Project Description

Project Title

Prediction of Gene Expression Using Histone Modification Levels at Promoters (1 kb upstream to 2 kb downstream of TSS).

Objectives

1. See maximum how much accuracy you can get.
2.What do you say about residuals i.e. what could compose of.
3. Compare it to RNA half-life.

Overview

This project aims to predict gene expression levels using histone modification signals at promoter regions. Promoter-associated chromatin marks are known to play a critical role in transcriptional regulation, and this study evaluates how much of gene expression variability can be explained using these epigenetic features.

Data and Preprocessing

RNA-seq data were processed in a Linux environment to obtain aligned reads and gene-level expression counts. Expression values were normalized to FPKM using gene length and library size. RNA-seq outputs contained Ensembl gene IDs, whereas histone modification datasets were annotated using gene symbols. To integrate these datasets, Ensembl gene IDs were mapped to HGNC gene symbols using Ensembl BioMart.

Methodology

Histone modification read counts (including marks such as H3K4me3 and H3K27ac) were extracted from promoter regions and used as predictive features. A random forest-based machine learning model was trained to predict FPKM values for approximately 29,598 genes in the GM12878 cell line.

Results

Correlation analysis showed strong associations between promoter histone marks and gene expression. Predicted versus actual FPKM values demonstrated good agreement, indicating effective prediction of transcriptional output from chromatin features. Residual analysis revealed variability not explained by promoter histone modifications alone.

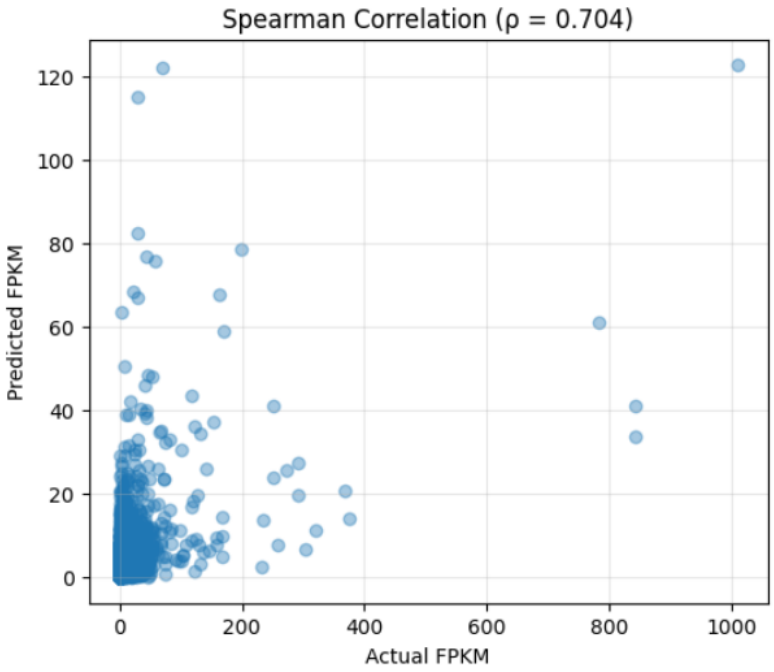

Figure 1. Predicted vs Actual FPKM values. The scatter plot demonstrates a strong agreement between predicted and observed gene expression levels, indicating that promoter-associated histone modifications have substantial predictive power for transcriptional output.

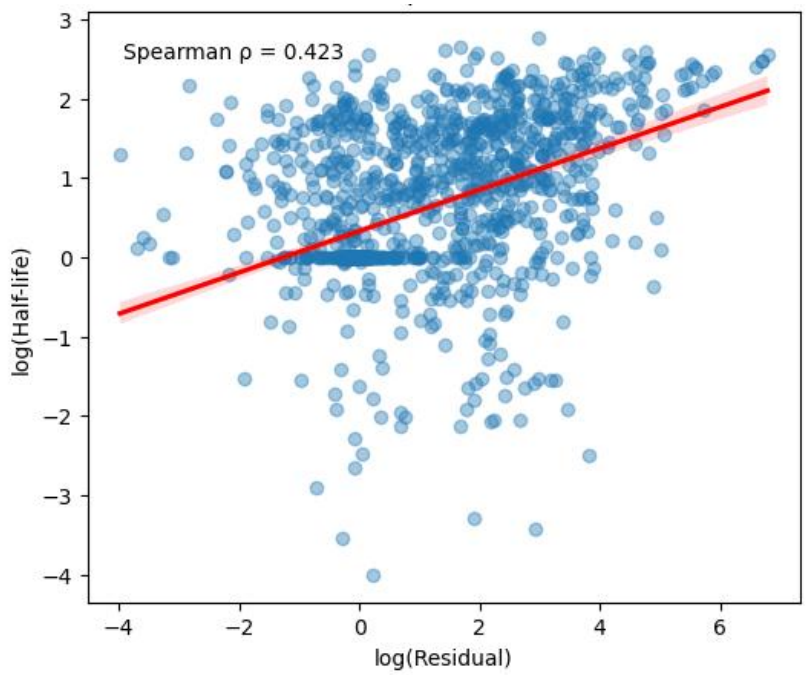

Figure 2. Relationship between log-transformed residuals and log-transformed RNA half-life. The observed correlation suggests that RNA stability and decay contribute significantly to gene expression variability that is not explained by promoter histone modifications alone.

Conclusion

This study demonstrates that histone modification levels at promoter regions exhibit a strong correlation with gene expression, as reflected by the high agreement between machine learning–predicted and observed FPKM values. These results confirm that promoter-associated chromatin features capture a substantial component of transcriptional regulation.

However, analysis of model residuals revealed systematic variability that could not be explained by promoter histone modifications alone. The significant correlation observed between log-transformed residuals and RNA half-life indicates that post-transcriptional mechanisms, particularly RNA stability and decay, contribute independently and substantially to gene expression variability.

The main finding of this study is that gene expression regulation is governed by a combination of transcriptional control mediated by promoter chromatin state and post-transcriptional regulation reflected in RNA half-life. While promoter histone modifications enable accurate prediction of expression trends, RNA decay processes account for a major fraction of the unexplained variation, highlighting the multi-layered nature of gene regulation.

Project Roles and Contributions

Ishika Gupta (MT25180) and Sunandini Chowdhury (MT25187) contributed equally to all stages of the project. Both members were jointly involved in RNA-seq preprocessing, FPKM calculation, gene identifier harmonization, feature extraction, model development, result analysis, interpretation, and preparation of documentation and presentation materials.