The School of Mathematics

THE UNIVERSITY
of EDINBURGH

Identifying the Influence of *Cryptococcus* Growth Media, Time, and K-mer Motifs on GAT201 Effect in Gene Expression Modelling

by

Adlensius Fransiskus Djunaedi

Dissertation Presented for the Degree of
MSc in Statistics with Data Science

August 2024

Supervised by
Dr Natalia Bochkina and Dr Simon Taylor

Executive Summary

This report aims to identify experiment design factors and k-mer motifs that significantly affect GAT201 effect on gene expression. Gene expression data of 6795 *Cryptococcus neoformans* genes and the frequencies of k-mer motifs were provided. Design factors include time samples were taken (0, 30, 120, or 240 minutes after inoculation), presence of GAT201 (WT if GAT201 is present or del if GAT201 is deleted), and growth media (YPD, RMPMI, or RPMI with serum) with 4 replicates of each sample.

Differential Gene Expression (DGE) analysis with DESeq2 was conducted by constructing generalised linear models (GLM) for each gene and assuming gene counts follow a negative binomial (NB) distribution. Design factors are deemed influential on the GAT201 effects when for many genes, the interaction term between them and GAT201 are significant. Then, gene expressions were modelled with k-mer motifs frequencies and design factors as features to obtain small sets of influential motifs. To avoid computational issues, a three-stage procedure is proposed: LFC computation with DGE analysis, motifs reduction with Lasso regression, and model fitting with GLM. Variables of interest are the interaction terms between motifs and GAT201. Finally, motif clustering was also done as a demonstration of dealing with multicollinearity. Model comparison was done with AIC to measure the models' performances.

Overall, results show that growth media does not significantly affect the GAT201 effect with only less than 1% of the genes exhibiting significant interaction terms between GAT201 and media. This concludes that GAT201 activation is serum-independent. Time, however, is relatively more influential to the GAT201 effect. An increasing trend in the number of genes showing significant interaction terms between time and GAT201 as time passes by from 8.10% at timepoint 30 to 16.40% at timepoint 240. This reveals that the GAT201 effects differ at different timepoints.

DGE analysis also demonstrated the significance of growth media and time in modelling gene expression. This result agrees with the principal component analysis results which show a clear separation of samples by timepoints and clustering based on media at certain timepoints. In particular, it was found that the GAT201 effect on gene expression is the strongest at timepoint 240. Therefore, analysis is conducted at timepoint 240 and its prior timepoint, 120, to identify the top influential k-mer motifs on GAT201. LFC computation, Lasso regression, and GLM fitting were done separately on 2-mer, 3-mer, 4-mer, and 5-mer data at both timepoints. Results show that the top 5 most influential are the 2-mer AT, TA, AC, CA, and AG motifs; the 3-mer AAT, TAA, CAA, GAA, and AAG motifs; the 4-mer AGAT, GTCA, GCTA, AATT, ACGC motifs; and the 5-mer AGATC, ATTGC, CAATA, AGGGA, and GTTTA motifs. Furthermore, motif clustering also successfully clustered highly correlated motifs into four groups. Three of the clusters were found to have significant interaction terms with GAT201. Model comparison revealed that the 4-mer model has the lowest AIC of 1561122, making it the most optimal model, while the cluster model has the highest AIC of 1577305, indicating its underperformance.

It is thus recommended for the Wallace lab to sample cells at other timepoints. As time was found to be influential on GAT201, sampling on other timepoints would reveal GAT201's effect in different timepoints. In addition, running the experiment for a longer timecourse could reveal the influence of media on GAT201 as more than four hours may be needed for the interaction effect to be apparent. Additionally, The Wallace lab could genetically engineer cells to focus on the top influential motifs and learn more about the GAT201 factor. Hopefully, the results of this report could help the Wallace lab in obtaining new insights on the behaviour of GAT201 and techniques in dealing with highly correlated k-mers data.

Nevertheless, the results provided are subject to several limitations. Firstly, the DESeq2 assumption on the inverse relationship between dispersion estimate and mean counts may be violated which could lead to unreliable results. Secondly, the implementation of model chaining may propagate errors throughout the procedure which may lead to misleading inferences. More advanced machine learning algorithms such as random forest and neural networks could be explored in future studies to deal with high-dimensional motif data. A more refined and rigorous motif clustering method could also be implemented to further improve its performance.

Acknowledgments

I would like to thank my supervisors, Dr Natalia Bochkina and Dr Simon Taylor, for their continuous support, guidance, enthusiasm, and generosity for staying beyond the meeting hours. I also would like to give special thanks to Dr Edward Wallace, for providing the opportunity and the dataset to work with along with his superb advice. His expertise and knowledge played a huge part in the completion of this report and this project would not have been possible without him. I also would like to thank PhD student helper Yiran Zhu for sharing his experience and help with programming and report writing. Finally, I would like to thank my friends for sharing their knowledge, views, and experience during discussions.

University of Edinburgh – Own Work Declaration

This sheet must be filled in, signed and dated - your work will not be marked unless this is done.

Name: Adlensius Fransiskus Djunaedi
Matriculation Number: S2591760
Title of work: Identifying the Influence of *Cryptococcus* Growth Media, Time, and K-mer Motifs on GAT201 Effect on Gene Expression

I confirm that all this work is my own except where indicated, and that I have:

- Clearly referenced/listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Not sought or used the help of any external professional academic agencies for the work
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Complied with any other plagiarism criteria specified in the Course handbook

I understand that any false claim for this work will be penalised in accordance with the University regulations (`https://teaching.maths.ed.ac.uk/main/msc-students/msc-programmes/statistics/data-science/assessment/academic-misconduct`).

Signature: Adlensius Fransiskus Djunaedi

Date: 8 August 2024

Contents

1 Introduction **1**
1.1 Background and Motivation . . . 1
1.2 Objective . . . 1

2 Exploratory Data Analysis **2**
2.1 Experiment Design Factors . . . 2
2.2 K-mer Motifs . . . 4

3 Methods **7**
3.1 Differential Gene Expression Analysis . . . 7
3.2 K-mer Motifs Analysis . . . 8
3.2.1 LFC Computation . . . 8
3.2.2 Lasso Regression for Motifs Reduction . . . 10
3.2.3 Generalised Linear Model . . . 10
3.2.4 K-mer Motif Clustering . . . 10

4 Results and Discussions **11**
4.1 Significant Experiment Design . . . 11
4.1.1 Effect of Growth Media . . . 11
4.1.2 Effect of Time . . . 12
4.2 Significant K-mer Motifs . . . 13

5 Conclusions **19**

Appendices **28**

A Word count **28**

List of Tables

1 Proportion of genes with adjusted p-values ≤ 0.05 for all growth media effect coefficient terms at different time points . . . 12
2 Proportion of genes with adjusted p-values ≤ 0.05 for all time effect coefficient terms . . . 13
3 Proportion of genes with adjusted p-values ≤ 0.05 for GAT201 effect coefficient terms . . 13
4 Full results of motifs with significant interaction term between the motifs and GAT201 . . 15
5 List of motifs clustered into four clusters . . . 23
6 Estimated coefficients, standard errors, and p-values of the interaction between top 5 most influential motifs and GAT201 . . . 24
7 AIC of GLM Models . . . 24

List of Figures

1 Distributions of gene expression according to different strains of genotype WT . . . 2
2 Distributions of gene expression according to different strains of genotype del . . . 3
3 PCA plot of PC1 vs PC2 colored according to Growth Media (Condition) . . . 3
4 PCA plot of PC1 vs PC3 colored according to Growth Media (Condition) and various shapes according to timepoint . . . 4
5 PCA plot of PC1 vs PC3 colored according to GAT201 presence and various shapes according to timepoint . . . 4
6 PCA plot of PC2 vs PC3 colored according to GAT201 presence and various shapes according to timepoint . . . 5
7 Heatmap of gene expression with hierarchical clustering . . . 5
8 PCA plot of 2-mer (top left), 3-mer (top right), 4-mer (bottom left), 5-mer (bottom right) motifs . . . 6
9 Correlation heatmap of k-mer motifs . . . 6
10 Mean vs variance of gene counts data. Red line indicates mean equal to variance. . . . 7
11 Histograms and Normal QQ plot of LFC at time 240 and 120. Most points in the QQ plots lie on the straight lines with some deviations around the tails shows that the assumption of normal distribution approximation of LFC is appropriate. . . . 9
12 Proportion of non-chosen motifs are not chosen in fifty simulated cases at timepoint 120 for 2-mer (top left), 3-mer (top right), 4-mer (bottom left), and 5-mer (bottom right) models 14
13 Proportion of non-chosen motifs are not chosen in fifty simulated cases at timepoint 240 for 2-mer (top left), 3-mer (top right), 4-mer (bottom left), and 5-mer (bottom right) models 14
14 Venn diagram indicating the count and percentage of 2-mer (top left), 3-mer (top right), 4-mer (bottom left), and 5-mer (bottom right) motifs that have significant interaction term with GAT201 at timepoint 120, 240, and the intersection between them . . . 16
15 Total within sum of square error for various number of clusters . . . 17
16 PCA plot of k-mer motifs clustered into 4 clusters . . . 17
17 Log transformed p-values of the interaction terms between 2-mer motifs and GAT201 . . 18
18 Log transformed p-values of the interaction terms between 3-mer motifs and GAT201 . . 18
19 Log transformed p-values of the interaction terms between 4-mer motifs and GAT201 . . 18
20 Log transformed p-values of the interaction terms between 5-mer motifs and GAT201 . . 19
21 Log transformed p-values of the interaction terms between clustered motifs and GAT201 . 19
22 Residual Plots for 2-mer model . . . 20
23 Residual Plots for 3-mer model . . . 20
24 Residual Plots for 4-mer model . . . 21
25 Residual Plots for 5-mer model . . . 21
26 Residual Plots for cluster model . . . 25

27 Screenshot of LaTeX word count of the main text 28

1 Introduction

1.1 Background and Motivation

Cryptococcus neoformans is a globally invasive fungal pathogen which causes Cryptococcosis infection [13]. Cryptococcosis in humans is mainly transmitted through inhalation of the spores from *Cryptococcus* where severe cases of pulmonary infection and meningitis are observed in immunocompromised hosts [15]. Thus, a thorough understanding of the factors that control the proliferation and dissemination of *Cryptococcus* is crucial in the development of treatments and preventive measures as they directly affect its virulence [9].

Cryptococcus environmental adaptation involves a trade-off between the metabolically demanding polysaccharide capsule synthesis as a defense mechanism and cellular proliferation [9, 16, 8]. A GATA family transcription factor, GAT201, was found to be a key regulator of Cryptococcal pathogenesis. Overexpression of GAT201 induces capsule production under a classically non-inducing environment while the deletion of GAT201 markedly promotes proliferation in RPMI media at alkaline pH [11, 9].

To achieve a comprehensive understanding of GAT201's behaviour, it is of interest to uncover the factors that could influence its effect on gene expression level in *Cryptococcus*. Furthermore, research has revealed the capability of DNA sequence and k-mers in explaining the variation of gene expression [2, 4, 14]. The identification of a small number of k-mers that could affect the GAT201 effect on gene expression could substantially reduce feature space and computational power requirement.

Differential gene expression (DGE) analysis of gene counts obtained through RNA sequencing will be employed to assess the effect of various features on GAT201. Fundamentally, DGE analysis involves a comparison of relative frequencies between two groups of gene counts. The challenges come from a large number of genes to be modelled paired with the small number of replicates for each gene which may introduce higher uncertainty on the power of tests [5]. Furthermore, common distributions such as binomial and Poisson are insufficient in modelling the variability of read counts [5]. The Negative Binomial (NB) distribution is thus a popular alternative distribution to be assumed [1].

1.2 Objective

This report aims to answer the following two research questions.

1. Which experiment design factors, i.e., growth media condition and amount of time passed after inoculation, significantly affect the GAT201 effect on gene expression level?

2. Which k-mer motifs significantly affect the GAT201 effect on gene expression level?

To address the first research questions, generalised linear models (GLM) will be constructed to perform DGE analysis and model gene expression levels across experiment design factors. Moreover, the main aim of the second research question is to model the relationship between motifs and gene expression separately for 2, 3, 4, and 5-mer motifs. It is also of interest to investigate the minimum required set of motifs from all 2, 3, 4, and 5-mer motifs combined that could significantly influence GAT201.

Statistically, both research questions aim to identify significant interaction effects between the features and GAT201 effect [6]. The results presented in this report are expected to provide the Wallace lab with a deeper insight into the behavior of GAT201 across various environments along with its relationship with a few significant motifs, thus allowing a more efficient analysis in the future.

The remainder of this report is organised as follows. Section 2 presents the exploratory data analysis which involves preliminary analysis and principal component analysis. Section 3 describes the methodology including differential gene analysis and k-mer motif analysis. Section 4 presents model performances and a discussion of the results. Finally, section 5 concludes the report.

2 Exploratory Data Analysis

2.1 Experiment Design Factors

RNA abundance measurements of 6795 unique genes from *C. neoformans* were collected before inoculation (0 minutes) and at three different time points, 30, 120, and 240 minutes following inoculation. Samples are either wildtype cells with GAT201 (WT) of strain a or A; or mutant cells with GAT201 deleted (del) of strain B or M. Moreover, Cells sampled at time 0 were taken in nutrient-rich growth medium (YPD) while cells sampled at other times were either in host-like RPMI-1640 with 10% Heat inactivated fetal calf serum (RS) or RPMI media only (R). 4 replicates, 2 replicates for each strain, were taken at each time point and media. Thus, there are a total of 56 samples of gene expression levels. Each sample is labeled in the form of `strain_media_timepoint_replicate`, e.g., `A_Y_0_1` represents strain A sample (i.e., genotype WT) with YPD growth media taken at timepoint 0 as the first replicate.

Firstly, genes with zero counts across all samples are excluded from all analysis as they are meaningless. Subsequently, data is normalised before exploratory data analysis (EDA) is conducted. Let $counts_{ij}$ be the count for gene i and sample j and m_j be the median count for sample j. Then, the normalised counts y_{ij} is

$$y_{ij} = \log(counts_{ij}) - \log(m_j) \tag{2.1}$$

Figure 1 and 2 illustrate the distributions of gene expression separated by different strains of genotype WT and del, respectively. It can be observed that the distributions between the two strains of WT and del cells are nearly identical for all design conditions. Therefore, no strain effect will be assumed in gene expression measurements.

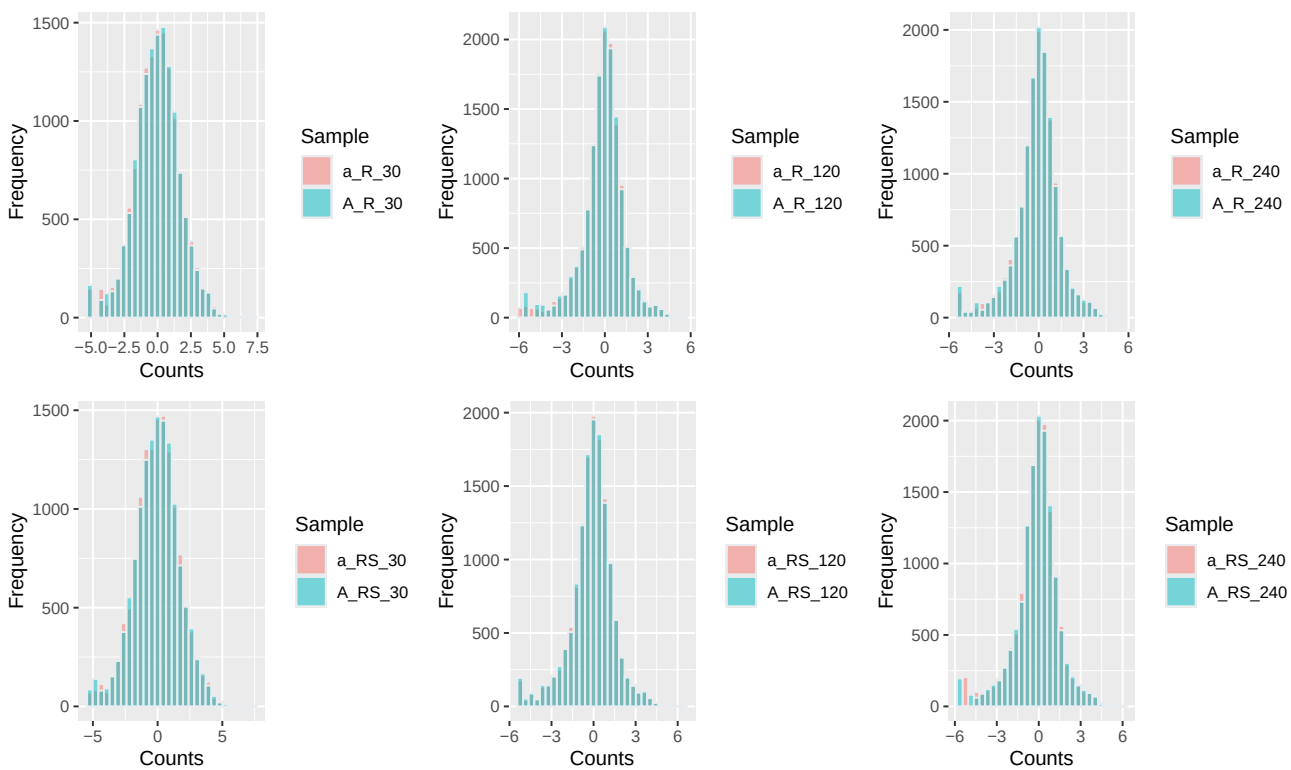

Figure 1: Distributions of gene expression according to different strains of genotype WT

Principal Component Analysis (PCA) was then conducted as a preliminary analysis. Results are shown in Figure 3, 4, 5, and 6. The percentages on the axis labels show the variance explained by the corresponding principal components (PC). Time is the most dominating feature with all four figures showing

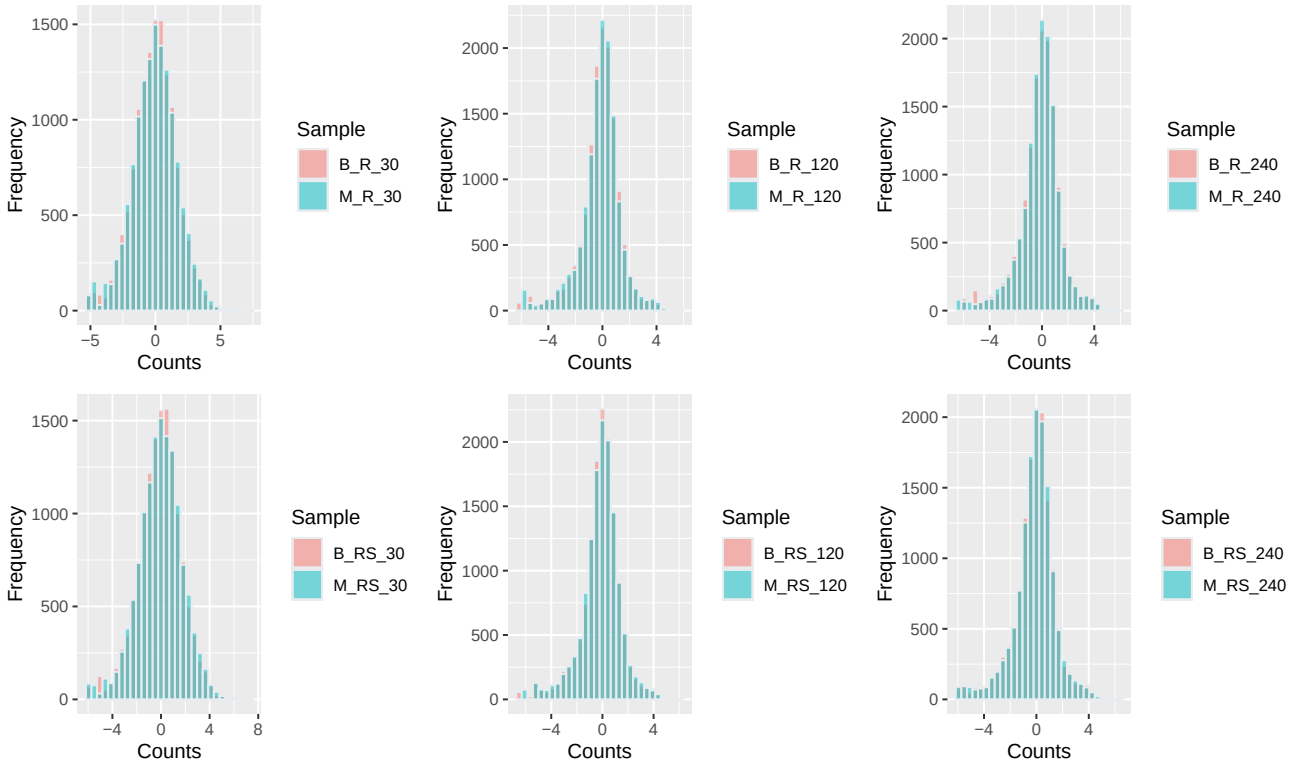

Figure 2: Distributions of gene expression according to different strains of genotype del

a clear separation of the four timepoints. PC1 seems to separate samples by time and condition/media feature. Similarly, PC2 appears to also describe the time of the samples. Figure 3 and 4 reveal that the condition/media effect may be present on time 30 and 120. Further, PC3 describes the GAT201 effect as shown in Figure 5 and 6 where it seems to be present at all time points. Overall, PC1, 2, and 3 can explain most of the variations (83.5%). The heatmap of gene counts in Figure 7 also shows similar results. Obvious clusters of timepoints can be seen followed by the GAT201 presence and growth media effects.

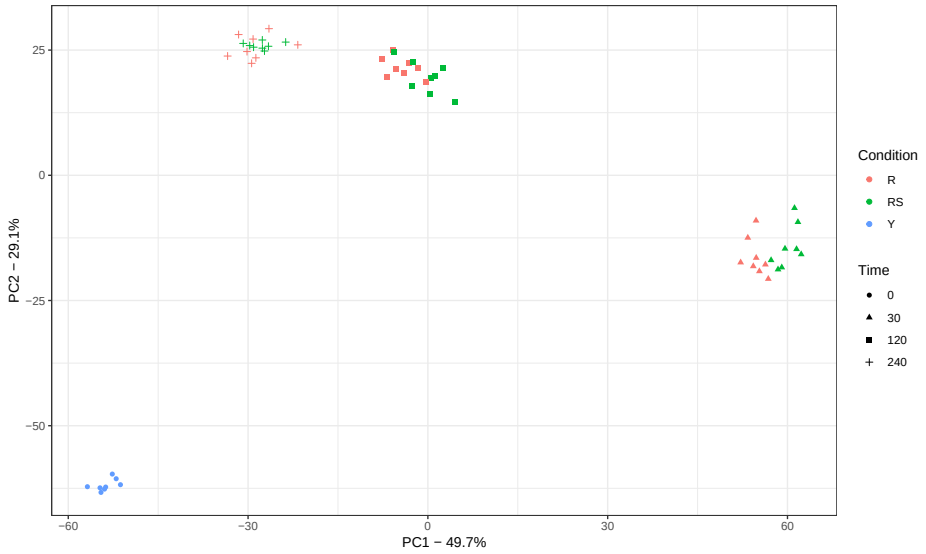

Figure 3: PCA plot of PC1 vs PC2 colored according to Growth Media (Condition)

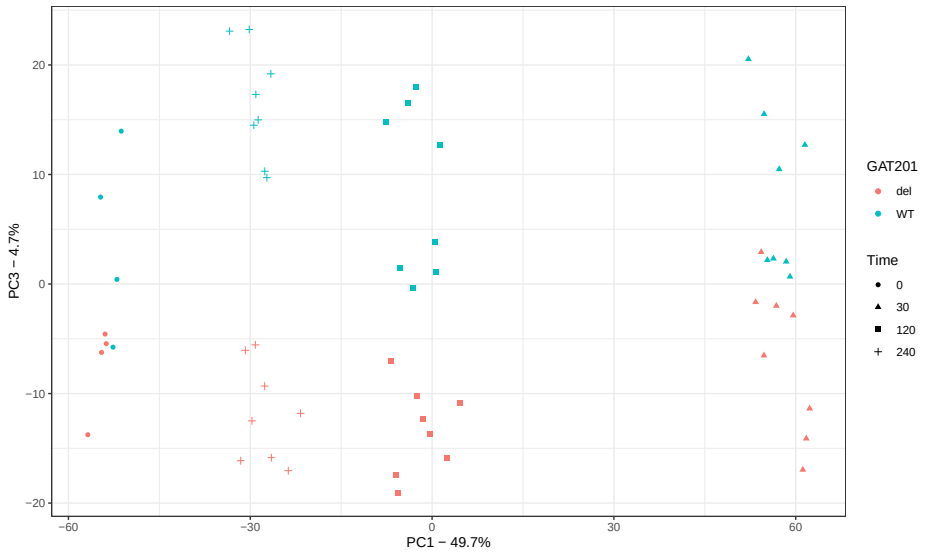

Figure 4: PCA plot of PC1 vs PC3 colored according to Growth Media (Condition) and various shapes according to timepoint

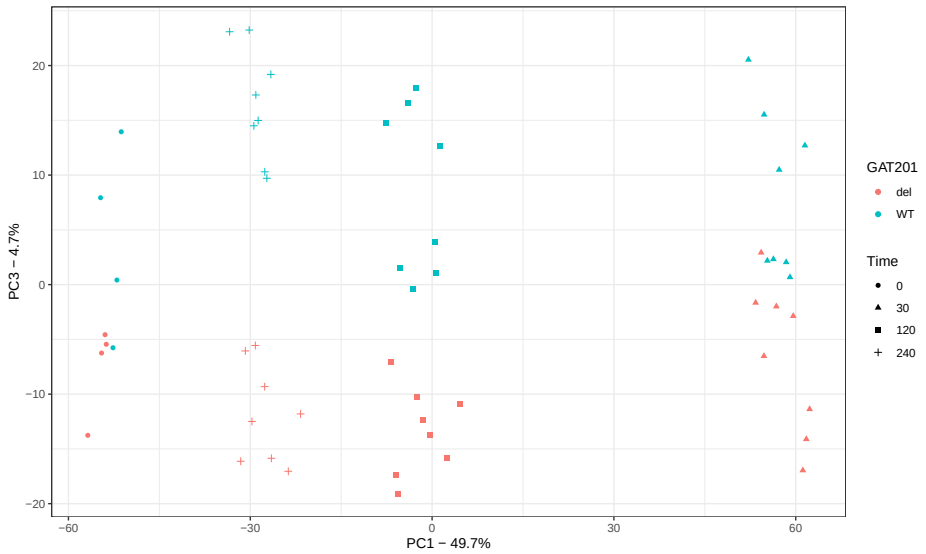

Figure 5: PCA plot of PC1 vs PC3 colored according to GAT201 presence and various shapes according to timepoint

2.2 K-mer Motifs

Frequencies of 2, 3, 4, and 5-mer motifs of each gene were provided. Due to the complementary nature of A-T and G-C nucleotides, a k-mer motif and its reverse complement are biologically equivalent. For example, the 5-mer motif ACGTT is equivalent to CCAGT. To address this, counts of each k-mer motif and their reverse complement for each gene are aggregated.

PCA was conducted on k-mer motif data to inspect similarly behaving motifs. Figure 8 shows the plot PC1 against PC2 for k-mer motifs. Some clusters can be observed for 2 and 3-mer motifs which provide preliminary evidence that there may exist groups of motifs that behave similarly such that a smaller significant set of motifs should be sufficient in explaining gene expression data. However, no clear

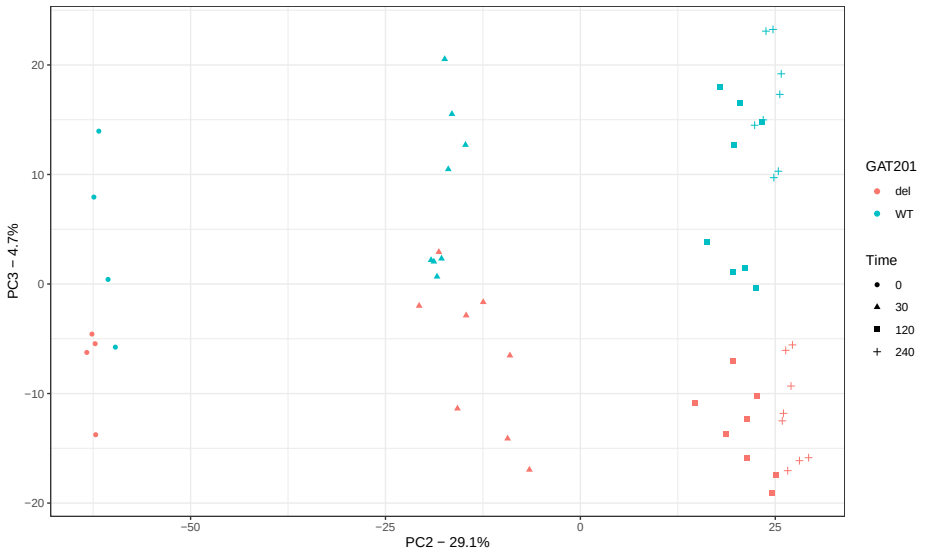

Figure 6: PCA plot of PC2 vs PC3 colored according to GAT201 presence and various shapes according to timepoint

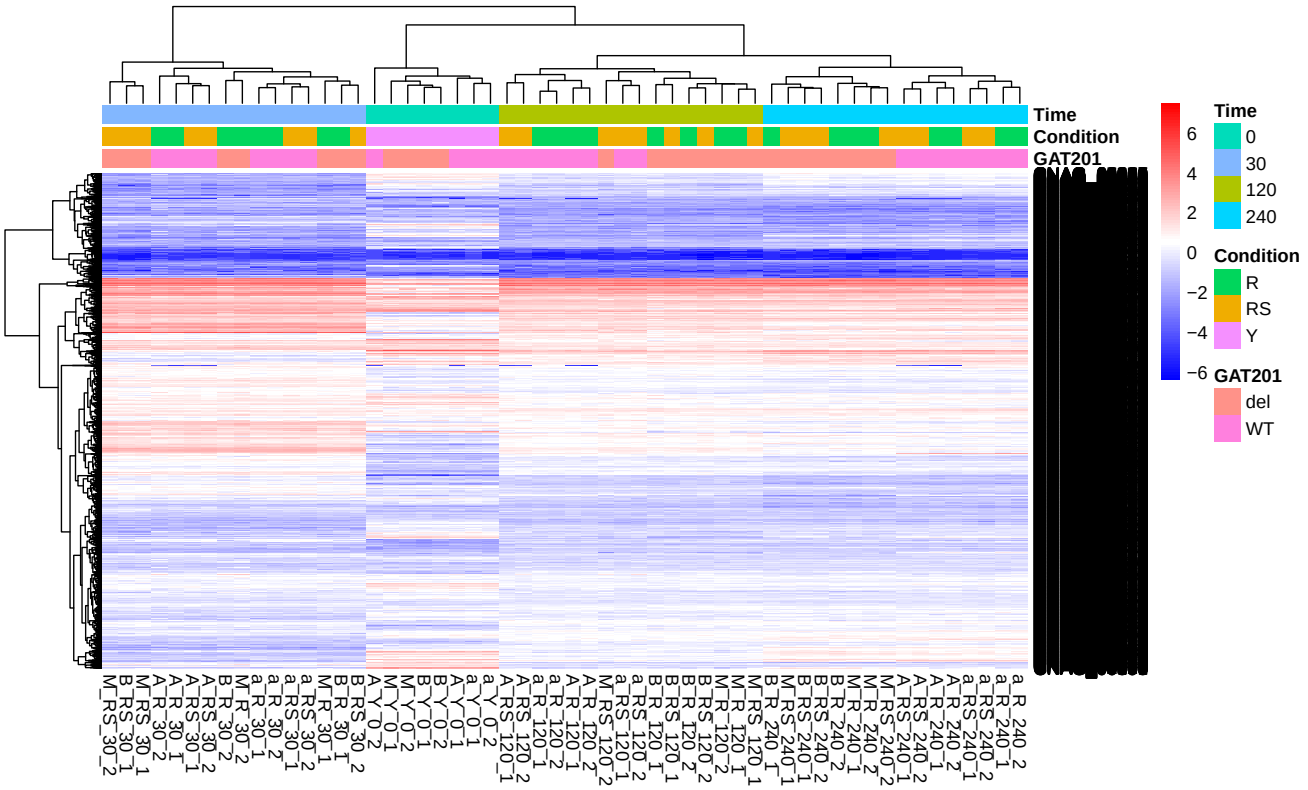

Figure 7: Heatmap of gene expression with hierarchical clustering

separation can be seen for the 4 and 5-mer motifs.

Furthermore, noting that (k+i)-mer motifs, $i \geq 1$, contain some k-mer motifs implies that a large amount of information is shared between the features when all 2, 3, 4, and 5-mer motifs are considered in a single model. High correlations among the features, as shown by the large amount of dark red and blue streaks in Figure 9, reflect this issue and may affect model stability [6].

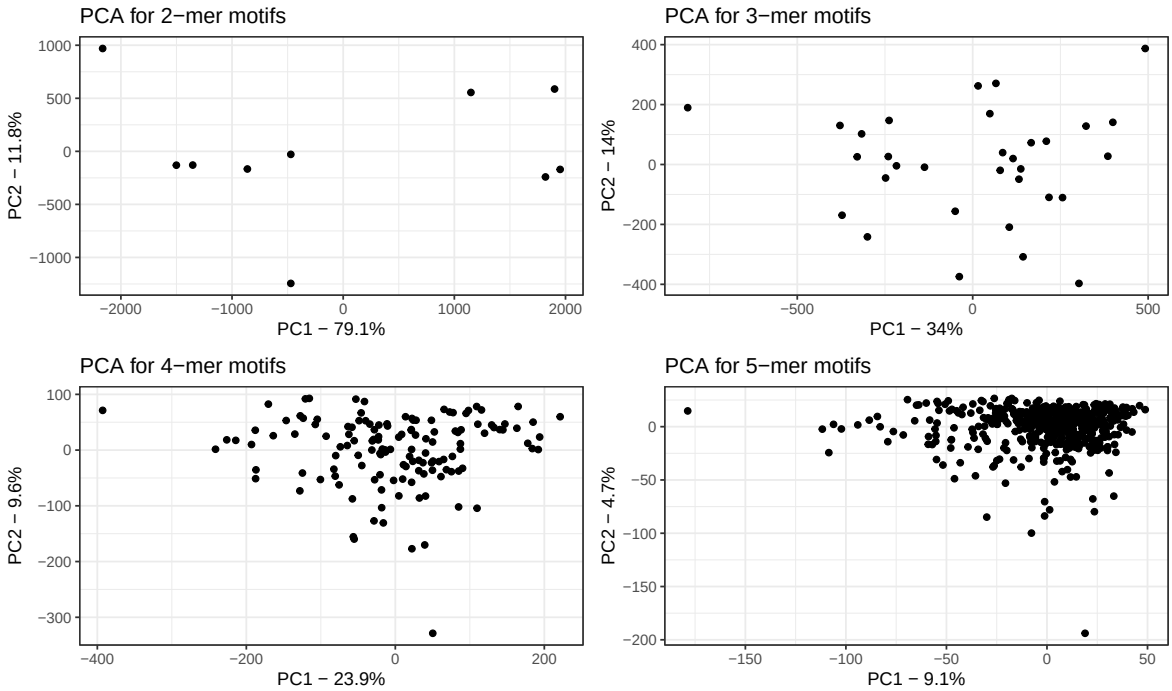

Figure 8: PCA plot of 2-mer (top left), 3-mer (top right), 4-mer (bottom left), 5-mer (bottom right) motifs

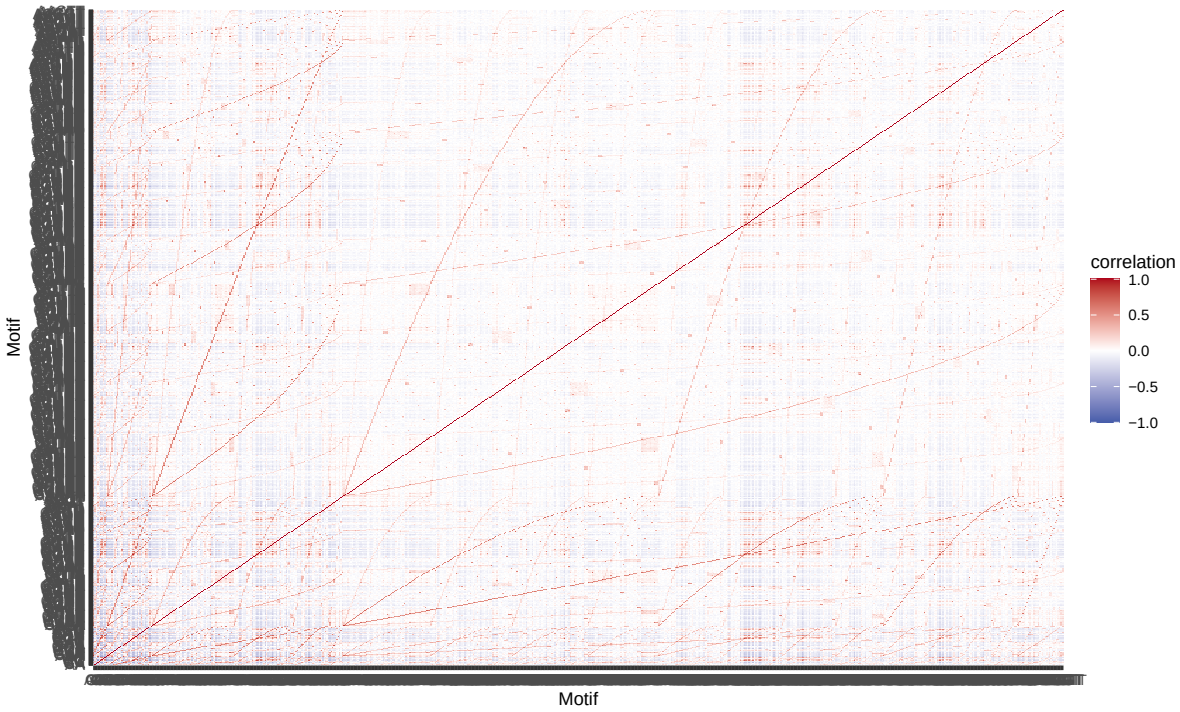

Figure 9: Correlation heatmap of k-mer motifs

3 Methods

DGE analysis will be performed to tackle the first research question while feature selection with Lasso regression along with GLM analysis to model gene expression will be the two main methods in addressing the second research question.

3.1 Differential Gene Expression Analysis

The DESeq2 method is employed to perform differential analysis of gene count data as it allows for correction on dispersion parameters by assuming the NB distribution and modelling its dependence with mean gene expression level to produce more stable estimates [12]. Let $X \sim NB(\mu, \alpha)$. Then, the probability mass function is

$$p(X = x) = \frac{\Gamma(\theta + x)}{\Gamma(\theta)x!} \left(\frac{\theta}{\theta + \mu}\right)^{\theta} \left(\frac{\mu}{\theta + \mu}\right)^{x}.$$

The current count data exhibits overdispersion as shown in Figure 10 indicated by the variance being larger than the mean. For an $n \times m$ count matrix K with row $i = 1, \ldots, n$ for each gene and column $j = 1, \ldots, m$ for each sample, the DESeq2 model is written as the following.

$$\begin{aligned} K_{ij} &\sim NB(\mu_{ij}, \alpha_i) \\ \mu_{ij} &= s_j q_{ij} \\ Var(K_{ij}) &= \mu_{ij} + \alpha_i \mu_{ij}^2 \\ \log_2(q_{ij}) &= \sum_r x_{jr} \beta_{ir}, \end{aligned} \tag{3.1}$$

where μ_{ij} and α_i are the mean and gene-specific dispersion parameters, respectively. x_{jr} and β_{ir} are the design matrix elements and coefficients, respectively. The mean μ_{ij} is modeled as a quantity q_{ij} proportional to the sample-specific normalisation/size factor s_j. Finally, GLM with a logarithmic link function is fitted.

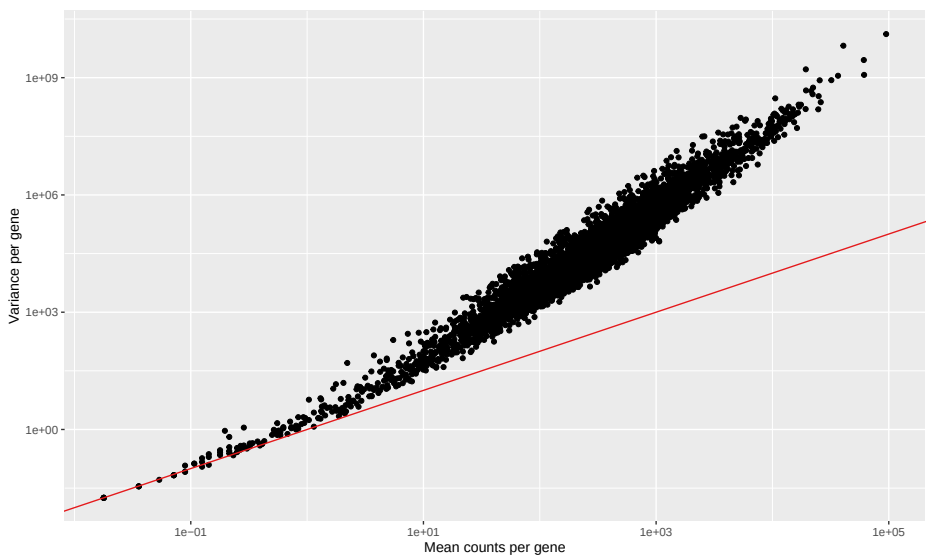

Figure 10: Mean vs variance of gene counts data. Red line indicates mean equal to variance.

DESeq2 tests the null hypothesis that the log 2-fold change (LFC) between treatment and control groups for a certain gene expression is equal to zero, i.e., the treatment effect does not exist for that gene.

Generally, the DESeq2 method consists of four steps, size factor μ_{ij} estimation, dispersion parameter α_j estimation, linear model fitting, and hypothesis testing.

First, the median-of-ratios method is employed to estimate the size factors s_j.

$$\hat{s_j} = median_i \left(\frac{K_{ij}}{(\prod_{\nu=1}^{m} K_{i\nu})^{1/m}} \right)$$

Then, the dispersion parameter α_i is estimated. Dispersion reflects the variability between replicates, hence the natural assumption that each gene would have its own dispersion parameter [12]. DESeq2 dispersion estimate has an inverse relationship with the mean as large spreads in variance estimate can be expected for genes with low mean counts while low spreads in variance estimate are found in high-mean-count genes [7]. DESeq2 assumes that dispersion parameters of genes with similar mean counts are similar to each other.

Gene-specific dispersion values α_i are first estimated by the maximum likelihood estimate (MLE). Then, a smooth curve for the mean of normalised counts against dispersion estimates is fitted to model their dependence. Dispersion estimates are then shrunk towards the curve to obtain the final estimates. This shrinkage would result in a significant reduction of false positives from underestimated dispersion values. Further, shrunken estimates would not be used if the dispersion MLE is above the curve for more than 2 standard deviations to avoid false positives.

GLM with the log link function are then fitted to each gene and Wald tests are conducted to test the significance of the model coefficients. This, however, would lead to the multiple hypothesis testing issue which requires the adjustment of p-values by the Benjamini and Hochberg procedure to control for the false discovery rate [3].

The interaction terms between the GAT201 and experiment design variables are of interest as this report aims to discover the factors that affect the GAT201 effect on gene expression. A large proportion of genes with significant interaction terms would indicate that the corresponding design factors are of high importance and shall be included in subsequent models. Data shall be filtered accordingly based on EDA results to show experiment design factor effects in relevant conditions.

3.2 K-mer Motifs Analysis

A small set of motifs that affect GAT201 can be determined by modelling gene counts as a linear combination of frequencies of k-mer motifs with design factors and identifying the significant interaction terms between GAT201 and the motifs. While directly fitting a GLM to model the gene counts with design factors and k-mer motifs as the features is ideal, this requires tremendous computational power. Therefore, this report proposes a three-stage procedure that comprises LFC computation, motifs reduction, and model fitting.

The idea is to initially reduce the number of motif features by modelling LFC against the k-mers with Lasso regression. A normal approximation is required to employ the Lasso regression and thus, a relatively normal-like LFC, as shown in Figure 11, is chosen instead of the gene counts which follow the negative binomial distribution. With the k-mer motifs features reduced, gene expression is then modeled with GLM. A limitation of this procedure is that it involves model chaining, i.e., each stage utilises estimates from prior stages. This could raise stability issues as there is no simple way to incorporate the standard error of the estimates in subsequent stages.

3.2.1 LFC Computation

To calculate the LFC, a DGE analysis with DESeq2 is conducted by modelling gene counts against GAT201. Mathematically, the model is written as follows.

$$\log_2(q_{ij}) = \beta_{i0} + GAT201_j \times \beta_{i1},$$

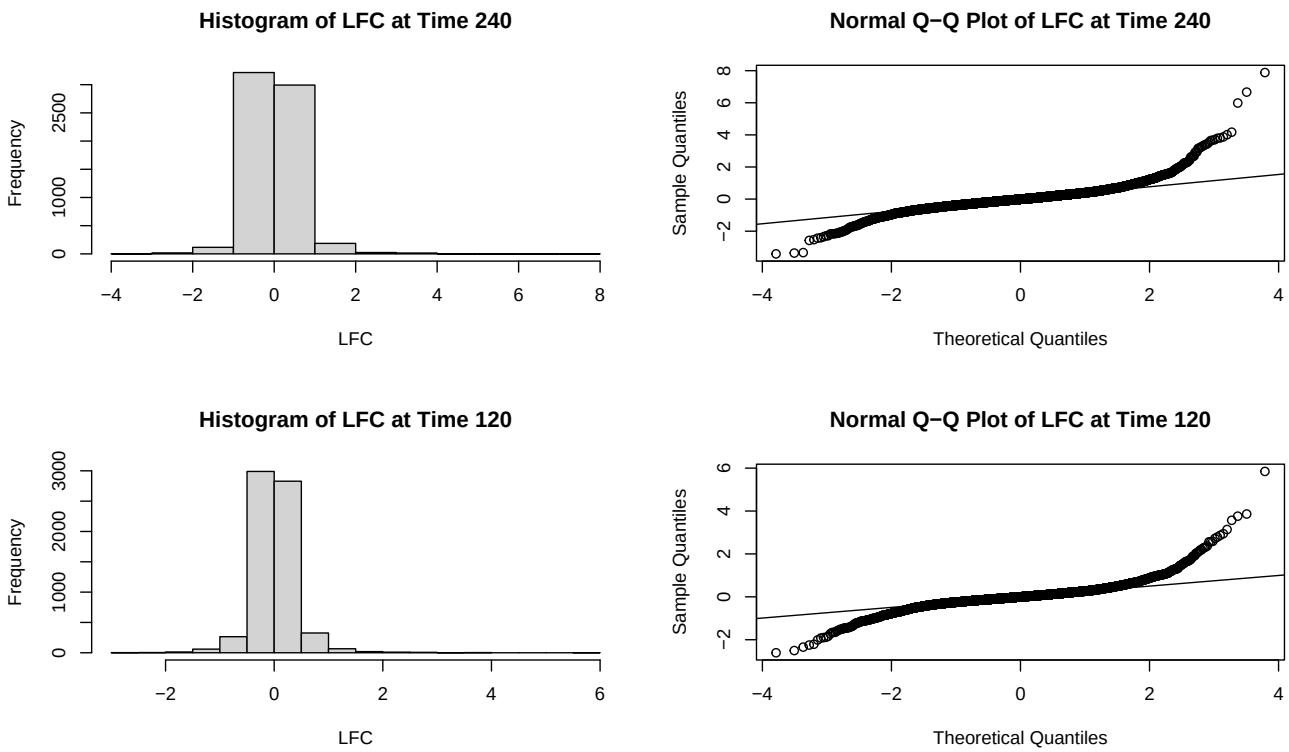

Figure 11: Histograms and Normal QQ plot of LFC at time 240 and 120. Most points in the QQ plots lie on the straight lines with some deviations around the tails shows that the assumption of normal distribution approximation of LFC is appropriate.

where $GAT201$ is a binary variable taking value 1 if GAT201 is present and 0 otherwise. The coefficients of GAT201, β_{i1}, $i = 1, \ldots, n$ for n number of genes which reflect the LFC between groups with GAT201 and groups with GAT201 deleted for each gene. As normal approximation of LFC will be required, only genes with total counts of more than or equal to 100 across all samples are included in the analysis.

While it is difficult to incorporate standard errors of the LFC estimate into the next stage, a crude stability check of the results could be done. Noting that the marginal distribution of the LFC asymptotically follows a normal distribution with mean of the point estimate of the LFC and standard deviation of the standard error of the estimated LFC, i.e., $\beta_{i1} \sim N(\hat{\beta}_{i1}, SE(\hat{\beta}_{i1}))$, we could simulate more LFC estimates and check whether using the estimated LFC would produce similar results in the next stage compared to if we were to use the simulated LFC.

3.2.2 Lasso Regression for Motifs Reduction

Lasso regressions are run with the estimated LFC as the response variable and k-mer motifs as the explanatory variable. The coefficient of a Lasso regression is obtained by optimising the following loss function.

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \mathcal{L}(y_i, f(X_{i.})) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{3.2}$$

where y_i is the (estimated) LFC, X is the training matrix, $\mathcal{L}$ is the logistic loss function, and λ is the regularisation parameter optimised by cross-validation. The best λ is chosen such that loss is minimum. Due to the nature of Lasso penalty, irrelevant features will be shrunk to zero which makes it ideal for motifs reduction. Shrunken motifs will be excluded when modelling the gene expression in the next stage.

It is worth noting, however, that under high correlation between features, Inference of Lasso regression may be unreliable due to the inflated standard error from multicollinearity [18]. This is a common issue with k-mer-based representation of genes when all 2, 3, 4, and 5-mer motif data are combined as shown in 9 [10]. Therefore, interpretations of results must be done carefully.

As a result stability check, Lasso regressions can be fitted with the simulated LFC values instead of the estimated LFC values. Results are said to be stable if the motifs that are not chosen are also not chosen in around $\geq 50\%$ or more simulated cases.

3.2.3 Generalised Linear Model

In the final stage, construct a GLM with the log link function assuming that the gene counts follow NB distributions. Let K_{ij} be the gene counts for gene i and sample j. Further, let m_j be the median count for sample j. Then,

$$\begin{aligned} K_{ij} &\sim NB(\mu_{ij}, \theta) \\ \log(\mu_{ij}) &= \sum_{r} x_{jr}\beta_{jr} + \log(m_j), \end{aligned} \tag{3.3}$$

where x_{jr} and β_{jr} are the features and corresponding coefficients, respectively. The model implies that all genes share a common dispersion parameter θ. Coefficient of interests are the interaction terms between k-mer motifs and GAT201. For each 2, 3, 4, and 5-mer motifs data, separate models will be fitted. A single model with all k-mer motifs data will also be fitted to obtain the smallest set of motifs that significantly affect GAT201 in explaining gene expression.

3.2.4 K-mer Motif Clustering

To address the multicollinearity issues in highly correlated features when all k-mer motifs are considered, a clustering technique shall be introduced. The method employed in this report is a modified clustering

approach proposed in [10] to work with k-mer-based representation in machine learning frameworks. The method comprises three steps, which are the screening, clustering, and learning steps.

In the screening step, Lasso regression is conducted with LFC as the response variable and k-mer motifs as the explanatory variables. Then, define a set of active features, a, as the set containing features with non-zero coefficients. That is,

$$a = \{i \in [1, \ldots, p], \text{ such that } |\beta_i| > 0\},$$

where β is a $p \times 1$ vector of the model coefficients and let $p_a = |a|$ be the number of coefficients in a. Then, construct a $p_a \times p$ correlation matrix G between the active features and original features. Finally, define the final set of features e containing features that are strongly related to at least one of the active features. That is,

$$e = \{i \in [1, \ldots, p], \text{ such that } \max(G_{.,i}) > s_1\}, \tag{3.4}$$

where s_1 is the screening threshold hyperparameter that defines strong correlation.

In the clustering step, the final selected features are clustered by k-means clustering due to its simplicity. the optimal number of clusters is chosen by the elbow method according to the decrease in total within the sum of square error.

Finally, in the learning step, each cluster is identified as the new features defined as the mean of the motifs frequencies in the corresponding cluster. With the significant reduction in the number of features and highly correlated features are clustered together, GLM model is constructed and steps in Section 3.2.3 are repeated with the clusters as the features instead of the motifs to obtain the significant interaction terms between the clusters and GAT201.

Model comparison will then be conducted to identify model that performs the best. AIC criteria which measures the balance between model fit and complexity will be used as models are non-nested.

4 Results and Discussions

This section is divided into two subsections where Section 4.1 addresses the first research question while Section 4.2 addresses the second research question. All models were implemented in R with DESeq2 package for differential gene analysis.

4.1 Significant Experiment Design

There are two experiment designs to be investigated, time and growth media. We are interested in examining the interaction terms between experiment design and GAT201 to identify factors that can influence the GAT201 effect in gene expression and uncover its behaviour in various environments. Nevertheless, the main effect and total effect will also be analysed as they will be relevant in building the model to address the second research question. Genes with counts less than 10 across all samples are excluded from the analysis to improve stability of results.

4.1.1 Effect of Growth Media

Three models are constructed to assess the effect of growth media, one for each timepoint 30, 120, and 240. Cells sampled at time 0 are always sampled from the YPD media and thus, media effect at this timepoint is meaningless to be examined. Following the notations in equation (3.1), for each time point, the model is as follows.

$$\log_2(q_{ij}) = \beta_{i0} + GAT201_j \times \beta_{i1} + Condition_j \times \beta_{i2} + GAT201 \times Condition_j \times \beta_{i3},$$

where $GAT201$ is a binary variable taking value 1 if GA201 is present and 0 otherwise. Further, $Condition$ is a binary variable taking value 1 for growth media RS and 0 for growth media R.

The results for the effect of growth media are summarised in Table 1 which shows the percentage of genes with significant adjusted p-values between GAT201 and growth media at different time points. The Factor column shows the features, the Effect column shows the type of effect measured by the corresponding factor, and the Time column shows the time point of analysis. LFC Up and Down show the proportion of genes with significant coefficient indicated by their adjusted p-values being ≤ 0.05. As the coefficient measures the LFC effect between the reference and non-reference group, LFC Up shows the proportion of genes that are up-regulated, i.e., genes with positive coefficients, while LFC Down shows the proportion of genes that are low-regulated, i.e., genes with negative coefficients. The Total column shows the total proportion of significant genes.

Factor	**Effect**	**Time**	**LFC Up**	**LFC Down**	**Total**
Condition$\times GAT201$	Interaction	30	0.00%	0.00%	0.00%
Condition$\times GAT201$	Interaction	120	0.05%	0.15%	0.20%
Condition$\times GAT201$	Interaction	240	0.39%	0.36%	0.75%
Condition	Main	30	1.60%	4.60%	6.20%
Condition	Main	120	1.00%	0.39%	1.39%
Condition	Main	240	5.20%	1.90%	7.10%
Condition$\times GAT201 + Condition$	Total	30	0.12%	0.74%	0.86%
Condition$\times GAT201 + Condition$	Total	120	12.00%	9.50%	21.50%
Condition$\times GAT201 + Condition$	Total	240	10.00%	8.80%	18.80%

Table 1: Proportion of genes with adjusted p-values ≤ 0.05 for all growth media effect coefficient terms at different time points

Results show that the interaction terms are not statistically significant for most of the genes, with the total proportions of genes with significant interaction terms being less than 1% at all time points. This implies that growth media does not significantly affect GAT201 effect for most genes. This result is consistent with previous research where activation of the GAT201 transcription factor is serum-independent [9]. However, this could be attributed to the short 4-hour time course as other research has previously concluded that serum could induce capsule growth of *Cryptococcus neoformans* [19].

While the GAT201 effect may be serum-independent, results show that growth media have a significant effect in modelling gene expression as indicated by the relatively high proportion of genes with significant main and total effects. The total effects at time 120 and 240, for example, are particularly high with 21.50% and 18.80% of the genes showing significant coefficients, respectively. these results are inline with the EDA result which shows some clustering based on growth media at some timepoints. Therefore, the growth media effect should be taken into account when constructing a linear model for gene expression.

4.1.2 Effect of Time

The following model is constructed to investigate the effect of time.

$$\log_2(q_{ij}) = \beta_{i0} + GAT201_j \times \beta_{i1} + \sum_{\substack{t=\{30,120,240\} \\ r=\{2,3,4\}}} Time(t) \times \beta_{ir} + \sum_{\substack{t=\{30,120,240\} \\ r=\{5,6,7\}}} Time(t) \times GAT201 \times \beta_{ir}, \quad (4.1)$$

where $Time(t)$ are binary variables taking the value 1 if $Time$ is at time point t and 0 otherwise. Time 0 is considered to be the starting point and it is assumed that no reaction has occurred which makes it the reference time point.

Table 2 summarises the effects of time. The columns can be interpreted in the same way as in Table 1. Results show that the GAT201 effect differs significantly at different timepoints according to the relatively

Factor	Effect	Time	LFC Up	LFC Down	Total
GAT201×$Time(30)$	Interaction	30	3.40%	4.70%	8.10%
GAT201×$Time(120)$	Interaction	120	4.80%	3.70%	8.50%
GAT201×$Time(240)$	Interaction	240	7.60%	8.80%	16.40%
Time(30)	Main	30	41.00%	42.00%	83.00%
Time(120)	Main	120	40.00%	40.00%	80.00%
Time(240)	Main	240	41.00%	40.00%	81.00%
GAT201×$Time(30) + Time(30)$	Total	30	40.00%	43.00%	83.00%
GAT201×$Time(120) + Time(120)$	Total	120	40.00%	40.00%	80.00%
GAT201×$Time(240) + Time(240)$	Total	240	40.00%	38.00%	78.00%

Table 2: Proportion of genes with adjusted p-values ≤ 0.05 for all time effect coefficient terms

high proportions of genes with significant interaction terms at all time points. An upward trend in the number of genes influenced by GAT201 is indicated by the rising number of significant interaction terms as time passes starting at 8.10% at time 30, rising to 8.50% at time 120, and finally peaking at 16.40% at time 240. Therefore, it is concluded that activation of GAT201 effect is time-sensitive.

An extremely high proportion of genes with significant main and total effects of time can be observed. For all timepoints, the main and total effects are all around 80%. This indicates a strong effect of time for most genes in modelling gene expression. PCA in Section 2.1 also reached a similar conclusion where a clear separation between timepoints can be seen. This indicates that Time dominates the variation of gene expression and thus should be properly addressed in gene expression modelling.

4.2 Significant K-mer Motifs

To efficiently conduct k-mer motif analysis, DGE analysis is conducted to obtain the timepoint where GAT201 effect is the most prominent. Utilising the model defined in equation (4.1), the total GAT201 effect is calculated for all timepoints. Results in Table 3 show that the GAT201 effect is the strongest at timepoint 240 with 40% of the genes showing a significant total GAT201 effect. Therefore, the k-mer motif analysis shall be conducted at time 240 and one timepoint before that, timepoint 120, to obtain stable results.

Factor	Effect	Time	LFC Up	LFC Down	Total
GAT201×$Time(30) + GAT201$	Total	30	11.00%	12.00%	23.00%
GAT201×$Time(120) + GAT201$	Total	120	9.10%	7.50%	16.60%
GAT201×$Time(240) + GAT201$	Total	240	18.00%	22.00%	40.00%

Table 3: Proportion of genes with adjusted p-values ≤ 0.05 for GAT201 effect coefficient terms

Firstly, 2, 3, 4, and 5-mer models shall be considered separately. Following the three-stage procedure described in Section 3.2, LFC for all genes at times 120 and 240 are calculated through DGE analysis of gene expression regressed against GAT201. Fifty Simulated LFC at both times are generated for results stability check.

Then, Lasso regressions with the estimated LFC as the response variable and k-mer motifs as the features are run for the initial screening of k-mer motifs. Stability check is also done to ensure result stability as multiple model chaining is involved. For convenience, the proportion that a motif is not chosen is investigated instead of the proportion that a motif is chosen as they are equivalent. Figure 12 and 13 illustrate the proportion that non-chosen motifs are not chosen in the simulated cases. It can be

observed that for both time points, all motifs that are not chosen are also not chosen for around 50% or more in the simulated case. At time point 240, the lowest proportions of 2, 3, 4, and 5-mer motifs that are not chosen are 100% for AA, 46% for GCA, 48% for TGAA, and 66% for ACGGC, respectively. At time point 120, the lowest proportions are 74% for CG, 100% for AAA, 62% for ACCG, and 52% for GTAAC. Therefore, since the minimum proportion for all non-chosen k-mer motifs is around $\geq 50\%$, it is concluded that the obtained results are relatively stable.

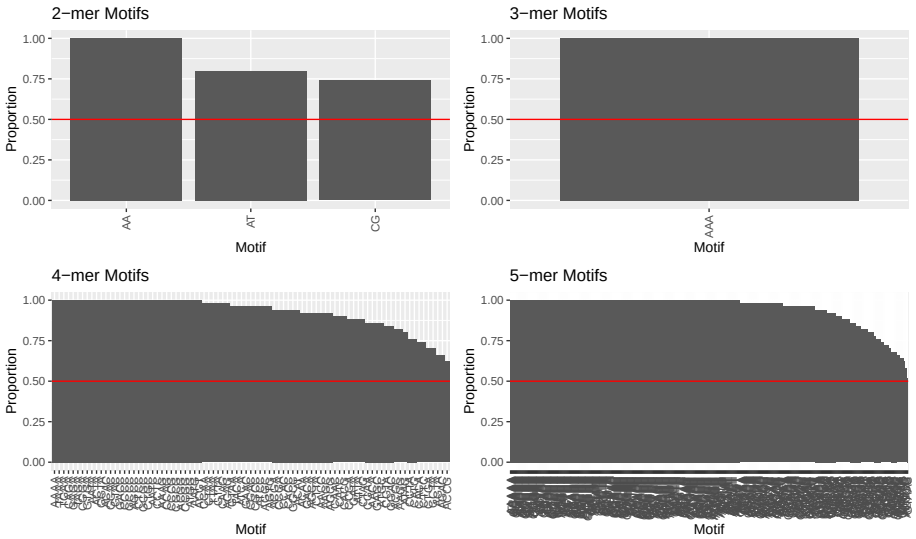

Figure 12: Proportion of non-chosen motifs are not chosen in fifty simulated cases at timepoint 120 for 2-mer (top left), 3-mer (top right), 4-mer (bottom left), and 5-mer (bottom right) models

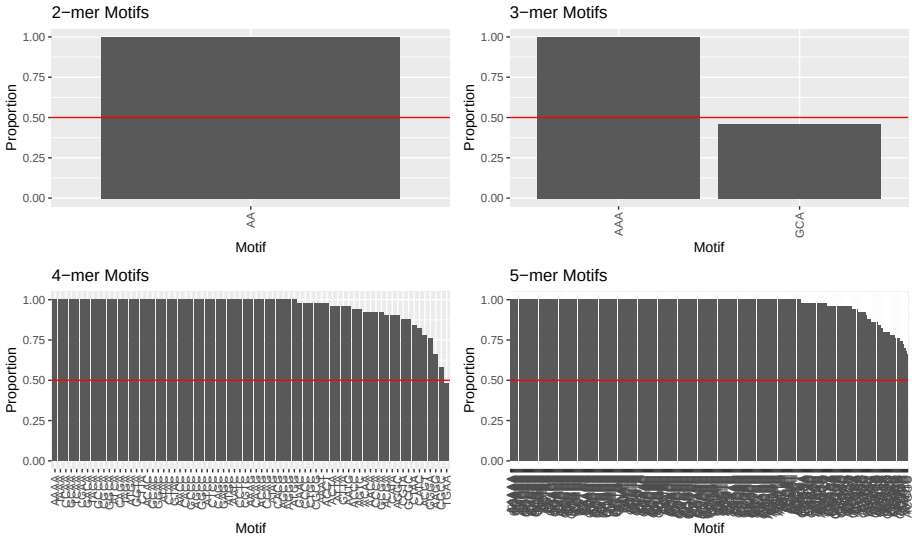

Figure 13: Proportion of non-chosen motifs are not chosen in fifty simulated cases at timepoint 240 for 2-mer (top left), 3-mer (top right), 4-mer (bottom left), and 5-mer (bottom right) models

Finally, GLM models are constructed for both time points with gene counts as the response variable. From Section 4.1, it is known that while the interaction of GAT201 and growth media is not significant for most genes, the total gene effect is still present in modelling gene counts. Furthermore, analysing data at a specific time point has addressed the time effect. Therefore, the explanatory variables for the

GLM models are growth media, GAT201, k-mer motifs, and the interaction between each k-mer motif and GAT201. Adopting the same notation in equation (3.3), the model can be formally written as the following.

$$K_{ij} \sim NB(\mu_{ij}, \theta)$$
$$\log(\mu_{ij}) = \beta_{0i} + GAT201_j \times \beta_{i1} + Conditon_j \times \beta_{i2} + \sum_r x_{ir}\beta_{ir} + \sum_z GAT201 \times x_{iz} \times \beta_{iz} + \log(m_j),$$

where x are the corresponding motifs.

Table 4 shows the motifs where the interaction between them and GAT201 is significant. The Time 240 and 120 column show the motifs which has significant interaction terms with GAT201 only at time 240 and 120, respectively. Similarly, the Time 240 and 120 column shows the motifs with significant interaction terms with GAT201 at both time 240 and 120. Motifs are sorted according to p-value in increasing order, e.g., the interaction term 5-mer motif CTGAC×GAT201 is significant with the smallest p-value while AAGTG×GAT201 is significant with the largest p-value. Figure 14 illustrates the number of significant interaction terms for each k-mer model at timepoints 120 and 240.

K-mer	Time 240	Time 120	Time 240 and 120
2-mer	AT, TA, AC, CA, AG, CC, GA		
3-mer	AAT, TAA, CAA, GAA, AAG, AAC, ACT		
4-mer	GTCA, GCTA, AATT, ACGC, GATC, CCTA, AACG, CCCG, AGAC, CCGC, AGCT, ACCG, CCGG, CCGA, ATCG, CGGC, GTGA, AATA, CGGA, ACGG, TGCA, ACTG, CAAC, TTAA, AGTA, ATTG, AAAC, CTAG, CGAG		AGAT
5-mer	CTGAC, CGATA, CCAAG, ATACG, ATCTG, AGATA, GATAA, GATCA, ATATC, TATCA, CTGCC, CCCAA, AAAGC, CAGGG, ACACT, CCTGA, GGTCA, GACTA, AGGAA, CAAAG, GTCAA, ACAGC, GATAC, AATTA, CCGTC, CTGTA, GACGC, GATGA, CAACA, CGGGC, GAAAA, CCGAA, CTTTA, GTCGA, CGGTA, AAGTG	TACTA, AGCCA, AACCC, ATCTA, AAATC, GAATA, AGATG, AATTG, TGTAA, AGTTG, ACAAT, CCCTC, AGAAT, CACAC, GCGAA, ATGTA, ATTGG, ACTGC	AGATC, ATTGC, CAATA, AGGGA, GTTTA

Table 4: Full results of motifs with significant interaction term between the motifs and GAT201

To choose the motifs that significantly affect the GAT201 effect the most, it is believed that the motifs with interaction terms that are significant in both timepoints 240 and 120 are the most influential (i.e., the intersection). This is followed by the motifs with interaction terms that are significant in time 240 as the GAT201 effect is most prominent at time 240, thus making its results more reliable than the results

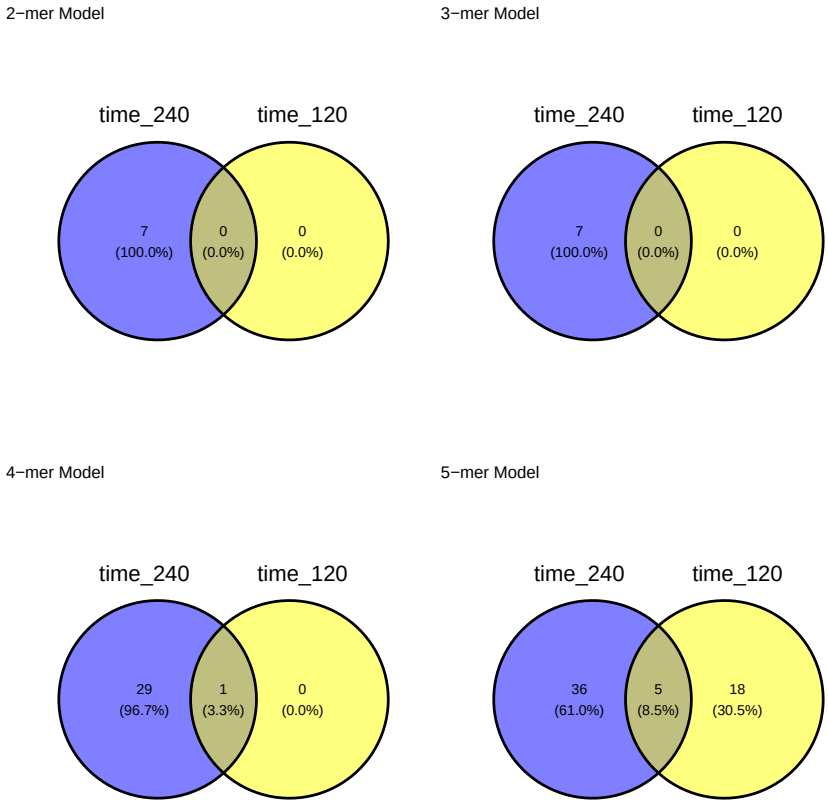

Figure 14: Venn diagram indicating the count and percentage of 2-mer (top left), 3-mer (top right), 4-mer (bottom left), and 5-mer (bottom right) motifs that have significant interaction term with GAT201 at timepoint 120, 240, and the intersection between them

at timepoint 120. The chosen top 5 most influential motifs on GAT201 effect are as follows.

1. 2-mer motifs: AT, TA, AC, CA, AG
2. 3-mer motifs: AAT, TAA, CAA, GAA, AAG
3. 4-mer motifs: AGAT, GTCA, GCTA, AATT, ACGC
4. 5-mer motifs: AGATC, ATTGC, CAATA, AGGGA, GTTTA

As an extension, we also consider 2, 3, 4, and 5-mer motifs at time 240 altogether to model gene expression. Motif clustering shall be employed to alleviate the multicolinearity issue from shared information. The active feature set a and final feature set e are first obtained with screening threshold s_1 set to 0.5. Figure 15 shows the total within the sum of square error for k number of clusters. By the elbow method, the figure implies that clustering the motifs into 4 clusters would give the optimal result. Figure 16 illustrates the PCA plot of the chosen k-mer motifs. It shows that the first PC could explain most of the variations in the data (89.3%) and clustering the motifs into 4 clusters seems appropriate. The first cluster has 174 motifs, the second cluster has 23 motifs, the third cluster has 2 motifs, and the fourth cluster has 79 motifs. Table 5 shows the full list of motifs in each cluster.

Finally, GLM model is constructed to identify the significant interaction terms between the clusters and GAT201. Results show that interaction terms between GAT201 and cluster 1, 3, and 4 are significant.

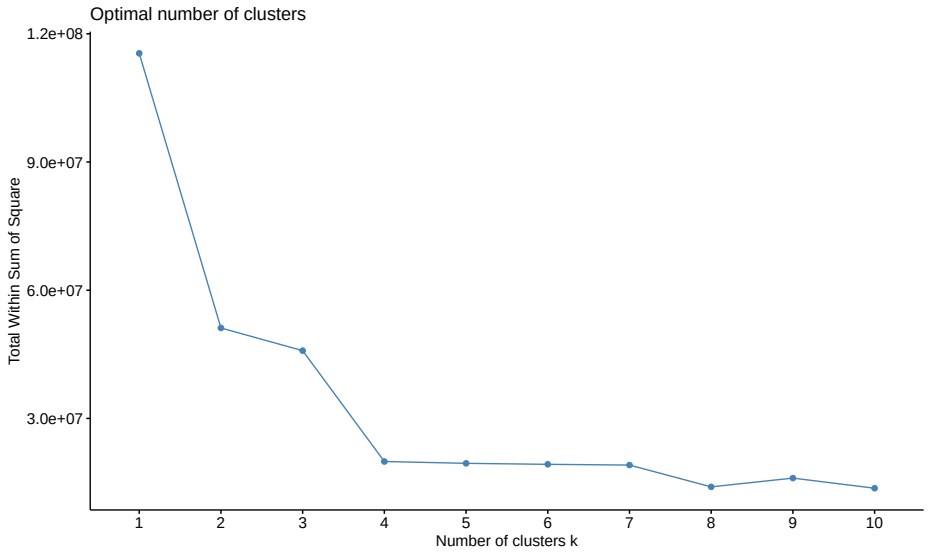

Figure 15: Total within sum of square error for various number of clusters

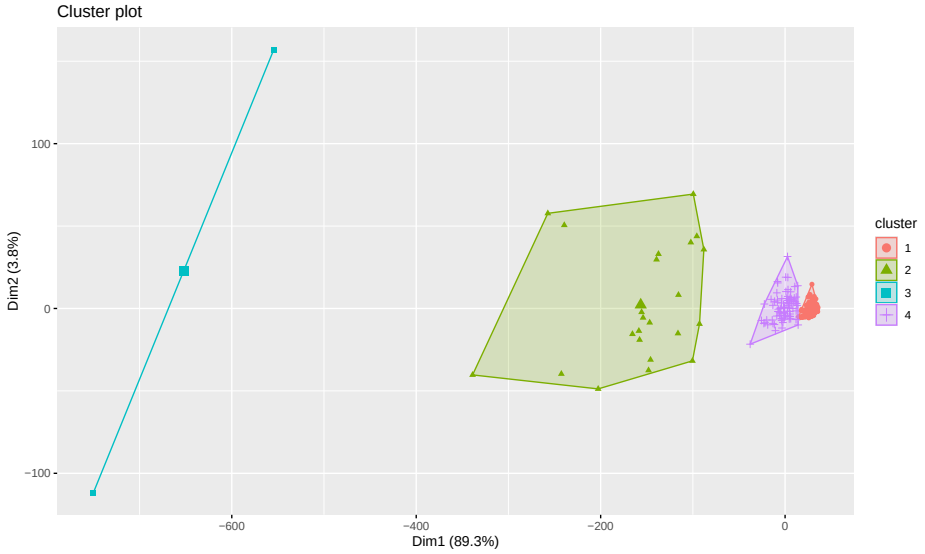

Figure 16: PCA plot of k-mer motifs clustered into 4 clusters

17, 18, 19, 20, and 21 illustrate the log-transformed p-values of the interaction terms in the 2-mer, 3-mer, 4-mer, 5-mer, and cluster model, respectively. Circled points are the top 5 most influential motifs. Note that in most models, motifs with the most significant interaction terms are chosen except in the 5-mer model.

Table 6 shows the coefficients, standard errors, and p-values of the interaction terms for the top motifs in each model. The $\log_2$ transformation on the sample concentration q_j implies that the interaction effect estimate β_{ij} can be interpreted as a $2^{\beta_{ij}}$ FC in the difference of GAT201 effect for a unit change in the frequency of the corresponding motif [6]. Therefore, the magnitude of the interaction term coefficient estimates corresponds to the strength of influence that the motif has on GAT201 effect.

Residual analyses for all models at timepoint 240 are conducted to assess their fit (see Figure 22, 23, 24, 25, 26 for residual plots of 2-mer, 3-mer, 4-mer, 5-mer, and cluster model, respectively). The goodness-of-fit for all models are relatively similar; not ideal but acceptable. Linearity and homoskedasticity seem to be violated by non-random scatters and uneven spread in residuals vs fitted and scale-location plots,

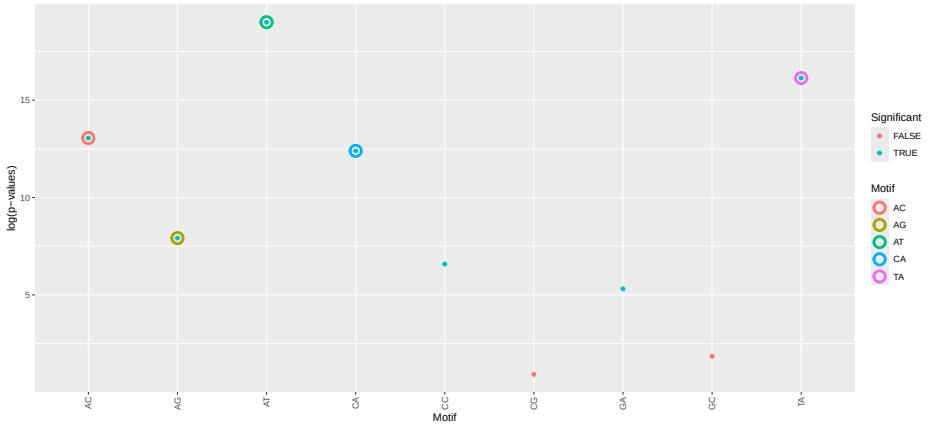

Figure 17: Log transformed p-values of the interaction terms between 2-mer motifs and GAT201

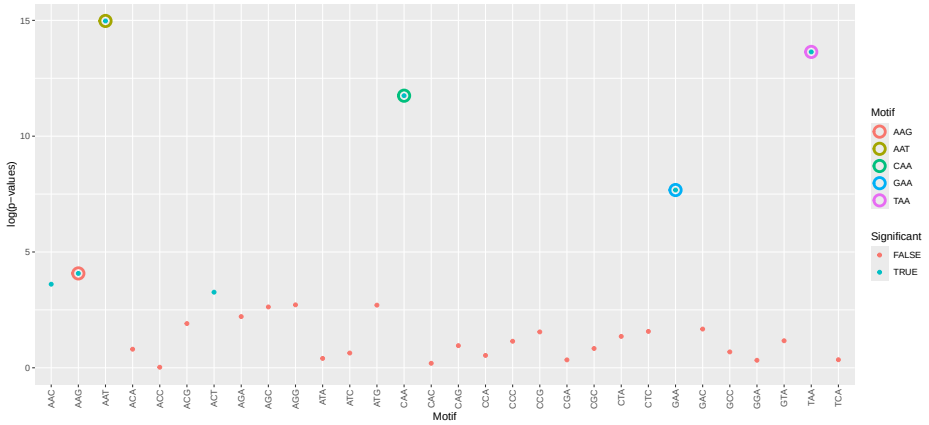

Figure 18: Log transformed p-values of the interaction terms between 3-mer motifs and GAT201

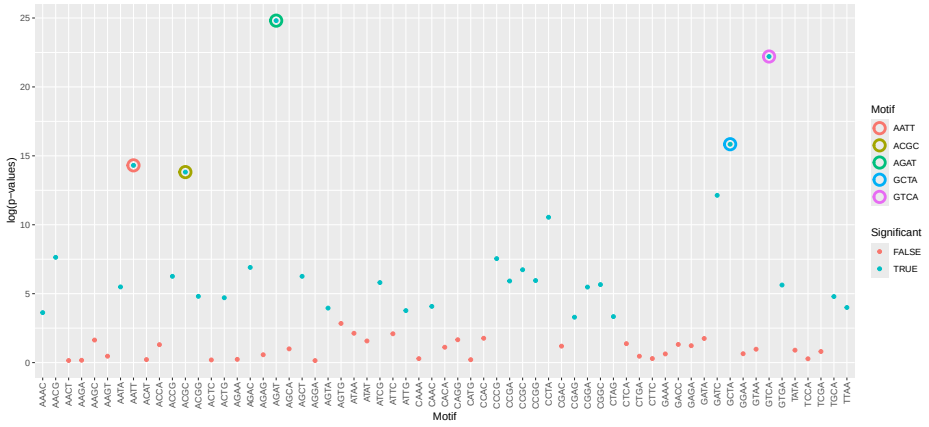

Figure 19: Log transformed p-values of the interaction terms between 4-mer motifs and GAT201

respectively. Further, points in QQ-plot form a straight line from the left tail but deviate heavily on the right tail. Additionally, no extreme outlier was detected in the data. The weak goodness-of-fit in all models might be due to the assumption of common dispersion parameter among all genes which may be violated as seen from previous DEG analysis which requires independent models for each gene.

Comparison of AIC between models shows that the cluster model has the highest AIC of 1577305

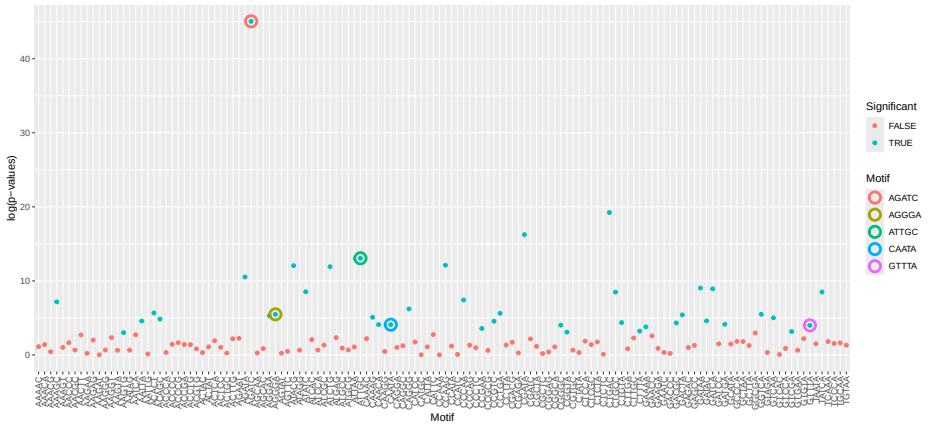

Figure 20: Log transformed p-values of the interaction terms between 5-mer motifs and GAT201

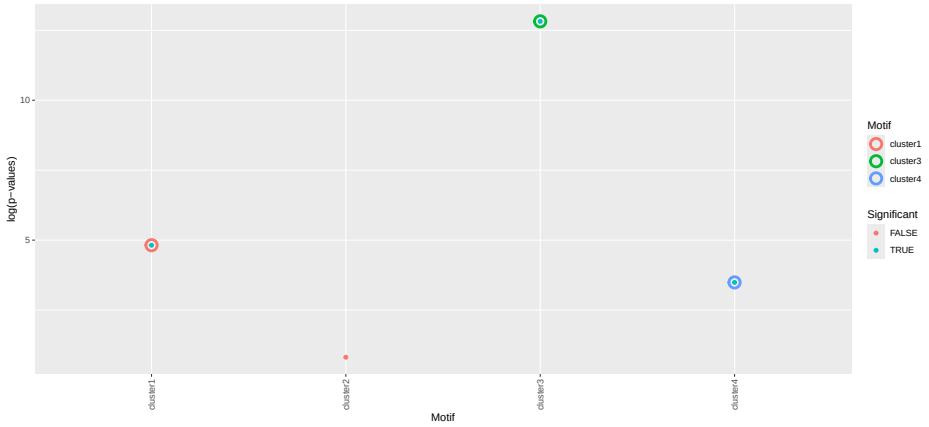

Figure 21: Log transformed p-values of the interaction terms between clustered motifs and GAT201

while the 4-mer model with the lowest AIC of 1561122 (see Table 7). These results suggest that the more parsimonious 4-mer model is the optimal model.

The underwhelming performance of the cluster model suggests room for improvement. The simplified approach of clustering may be the cause of the poor performance. Nevertheless, one clear advantage of the cluster model is it requires the least amount of computational power due to the significant reduction in the feature space.

5 Conclusions

The objective of this report is to identify design factors and k-mer motifs that could significantly influence GAT201 factors in gene expression. DEG analysis showed that growth media does not significantly affect GAT201 for most genes as the activation of GAT201 is serum-independent with only less than 1% of the genes showing significant interaction terms at all timepoints. Time, however, significantly affects GAT201 for many genes with 240 minutes being the most influential timepoint with 16.4% of the genes revealing significant interaction effects. Therefore, the Wallace lab could consider studying the GAT201 effect at new timepoints and, for example, find out the timepoint that has a peak influence on GAT201. They should also consider running the experiment in a longer timeframe as four hours may not have been enough for the media to exert its influence on GAT201.

Modelling gene expression with k-mer motifs when the GAT201 effect is most prominent revealed the motifs that could significantly affect GAT201. The top 5 most influential motifs are: the 2-mer AT, TA,

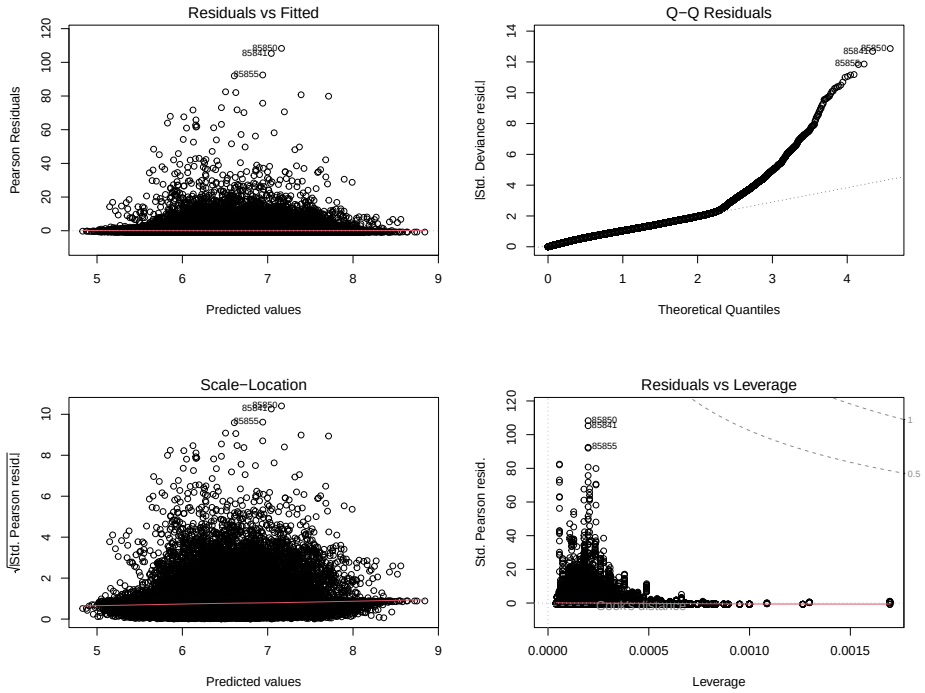

Figure 22: Residual Plots for 2-mer model

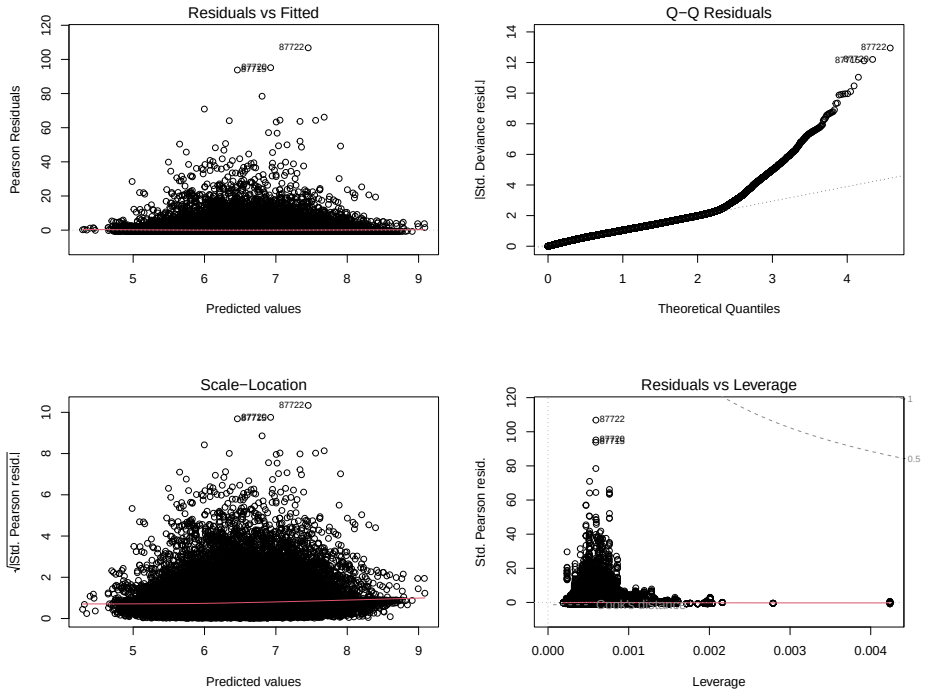

Figure 23: Residual Plots for 3-mer model

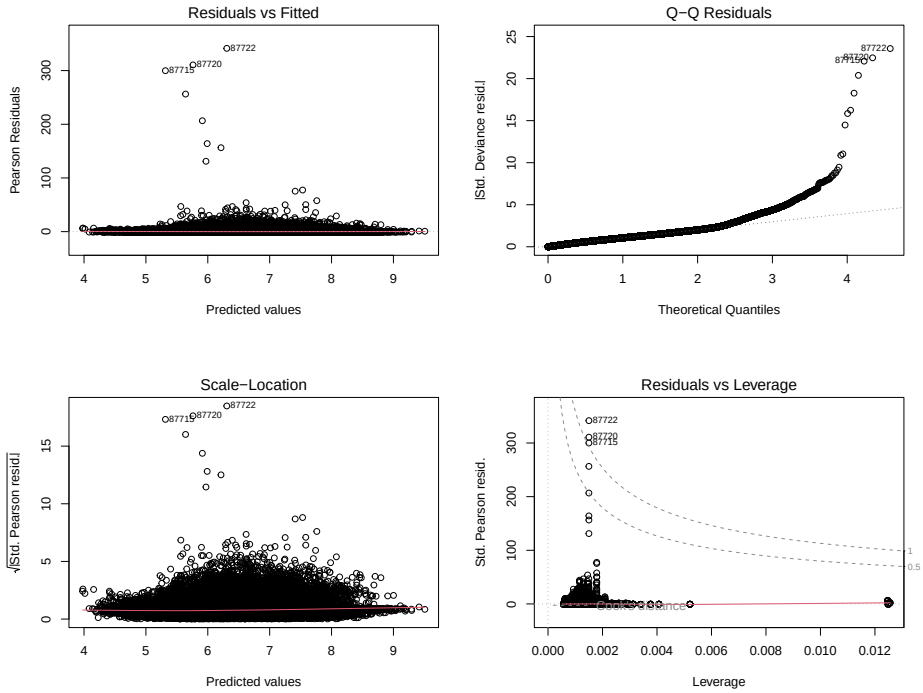

Figure 24: Residual Plots for 4-mer model

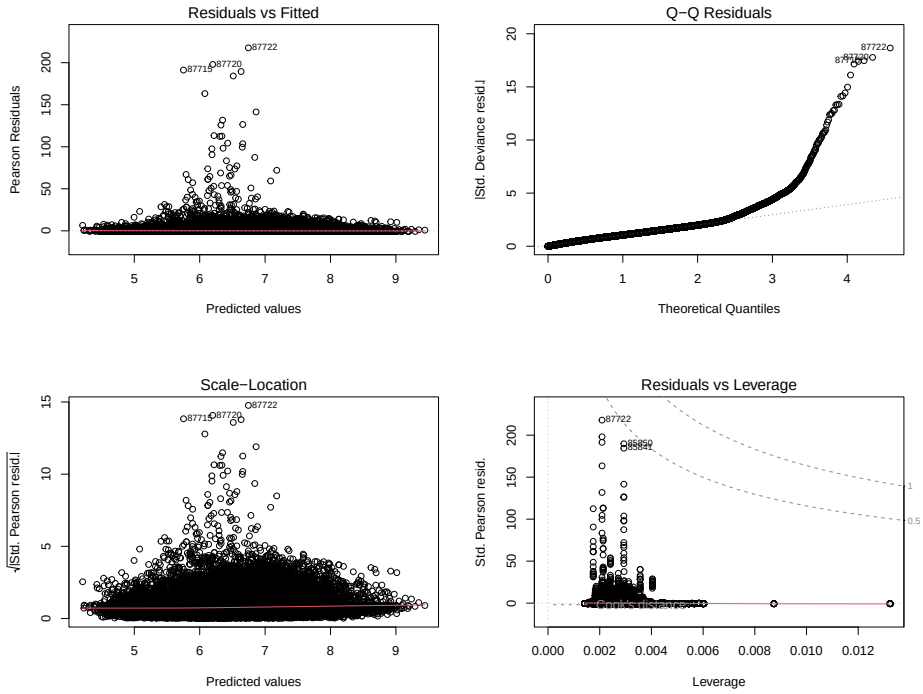

Figure 25: Residual Plots for 5-mer model

AC, CA, and AG motifs; the 3-mer AAT, TAA, CAA, GAA, and AAG motifs; the 4-mer AGAT, GTCA, GCTA, AATT, ACGC motifs; and the 5-mer AGATC, ATTGC, CAATA, AGGGA, and GTTTA motifs. Motif clustering of all motifs was employed to tackle multicollinearity of the k-mer features. Motifs were clustered into four clusters (see Table 5) and GLM analysis showed that clusters 1, 3 and 4 have the most significant interaction terms with GAT201. Model comparison showed that the 4-mer model was the optimal model while the cluster model underperformed compared to the other models. Hopefully, these results could help the Wallace lab to genetically engineer cells and concentrate on the top influential motifs to learn more about the GAT201 effect on gene expression.

There are, however, some limitations to the presented methods. Firstly, DESeq2 assumes an inverse relationship between dispersion estimates and the mean counts which could be found lacking in the current data. This may result in less accurate dispersion estimates and DGE analysis results. Additionally, the employed motif selection method involves model chaining without a rigorous way of addressing the uncertainty of the estimates. Ignoring standard errors of estimate in model chaining may ensue propagation of errors in subsequent stages which could lead to misleading statistical inference.

Future research could investigate more advanced machine-learning techniques for modelling gene expression with the k-mer motifs. Algorithms such as interpretable random forests and neural networks are more flexible in dealing with high-dimensional motifs data. More refined motif clustering methods could also be investigated to further improve its performance, for example, through cross-validation of hyperparameters.

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
CTAA, TTAA, TGCA, TCGA, TATA, ACTA, GTAC, GATC, CGCG, CAAAA, GAAAA, ACAAA, TCAAA, AGAAA, GGAAA, AACAA, CCCAA, AGCAA, TGCAA, GTCAA, AAGAA, CCGAA, AGGAA, GGGAA, TGGAA, GATAA, TATAA, GCTAA, TGTAA, ATTAA, AAACA, CAACA, AGACA, ACCCA, GCCCA, ATCCA, GTCCA, CCGCA, GATCA, TATCA, ACTCA, GGTCA, AAAGA, GAAGA, AGAGA, AACGA, ACCGA, GCCGA, TCCGA, ATCGA, GTCGA, CAGGA, GCGGA, AGGGA, ATGGA, ACTGA, CCTGA, AGTGA, ATTGA, CTTGA, CAATA, TAATA, ACATA, AGATA, CGATA, ATATA, CTATA, GTATA, GACTA, CAGTA, CGGTA, ATGTA, CTGTA, GTGTA, AATTA, CTTTA, AAAAC, CAAAC, ACAAC, AGCAC, ATCAC, AAGAC, GAGAC, CCGAC, CTGAC, GATAC, GAACC, AACCC, GACCC, ACCCC, CCCCC, CGCCC, CTCCC, GAGCC, CCGCC, CGGCC, CTGCC, CATCC, ACTCC, CCTCC, CTTCC, AAAGC, ACAGC, ACCGC, CCCGC, GCCGC, CTCGC, CAGGC, CCGGC, CGGGC, ACTGC, CCTGC, ATTGC, CTTGC, CAATC, ACATC, CCATC, AGATC, CGATC, ATATC, CACTC, CCCTC, CGCTC, ATCTC, CTCTC, CAGTC, CCGTC, CGGTC, ATGTC, CGTTC, CTTTC, AAAAG, CAAAG, CCCAG, AAGAG, AATAG, AAACG, CAACG, AGACG, CGACG, CACCG, ACCCG, CCCCG, CGCCG, CCGCG, AATCG, CATCG, ACCGG, AGCGG, ATCGG, AAGGG, AGGGG, AGTGG, ACATG, AGATG, ATATG, AACTG, ACCTG, ATCTG, AAGTG, AATTG, ATTTG, ACAAT, AGAAT, ATAAT, AACAT, AAGAT, AATAT, ACTAT, AGTAT, ACACT, ACAGT, AAGGT, AACTT	TA, GC, CG, AT, AAA, CAA, GAA, TAA, AGA, CGA, GGA, ATA, GTA, AAC, CCC, GCC, CGC, ATC, CTC, AAG, CCG, ATG, AAT	AA, CC	AAAA, CAAA, GAAA, CCAA, GCAA, TCAA, AGAA, GGAA, GTAA, AACA, TACA, CCCA, TCCA, AGCA, GGCA, ATCA, GTCA, AAGA, CAGA, GAGA, CCGA, GCGA, AGGA, GTGA, AATA, GATA, AGTA, CGTA, ATTA, AAAC, CAAC, GCAC, AGAC, CGAC, GGAC, ATAC, GACC, ACCC, CCCC, GCCC, CGCC, ATCC, CTCC, AAGC, CAGC, GAGC, CCGC, CGGC, CATC, ACTC, CCTC, CGTC, ATTC, AAAG, CAAG, ACAG, CCAG, AGAG, CGAG, ATAG, AACG, ACCG, CCCG, AGCG, ATCG, AAGG, CAGG, ACGG, AGGG, ACTG, AGTG, ATTG, ACAT, AGAT, ATAT, AACT, ACCT, AAGT, AAAAA

Table 5: List of motifs clustered into four clusters

Model	Variable	Estimate	Std. Error	p-value
2-mer	AT×GAT201	0.093	0.016	0.000
	TA×GAT201	-0.085	0.016	0.000
	AC×GAT201	0.052	0.011	0.000
	CA×GAT201	-0.050	0.011	0.000
	AG×GAT201	0.041	0.012	0.000
3-mer	AAT×GAT201	0.072	0.014	0.000
	TAA×GAT201	-0.070	0.014	0.000
	CAA×GAT201	-0.065	0.015	0.000
	GAA×GAT201	-0.073	0.021	0.000
	AAG×GAT201	0.051	0.021	0.017
4-mer	AGAT×GAT201	0.032	0.005	0.000
	GTCA×GAT201	-0.028	0.004	0.000
	GCTA×GAT201	0.028	0.005	0.000
	AATT×GAT201	-0.031	0.006	0.000
	ACGC×GAT201	0.021	0.004	0.000
5-mer	AGATC×GAT201	0.076	0.008	0.000
	ATTGC×GAT201	0.044	0.009	0.000
	CAATA×GAT201	0.020	0.008	0.017
	AGGGA×GAT201	-0.026	0.009	0.004
	GTTTA×GAT201	-0.023	0.010	0.018
Cluster	cluster1×GAT201	0.281	0.106	0.008
	cluster3×GAT201	-0.006	0.001	0.000
	cluster4×GAT201	0.112	0.052	0.031

Table 6: Estimated coefficients, standard errors, and p-values of the interaction between top 5 most influential motifs and GAT201

Model	AIC
Cluster	1577305
2-mer	1572308
3-mer	1564416
4-mer	1561122
5-mer	1564772

Table 7: AIC of GLM Models

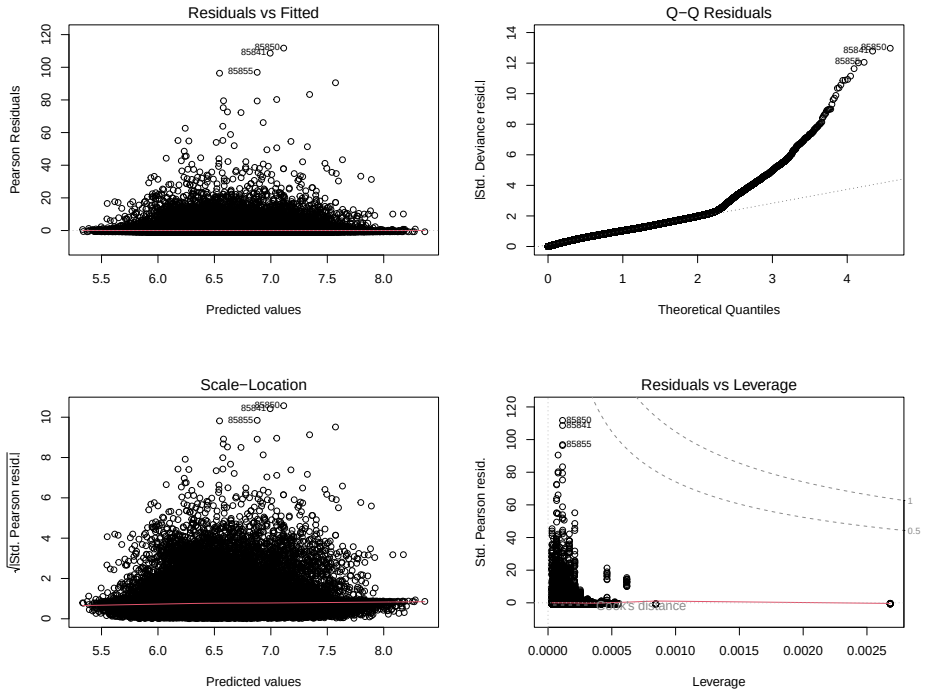

Figure 26: Residual Plots for cluster model

References

[1] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.

[2] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble. The meme suite. *Nucleic acids research*, 43(W1):W39–W49, 2015.

[3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[4] J. Cheng, K. C. Maier, Ž. Avsec, P. Rus, and J. Gagneur. Cis-regulatory elements explain most of the mrna stability variation across genes in yeast. *Rna*, 23(11):1648–1659, 2017.

[5] Y. Di, D. W. Schafer, J. S. Cumbie, and J. H. Chang. The nbp negative binomial model for assessing differential gene expression from rna-seq. *Statistical applications in genetics and molecular biology*, 10(1), 2011.

[6] J. C. Duda, C. Drenda, H. Kästel, J. Rahnenführer, and F. Kappenberg. Benefit of using interaction effects for the analysis of high-dimensional time-response or dose-response data for two-group comparisons. *Scientific Reports*, 13(1):20804, 2023.

[7] N. Eling, A. C. Richard, S. Richardson, J. C. Marioni, and C. A. Vallejos. Correcting the mean-variance dependency for differential variability testing using single-cell rna sequencing data. *Cell systems*, 7(3):284–294, 2018.

[8] R. García-Rodas, R. J. Cordero, N. Trevijano-Contador, G. Janbon, F. Moyrand, A. Casadevall, and O. Zaragoza. Capsule growth in cryptococcus neoformans is coordinated with cell cycle progression. *MBio*, 5(3):10–1128, 2014.

[9] E. S. Hughes, Z. He, L. R. Tuck, E. R. Ballou, and E. W. Wallace. A trade-off between proliferation and defense in the fungal pathogen cryptococcus at alkaline ph is controlled by the transcription factor gat201. *bioRxiv*, 2023.

[10] M. Jaillard, M. Palmieri, A. van Belkum, and P. Mahé. Interpreting k-mer–based signatures for antibiotic resistance prediction. *Gigascience*, 9(10):giaa110, 2020.

[11] O. W. Liu, C. D. Chun, E. D. Chow, C. Chen, H. D. Madhani, and S. M. Noble. Systematic genetic analysis of virulence in the human fungal pathogen cryptococcus neoformans. *Cell*, 135(1):174–188, 2008.

[12] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15:1–21, 2014.

[13] R. C. May, N. R. Stone, D. L. Wiesner, T. Bicanic, and K. Nielsen. Cryptococcus: from environmental saprophyte to global pathogen. *Nature Reviews Microbiology*, 14(2):106–117, 2016.

[14] C. Moeckel, M. Mareboina, M. A. Konnaris, C. S. Chan, I. Mouratidis, A. Montgomery, N. Chantzi, G. A. Pavlopoulos, and I. Georgakopoulos-Soares. A survey of k-mer methods and applications in bioinformatics. *Computational and Structural Biotechnology Journal*, 2024.

[15] R. Negroni. Cryptococcosis. *Clinics in Dermatology*, 30(6):599–609, 2012. Systemic Mycoses.

[16] T. R. O'Meara and J. A. Alspaugh. The cryptococcus neoformans capsule: a sword and a shield. *Clinical microbiology reviews*, 25(3):387–408, 2012.

[17] M. D. Robinson and G. K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332, 2008.

[18] K. P. Vatcheva, M. Lee, J. B. McCormick, and M. H. Rahbar. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale, Calif.)*, 6(2), 2016.

[19] O. Zaragoza, B. C. Fries, and A. Casadevall. Induction of capsule growth in cryptococcus neoformans by mammalian serum and co2. *Infection and immunity*, 71(11):6155–6164, 2003.

Appendices

A Word count

Word count of the main text: 4994

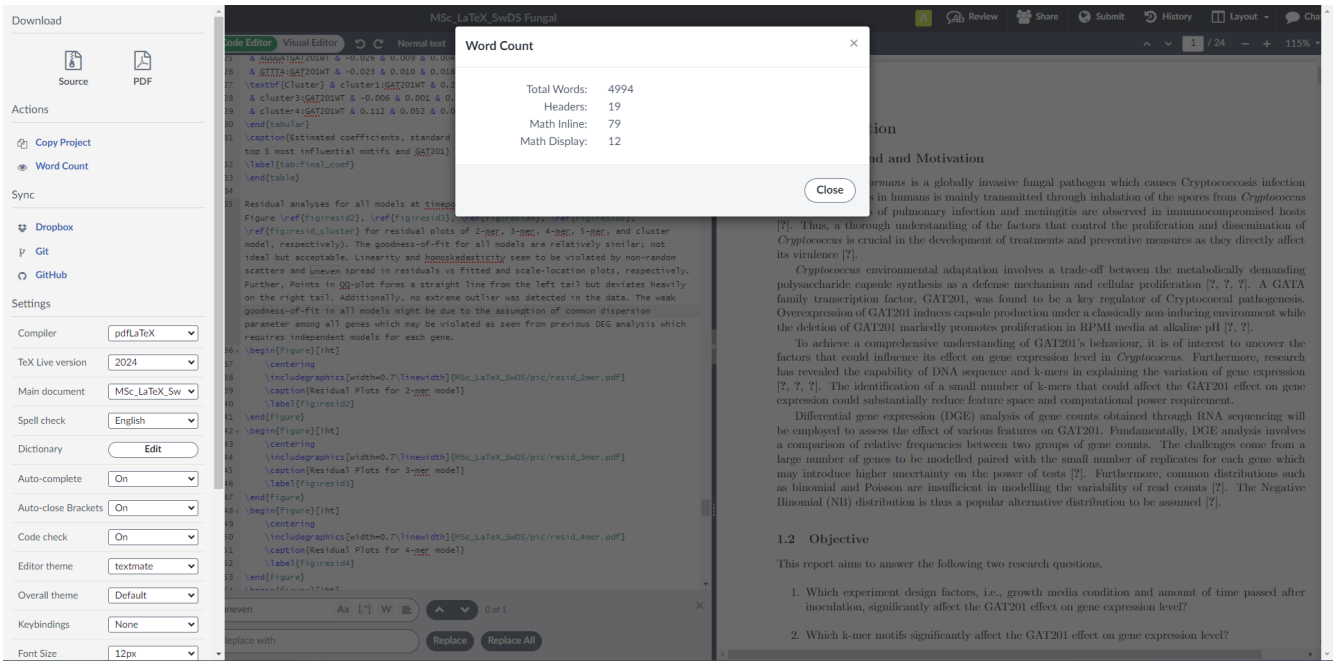

Figure 27: Screenshot of LaTeX word count of the main text