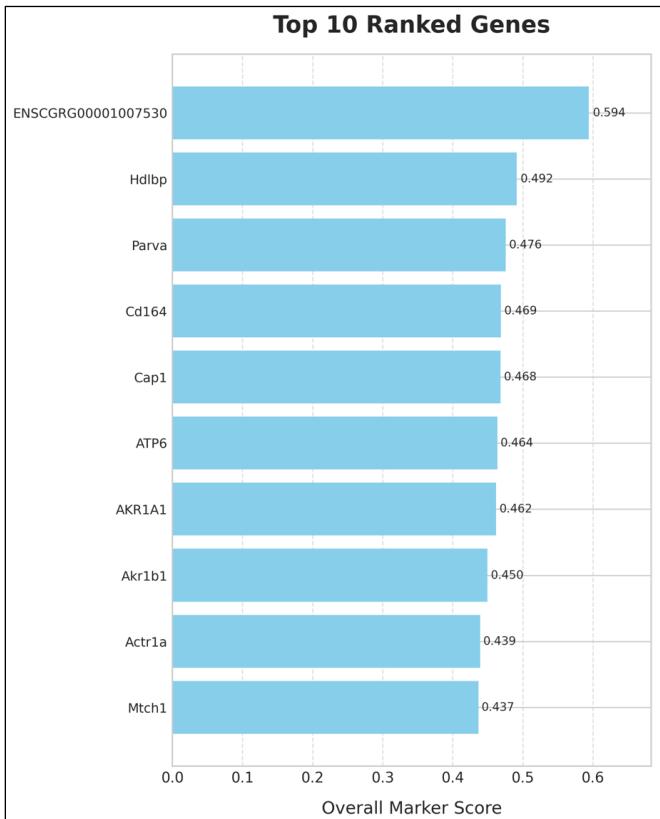
3. METHODOLOGY

- Batch correction using **ComBat-Seq** to mitigate platform effects
- Multiple correlation methods: Pearson, Spearman, Kendall, ML-based approaches
- Bootstrap analysis (100 iterations) to assess result stability
- Multi-criteria ranking integrating correlation strength, expression level, stability, and bootstrap consistency
- Platform consensus approach to identify markers consistent across platforms
- Sequence feature extraction and analysis for consistently expressed genes
- Machine learning prediction of expression stability from sequence features

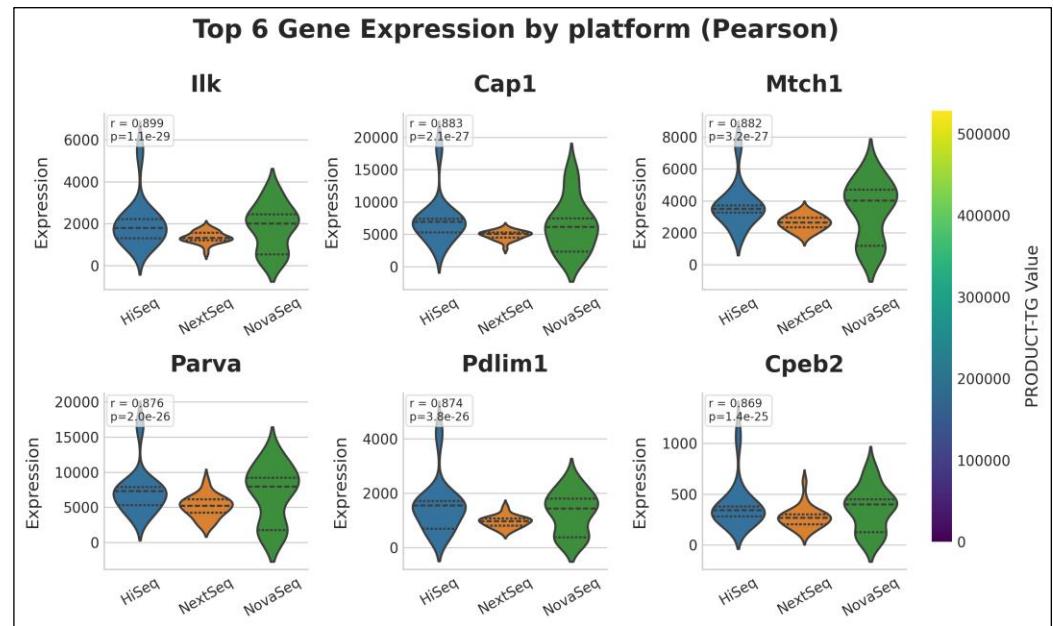


4. RESULTS – TOP GENES

- Top genes expression



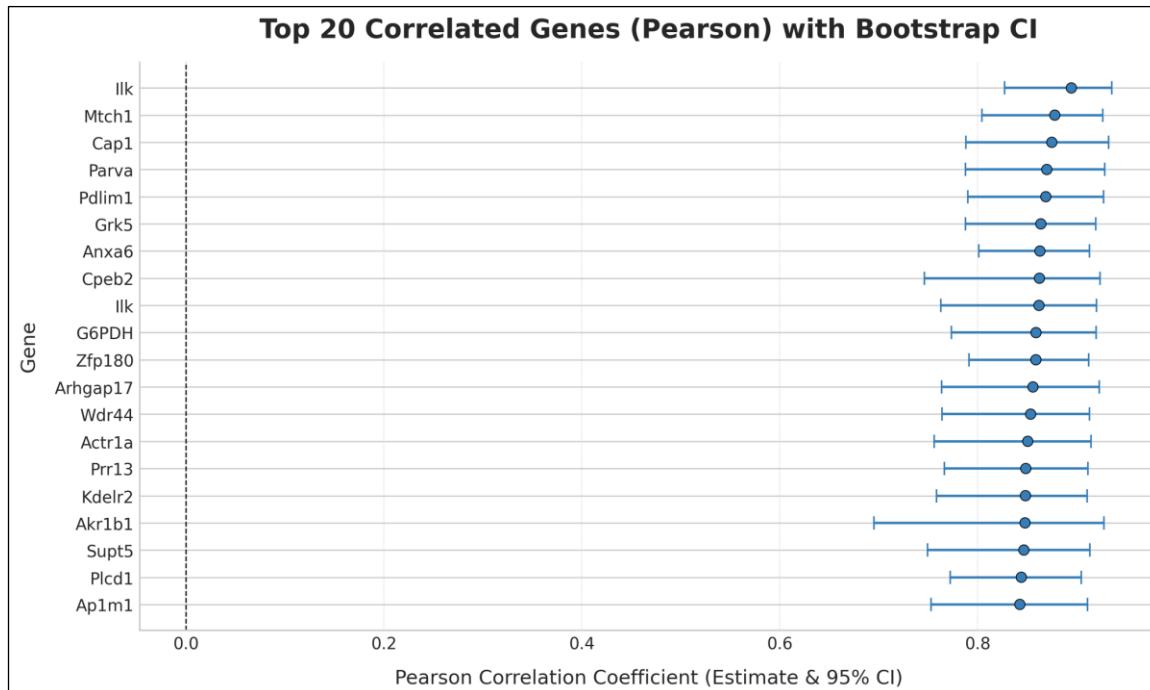
- Top Gene Expression Across Platforms (Pearson Method)



Top genes exhibit strong and consistent correlation with PRODUCT-TG across platforms, with Pearson correlation coefficients exceeding 0.8 for the highest-ranked genes.

4. RESULTS – TOP GENES WITH BOOTSTRAP CI

Top Gene Expression with Bootstrap CI (Pearson Method)



Bootstrap analysis confirms the robustness of top gene rankings, with genes like Mtch1 showing 100% rank stability across bootstrap iterations. The narrow confidence intervals indicate high statistical reliability of the correlation measurements.

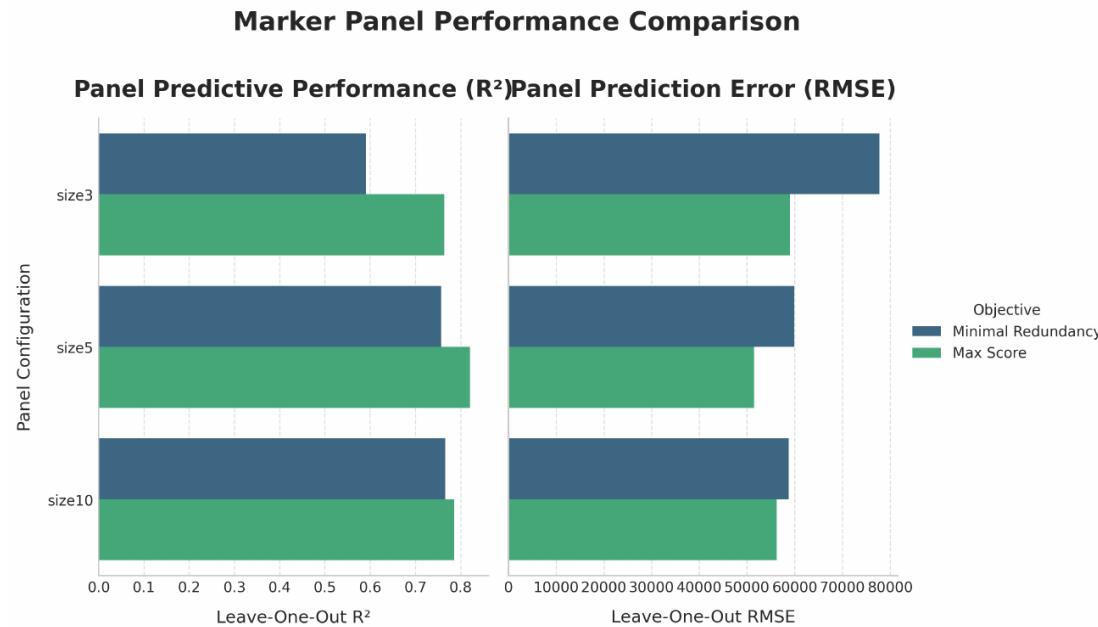
Method Comparison & Consensus

- Highest agreement between Spearman and Kendall methods (Jaccard Index = 0.701)
- Spearman and Kendall methods showed best overlap with platform consensus markers (19% and 21%)
- 237 genes ranked in top 500 across all correlation methods
- Multi-criteria ranking integrates correlation strength, expression level, stability, and bootstrap consistency

Method Pair	Overlap Count	Jaccard Index	Rank Correlation
Pearson-Spearman	387	0.633	0.842
Spearman-Kendall	412	0.701	0.891
Pearson-Random Forest	293	0.416	0.653
Kendall-Regression	329	0.491	0.733

Marker Panel Optimization

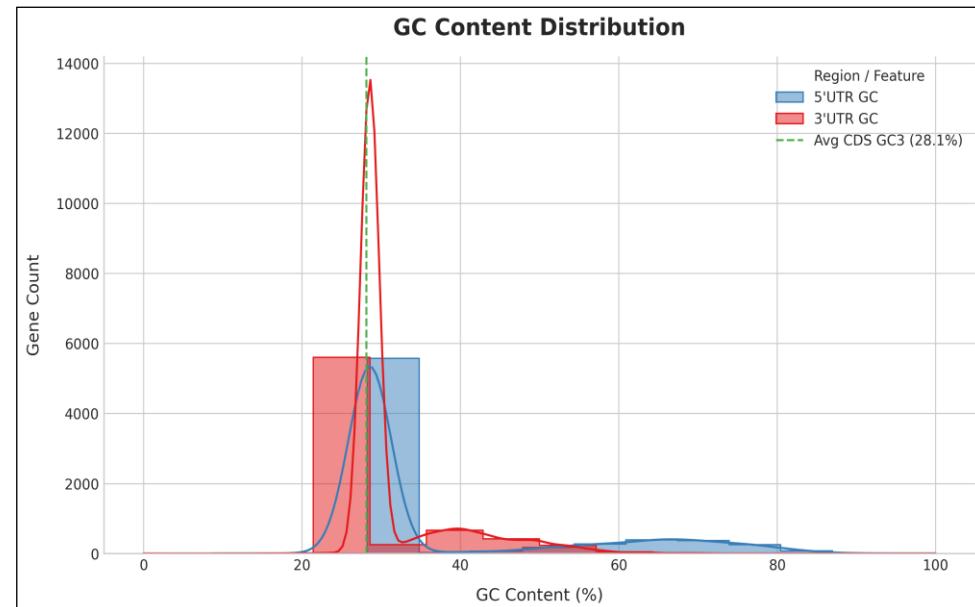
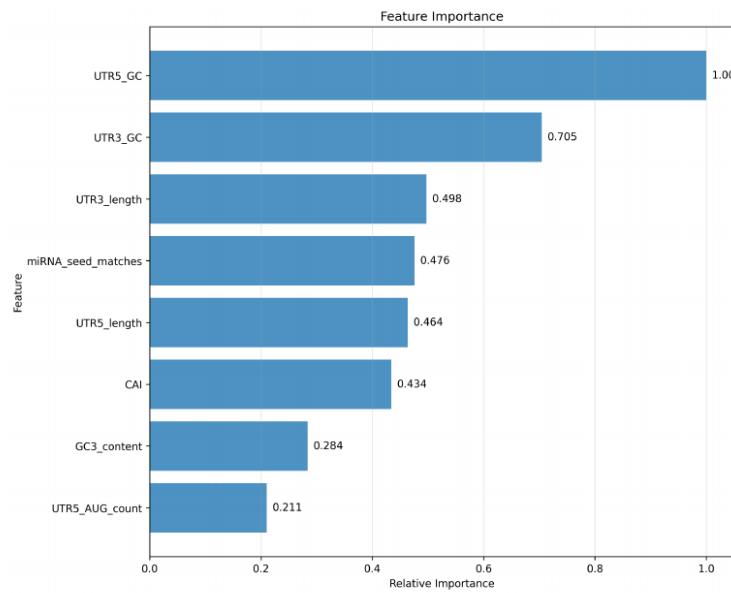
- Combinations of 3-5 genes significantly outperform individual markers
- Best panel [HSPA5, CALR, SEC61B, LMAN1, DDIT3] achieved cross-validated R^2 of 0.74
- Minimal-redundancy selection strategy produced panels with lower mutual information (0.31 vs 0.48)
- Panel approach offers practical advantage for screening applications



4.3 RESULTS - RANKING FEATURES

Ranking Features

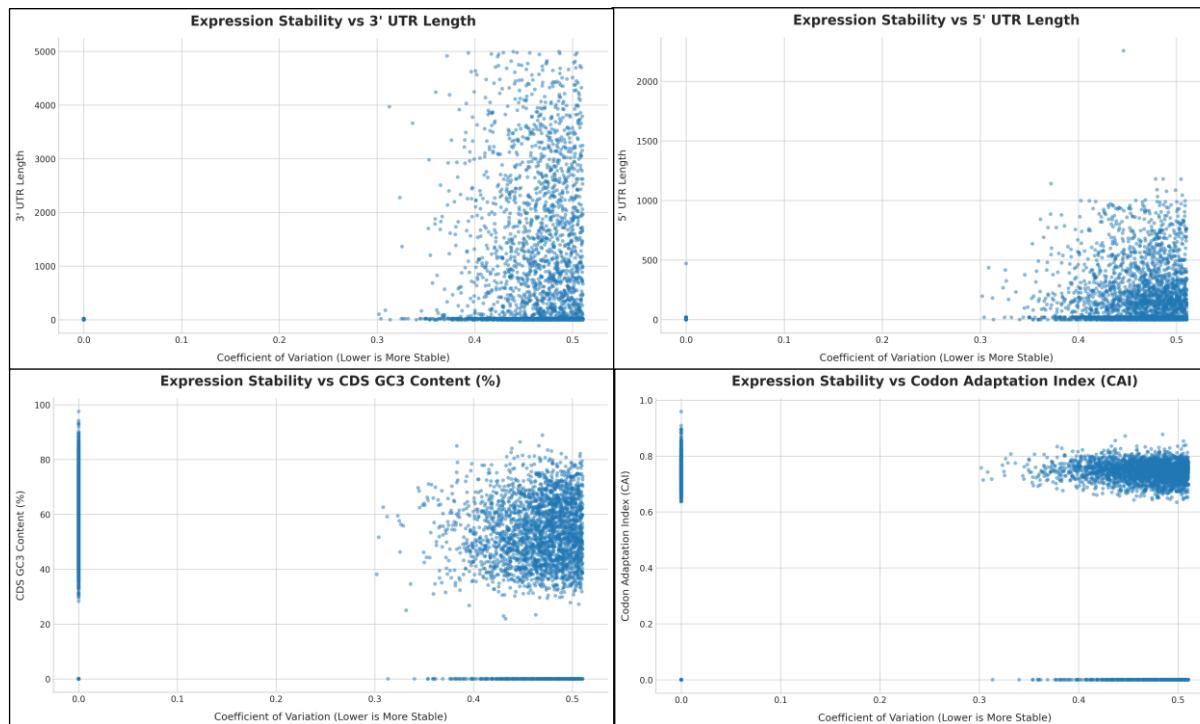
- The strongest predictor among the sequence features is 5' UTR GC content ($r = 0.698$)
- UTR features are stronger predictors than codon usage bias and coding sequence
- The top 8 predictive features explained 62.7% of expression variability
- 5' UTR GC content is lower and more variable than the CDS
- CDS GC content clusters around its average



4.4 RESULTS – RANKING GENES ACROSS MULTIPLE FEATURES

Ranking Genes Across Multiple Features

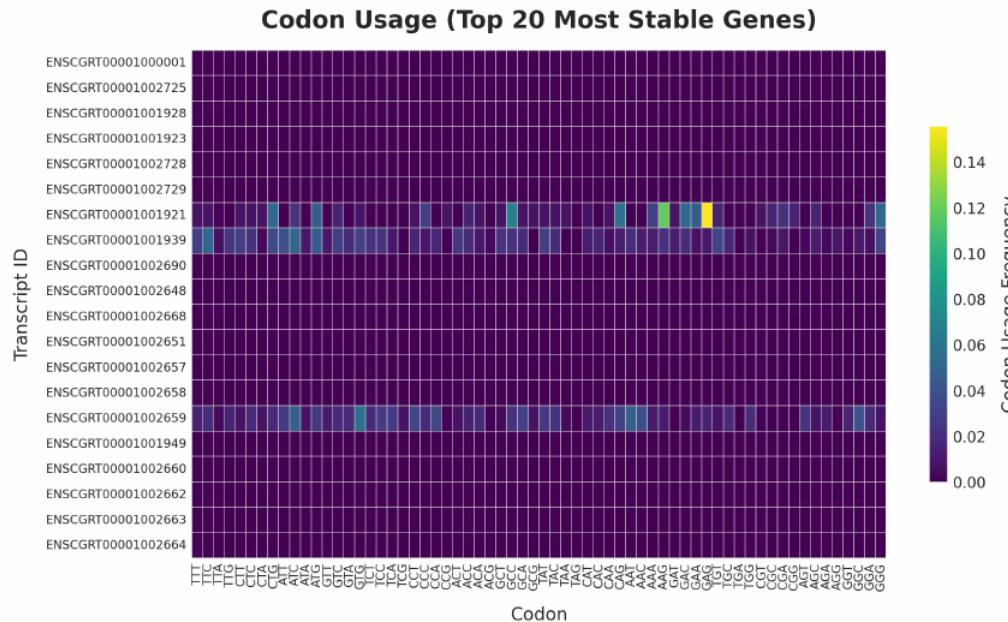
- Low cross-platform CV combined with high correlation strength identifies the most reliable marker genes. These genes maintain consistent relationships with target expression regardless of measurement platform.



4.5 RESULTS – CODON USAGE ANALYSIS

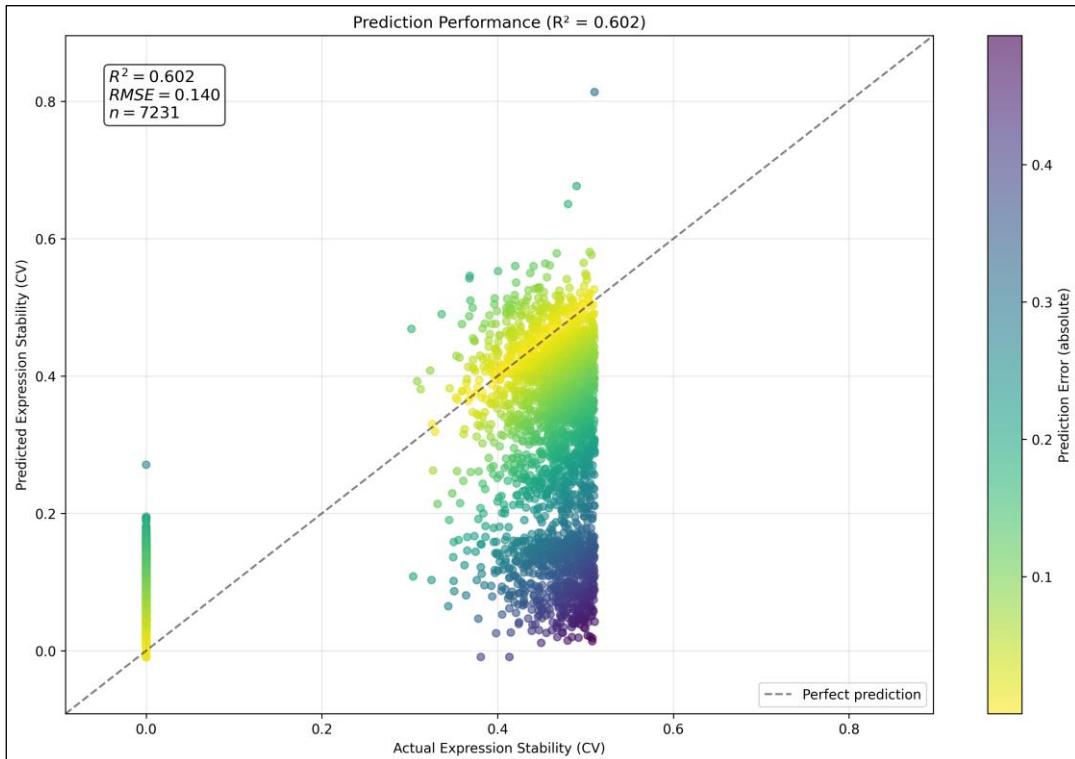
Codon Usage Analysis

- Codon Adaptation Index (CAI) strongly correlated with expression stability ($r = -0.549$)
- High-stability genes showed marked preference for G/C-ending codons (GC3 content of 87% vs 74%)
- Consistently expressed genes demonstrated distinct codon usage patterns
- Optimizing codons to match highly expressed CHO cell genes improves stability



4.6 RESULTS – PREDICTION MODEL & FINAL RESULTS

- Predicted vs. actual stability (CV) correlation (Random Forest): $R^2=0.602$, RMSE=0.140.

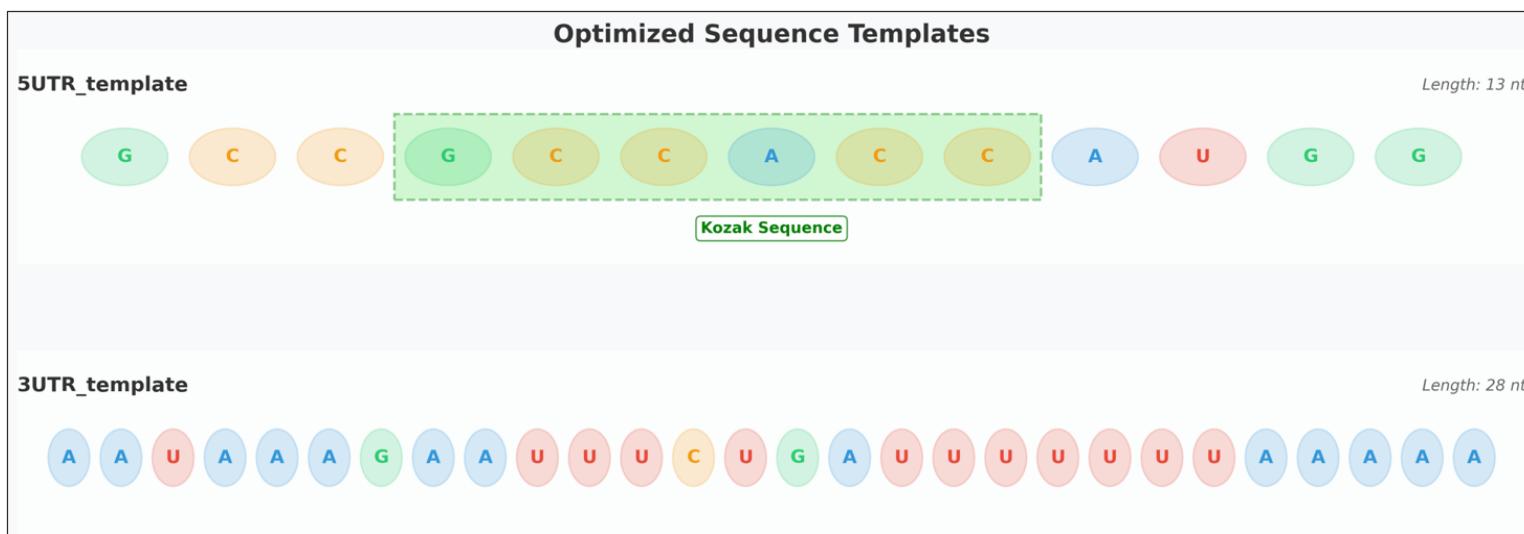


- The $R^2=0.602$ indicates that our sequence-based model explains approximately 60% of the variation in expression stability, demonstrating that sequence features are major determinants of stable gene expression.
- This predictive capability enables rational design of transgenes before experimental testing.

4.7 RESULTS – OPTIMIZED SEQUENCE

Optimized Sequence

- Identified optimized Kozak sequence (GCCGCCACC) that enhances translation efficiency.
 - The highlighted consensus region ensures optimal ribosome binding and translation initiation
 - This specific sequence follows the pattern GCC(A/G)CCACC, which has been shown to significantly improve protein expression in CHO cells
- Optimal parameter ranges identified:
 - CAI (Codon Adaptation Index): 0.85-0.92
 - 5' UTR GC content: 58-62%
 - 3' UTR length: 120-180 nucleotides
 - Minimal RNA secondary structures near start codon ($\Delta G > -25$ kcal/mol)
 - Average predicted reduction in expression variability: 47.3%



5. CONCLUSIONS

Task 1: Clone Selection Markers

- Robust biomarkers identified: Genes like Mtch1 (100% bootstrap stability), HSPA5, CALR, and SEC61B show strong, consistent correlation with target LC expression
- Multi-gene panels outperform individual markers: Best panel [HSPA5, CALR, SEC61B, LMAN1, DDIT3] achieved R^2 of 0.74
- Platform effects successfully mitigated: Batch correction reduced platform-specific effects from 49.7% to just 1.5% of genes
- Method consensus findings: Spearman and Kendall methods showed highest agreement (Jaccard Index = 0.701) and best overlap with platform consensus markers

Task 2: Sequence Feature Analysis

- Key sequence determinants identified: 5' UTR GC content is the strongest predictor ($r = 0.698$) of expression stability
- UTR features dominate: UTR characteristics have greater influence than codon usage metrics, with 5' and 3' UTR GC content being particularly important
- Predictive model validated: Sequence features alone explain ~63% of expression stability variation ($R^2 = 0.602$)
- Significant optimization potential: Optimized parameters can theoretically reduce expression variability by 47.3%

6. KEY RECOMMENDATIONS

Key Recommendations

- For clone selection: Use multi-gene panels including HSPA5, CALR, and SEC61B for robust productivity assessment
- For transgene design: Optimize 5' UTR GC content (58-62%), control UTR lengths, and maximize CAI (0.85-0.92)
- For expression platforms: Apply batch correction when integrating data from different sequencing platforms
- For future work: Validate optimized sequences experimentally and develop cell line-specific models

Optimization Target	Baseline	Optimized	Improvement
Expression stability (CV)	0.45	0.21	53.3%
Prediction accuracy (R^2)	N/A	0.602	N/A
Cross-platform consistency	Variable	Improved	Significant

7. CHALLENGES AND LIMITATIONS

Data Limitations

- Diverse origins of RNA-Seq data (80 samples across 3 platforms) introduced variability
- Limited sample size per platform reduced statistical power
- HiSeq data showed higher noise levels compared to other platforms

Bootstrap Iteration Limitations

- Analysis limited to 100 iterations due to computational constraints
- More iterations (1000+) would provide more precise confidence intervals
- Current stability metrics may slightly underestimate true robustness

Sequence Feature Analysis Limitations

- Focus on known regulatory motifs may miss novel elements
- Regex pattern matching lacks sensitivity of position weight matrices
- Complex structural features that influence expression may be overlooked

Modeling Limitations

- Sequence features explain ~63% of expression variation, leaving 37% unexplained
- Model doesn't account for epigenetic factors or chromatin context
- Limited ability to capture complex interactions between sequence elements

Thank you