The Report Committee for Xiwei Yan Certifies that this is the approved version of the following report:

Genomics Analysis on the Responses of *E. coli* cells to Varying Environmental Conditions

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor : ______________________
Claus O. Wilke

____
Lizhen Lin

Genomics Analysis on the Responses of *E. coli* cells to Varying Environmental Conditions

by

Xiwei Yan, B.S.; B.S.; M.A.

Report

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fufillment
of the Requirements
for the Degree of

Master of Science in Statistics

The University of Texas at Austin
May 2016

Genomics Analysis on the Responses of *E. coli* cells to Varying Environmental Conditions

Xiwei Yan, M.S.Stat.

The University of Texas at Austin, 2016

Supervisor: Claus O. Wilke

The natural living environments of *E. coli* cells are diverse, varying from mammalian gastrointestinal tracts and soil. Each environment might require distinct metabolic pathways and transporter systems, and long-term evolution has established elaborate regulatory system for *E. coli* cells to quickly adapt to the changing conditions. Sensing outside stresses and then adopting a different phenotype enable them to take advantage of any possible nutrients and defend against hostile environment. A lot of regulatory mechanisms have been identified by genetic, biochemical and molecular biology methods, and our study aim to build a systematic view on the response of the whole genome to four different environmental conditions. We used statistical tests including Pearson's tests and Spearman's tests and multiple testing adjustments to identify feature genes that are induced or repressed significantly across treatment levels. The feature genes identified were partially supported by previous literatures, and some of the novel genes not found in any previous studies may infer a potential research blind spot. Additionally, we compared the correlation tests to the implementation of machine learning algorithms, and discussed the advantage and drawbacks of each method.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Adaptive Response of *E. coli* to the Diverse Environment

The widely used model organism, *Escherichia coli,* is one of the best-studied microorganisms and is of interest both industrially and pathologically. *E. coli* cells have been found in diverse environment, including mammalian gastrointestinal tracts and soil. They are also challenged by diverse environmental stress like high osmolality, lack of nutrient and extreme temperatures. Consequently, the capability of *E. coli* to sense hostile environmental stress and change its physiological and biochemical properties for adaptation is the key to survive. In other words, *E. coli* cells are able to quickly switch to a different nutrient catabolic pathway or turn on different transporting systems through the elaborate regulation of gene expressions.

Each time the living condition changes in the environment, an extensive response is usually required, involving a group of genes and a variety of compounds. Different environment might require a distinct set of genes to be expressed or repressed. And the effect of the four conditions involved in our study will be discussed in the next few sections.

1.1.1. Carbon Source

Similar to mammals and many other organisms, glucose is the primary carbon source for *E. coli* cells. When glucose is present in the environment, *E. coli* cells will repress the alternative carbon source catabolism through a global regulation system known as carbon catabolite repression(Saier, 1998).

However, when glucose is limited in the environment, the *E. coli* cells will begin to relieve the carbon catabolite repression and activate other nutrient catabolic pathways, which usually involve the global transcription factor cyclic Amp (cAmp) receptor protein (Crasnier, 1996; Hengge-aronis, 1996). Once the alternative carbon source metabolic pathways are activated, the cells are able to thrive on the nutrients available in environment.

The lac operon induction is probably the most classic and well-studied regulatory system in *E. coli*. Lactose, or its experimental alternative IPTG, is able to induce the lactose metabolism genes by binding and removing the repressor protein from the lac operon promoter region (Chuang et al., 1993; François Jacob, 1961). This experimental system is now widely used in all research laboratory and genetic engineering industry. The substrate-specific induction mechanism now generalizes to the regulation of many other carbonhydrate catabolic genes (Brückner and Titgemeyer, 2002). Multiple such regulatory mechanisms allow *E. coli* cells to activate pathways of alternative carbon source (Liu et al., 2005).

Liu et al. reported the effect of 5 different carbon sources on global gene expression. The transcriptional profiles of the 5 alternative carbon sources were compared with that of glucose, and 50-270 genes were identified as differentially expressed genes (Liu et al., 2005).

1.1.2 Growth Phase

The growth culture of bacterial cells in labs usually goes through 4 steps: lag phase, exponential phase, stationary phase and death phase. Cells divide and reproduce rapidly only in the exponential phase, and begin to accumulate toxic compounds at stationary phase and end their lives at death phase. However, the ambient, nutrient-rich environment as in the exponential phase of cell culture is not representative of *E. coli* natural growth environment. On the contrary, *E. coli* cells in their natural habitat often have limited nutrient and hostile conditions, and have to develop stationary-phase properties to survive the periods of starvation (Hengge-aronis, 1996). Stationary-phase *E. coli* cells exhibit some morphological and physiological properties that are not present in the exponential-phase cells, including shrinkage of cell size, condensation of cytoplasm, accumulation of storage compounds and protective substances, and structural changes in many cellular component and cell membrane (Jenkins et al., 1990; Kolter et al., 1993; Siegele and Kolter, 1992). Such transition requires global gene regulation and large-scale protein synthesis that profoundly changes cell composition and physiology. And one of these master regulators is the rpoS-encoded sigma factor, the σ^s subunit of RNA polymerase.

The σ^s subunit, as an crucial part in transcriptional regulation, is strongly induced in stationary-phase and is essential for the expression of many stationary-phase responsive genes (Hengge-aronis, 1996). Up to 10% of *E. coli* genes have been identified in genomic study downstream of σ^s-dependent regulation, and more than 80 of them have been confirmed using genetic and molecular biology approach (Weber et al., 2005).

Aside from σ^s-dependent regulation, a group of nucleoid proteins related to DNA replication and transcription were also reported differentially expressed in stationary phase, indicating their regulatory roles in DNA functions such as replication, repair, recombination and protection (Ali Azam et al., 1999).

1.1.3 Sodium Levels

High sodium level in the environment represents the osmotic stress, which is quite common in the mammalian gastrointestinal tracts environment where *E. coli* cells usually stay (Weber et al., 2006). The cells have evolved a system to adapt to the high osmotic conditions. The consequences of sudden exposure to high osmolality include loss of water, reduce in respiration, and increase in pH (Weber and Jung, 2002). To counteract these changes, the cells will increase potassium uptake, accumulate osmoprotectants (betaine and trehalose) to balance external osmolarity and protect proteins from denaturation (Higgins et al., 1988). All these responses require a regulatory process involving activation of biosynthetic pathways and transport systems for the osmoprotectant (Purvis et al., 2005; Weber and Jung, 2002). And the same high osmotic effect also applies to cells when high external magnesium is present in the surroundings.

Extensive studies have been performed to investigate the mechanism of adaptation to high osmolality environment. A genome-wide study detected 152 genes with altered transcription levels in response to the high sodium condition, indicating a global effect of high osmality on gene expression. And several selected genes were confirmed by biochemical experiment (Weber and Jung, 2002).

1.1.4 Magnesium Levels

Besides the effect on osmolality as discussed in Section 1.1.3, the magnesium ions also play a central role in many cellular activities, including ATP-consuming reactions, DNA replication and transcription. Thus, any change in the external Mg^{2+} levels in the environment should be sensed and adapted by the bacterial cells quickly.

The *E. coli* cells adopt two-component signal transduction mechanism to regulate cellular responses towards a spectrum of environmental stimuli. The phoP/phoQ two-component system senses the external Mg^{2+} concentration change and mediates the transcription activation of Mg^{2+}-responsive genes. In this system, phoQ acts as the sensor protein, phosphorylate phoP regulator protein at low Mg^{2+} concentration, and induces around 30 genes (Soncini and Groisman, 1996). The phoP/phoQ genes showed up in the list of regulated genes in our statistical tests, which will be discussed in Chapter 4.

Previous research has shown that 232 genes with 0.75 fold change in expression levels with external Mg^{2+}, and 13 of these genes have conserved DNA sequence in the promoter region (Kato et al., 1999; Minagawa et al., 2003). This result is also supported by our study, as discussed later in Chapter 3.

1.1.5 Evolutional Significance of the Adaptive Gene Regulation

E. coli cells live in diverse and usually hostile environment. In order to survive and reproduce in various conditions, the cells have evolved to efficiently use all types of nutrients, and survive when encountered with extreme environment. Meanwhile, specific metabolic pathways will be turned off whenever the substrate is no longer available (Liu et al., 2005). This elaborate gene regulation system enables the utilization of cellular resources without wasting energies on producing useless enzymes or transporters.

1.2 Analysis of Transcriptomic Data Using Statistical Method and Machine Learning Models

In the past centuries, various genetic, biochemical and molecular biology approaches like Western blot, Northern blot and 2-D electrophoresis have been used to investigate the responses of *E. coli* cells to various environmental stimuli (Ali Azam et al., 1999; Chuang et al., 1993; Weber and Jung, 2002; Weber et al., 2006, 2005). However, these approaches are commonly blamed for their low-efficiency, low-throughput, arbitrary criteria and usually lack of statistical evidence. On the other side, the systematic biology datasets, including genomic, proteomic and transcriptomic data, are attracting more interest as experimental and computational techniques progress. The '-omics' data from experiments like microarray, RNA-seq, Mass-spectrometry are being generated and analyzed everyday, and the researchers are taking advantage of the vast amounts of data to establish a more comprehensive and robust model to study cellular activities (Chuang et al., 1993; Eguchi et al., 2004; Minagawa et al., 2003; Weber et al., 2005).

The traditional criterion of defining differential gene expression is either by visual inspection or setting a threshold for mean signal intensity. For example, some studies pick genes with larger than 3-fold change or larger than 1.4 log2 fold change, and such kind of criteria are quite arbitrary and baseless (Liu et al., 2005; Weber and Jung, 2002). The adoption of statistical methods in genetic analysis introduces several advantages: (1) It establishes a well-defined rule for identification of the differentially expressed genes; (2) It controls for false positive discovery rate for the large number of parallel tests; (3) It has a mathematical foundation and higher accuracy.

Meanwhile, more and more machine learning algorithms have been introduced into the field of genomics studies (Reed et al., 2003; Trentini et al., 2013). In our situation, the problem of identifying genes that differ in expression levels is then converted to a classification or clustering problem. Genes that are significantly up-regulated or down-regulated will produce higher accuracy in terms of predicting living conditions from gene expression levels. Thus, the Bayesian regression model and the Gaussian mixture model are used in our study to identify genes that may not be detected from classic statistical tests.

Chapter 2 Experiment Setting and Methods

2.1 Experiment Setting

The gene expression data were collected from143 *E.coli* samples cultured in 34 distinct growth conditions. In each sample, the gene expression levels of 4279 genes were measured and the associated treatment/growth conditions were recorded. These manipulated environmental factors include: growth time, carbon source, Mg^{2+} level, and Na^{+} level. Except for carbon source, the other three factors were recorded in both continuous scales (grow time in hours, sodium/magnesium concentration in millimolar) and categorical forms (growth phase as exponential/stationary/late_stationary, sodium/magnesium level as low/base/high).

2.2 DESeq2 package and Data Preprocessing

The DESeq2 package (Version 1.6.3) available in R (Version 3.2.0) was used in our analysis to study the differential gene expression patterns in bacteria when treated with different growth environment. The package DESeq2 is based on a negative binomial generalized linear model, and is well suited for handling raw count data set from techniques like RNA-Seq (Anders and Huber, 2010). DESeq2 package contains functions including estimation of size factor, auto-filtration of outlying data points, normalization of data by the estimated size factor, and variance stabilizing transformation. Our raw dataset collected from the 143 samples over 4279 genes were pre-processed by the DESeq2 package as mentioned above, and a logarithmic transformed and standardized dataset were obtained. These pre-processing steps are required for dataset coming from different samples, as the raw count data will be significantly affected by confounding factors like sequencing depth.

2.3 Estimation of log2 fold change

The DESeq2 package was used to estimate the logarithmic fold change of gene expression in the treatment group compared to that of the control group. However, DESeq2 package by default takes the first level in alphabetical order as the control group, and make comparison between the control group and all other groups. In order to detect a significant up-regulation/down-regulation occurred in any treatment levels, logarithmic ratios were calculated from all possible combinations of treatment groups by repeatedly re-coding the treatment levels before using the DESeq2 results function.

2.4 Correlation Test

Correlation tests were used to detect whether the gene expression responds to some changes in the growth conditions.

For the continuous variables, including bacterial growing time (in hours), the magnesium concentration (in millimolar) and sodium concentration (in millimolar), the correlation between the continuous factor and the gene expression was analyzed using both Pearson's correlation test and Spearman's rank correlation test, with or without logarithmic transformation.

For the categorical variables, including carbon source in the bacteria growth medium, growth phase the bacteria culture was collected, and the magnesium/sodium levels, the correlation between these variables and gene expression levels were analyzed using the same types of correlation tests. But in order to perform the tests, the categorical variables were transformed into numeric variables using two different methods: (1) coding each factor level by consecutive integer numbers; (2) coding the factor levels by binary numbers. Then the correlation tests were performed on all possible permutations of the factor levels, and the most significant test in each coding scheme were kept as the result. Use the Carbon Source factor as an example, method (1) will code factor levels gluconate, glucose, glycerol, lactate as any permutation of {1,2,3,4} and choose the permutation that generate the most smallest p-value, and method (2) will code factor levels as any permutation of {1,0,0,0} and also choose the permutation that generate the most significant result.

2.5 Multiple Testing Adjustment

The correlation tests were performed on all 4279 genes, and the multiple testing problem will greatly inflate the type I error. Thus, p-values were adjusted using the Holm–Bonferroni method to control for the false discovery rate.

2.6 Bayesian Regression Model

The relationship between continuous environmental factors (Na^+ in mM, Mg^{2+} in mM and growth time in hour) and the standardized gene expression levels was also modeled by Bayesian regression method. The likelihood function is shown as follow:

$$L(Y_i \mid \beta_0, \beta_1, \sigma) \sim N\,(\beta_0 + \beta_1 X_i, \sigma^2)$$

where Y_i represents the logarithmic transformed continuous factor value, and X_i represents the corresponding standardized gene expression level. β_0 and β_1 represents the parameters to be estimated in the Bayesian regression model. Conjugate prior distributions for β_0, and β_1 were assumed to be standard normal, while conjugate prior for σ^2 were assumed to be inverse-gamma with pre-defined hyper-parameter. Gibbs sampling algorithm was used to estimate β_0, β_1 and σ^2, and the estimated parameters were then used to calculate the fitness of model.

2.7 Gaussian Mixture Clustering Algorithm

The clustering algorithm was used to detect gene expression differences among categorical factors levels, by assuming the gene expression data follow a Gaussian mixture model as follow:

$$\text{Yi} \sim \sum_{i=1}^{k} w_i \, \mathrm{N}(\mu_i, \sigma_i^2)$$

where the standardized gene expression Y_i is assumed as an i.i.d. sample from a mixture of k Gaussian models with mean u_i and standard deviation σ_i, with k being the number of categorical levels. The Gaussian models were combined with weight w_i. According to the Gaussian mixture model, the standardized gene expression data from ith level were assumed to follow a Gaussian distribution with center μ_i and variance σ^2, and the probability of belonging to the ith treatment group is the weight w_i.

The prior distributions of the parameters u_i's, σ_i's, and w_i's were assumed to be standard normal distributions, inverse-gamma distributions, and dirichlet distribution, respectively. A set of latent variables Z_i was added to facilitate the sampling process and the classification of clusters.

The Gibbs sampling algorithm was implemented to estimate the posterior distributions of the standardized gene expression in each treatment group, and a distinguishable change in gene expression level can be detected using metrics discussed in the next section.

2.8 Performance Metrics of the Bayesian Regression and Gaussian Mixture Models

In order to represent the fitness of the Bayesian regression models, the likelihood of observed data given estimated model was calculated as below:

$$L(Y|Model) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y_i-\beta_0-\beta_1 X_i)^2}{2\sigma^2}}$$

where parameters β_0, β_1 and σ^2 were estimated as described in Section 2.6. Since the standardized gene expression levels were used as predictor variables, larger likelihood indicates a better prediction of the continuous factor values based on the model. And the better predictions are usually associated more distinguishable gene expression levels among factor levels.

For the Gaussian Mixture model, the likelihood of each data is calculated as the probability it comes from its associated cluster, given the mixture model with the estimated mean and standard deviation. For example, if Y_i comes from factor level 1, then the likelihood of Yi will be derived from the normal distribution $N(\mu_1, \sigma_1^2)$. To overcome the dominant effect of one or a few huge probability density values, the likelihood was standardized by the summation of likelihood in all clusters. A large likelihood usually indicates that gene expression data from different factor levels are well separated and very distinguishable.

Another performance metric used for Gaussian Mixture model is the accuracy of classification. Given the clusters with estimated mean and variance, each data was classified into the cluster that generates the largest likelihood. The accuracy is then calculated by comparing the predictions with the actual observations. Higher accuracy indicates better performance in classifying and predicting the categorical factor labels, and consequently is usually associated with differentially expressed genes.

Chapter 3 Result

3.1 DESeq2 Log2 Fold Change Estimation and Statistical Testing

The DESeq2 package method takes the raw count matrix of gene expression and an experiment design matrix as input, performs automatic data filtration, data transformation and data normalization by the estimated size factor, and outputs log2 fold change and p statistics from the negative binomial Wald Test between the default control group and all other treatment groups. The limit of DESeq2 is that it takes only use categorical experiment design variables, so only the categorical factors (growth phase, carbon source, Na^+ levels and Mg^{2+} levels) were used in this DESeq2 analysis.

As shown in Figure 3.1, the DESeq2 method successfully identified groups of genes with significant changes in expression levels in all four experiments. The log2 fold change is correlated with adjusted p-values, as genes with large log2 fold change (>2) always have significant adjusted p-value, and genes with small log2 fold change (close to 0) are always non-significant. However, the genes with intermediate log2 fold change (>0.5 and <2) appear to have a weaker relationship with the test result. Our result disagrees with previous statement by Weber and Jung about the lack of correlation between n-fold differences and significance, but also indicates that the traditional threshold set by certain fold change or certain log2 fold change may not be statistically correct (Weber and Jung, 2002). Although not suitable as an arbitrary criterion, setting a filter based on the log2 fold change might be helpful in getting rid of some noises in the data, and controlling for the same false discovery rate with less stringent adjusted p-values.

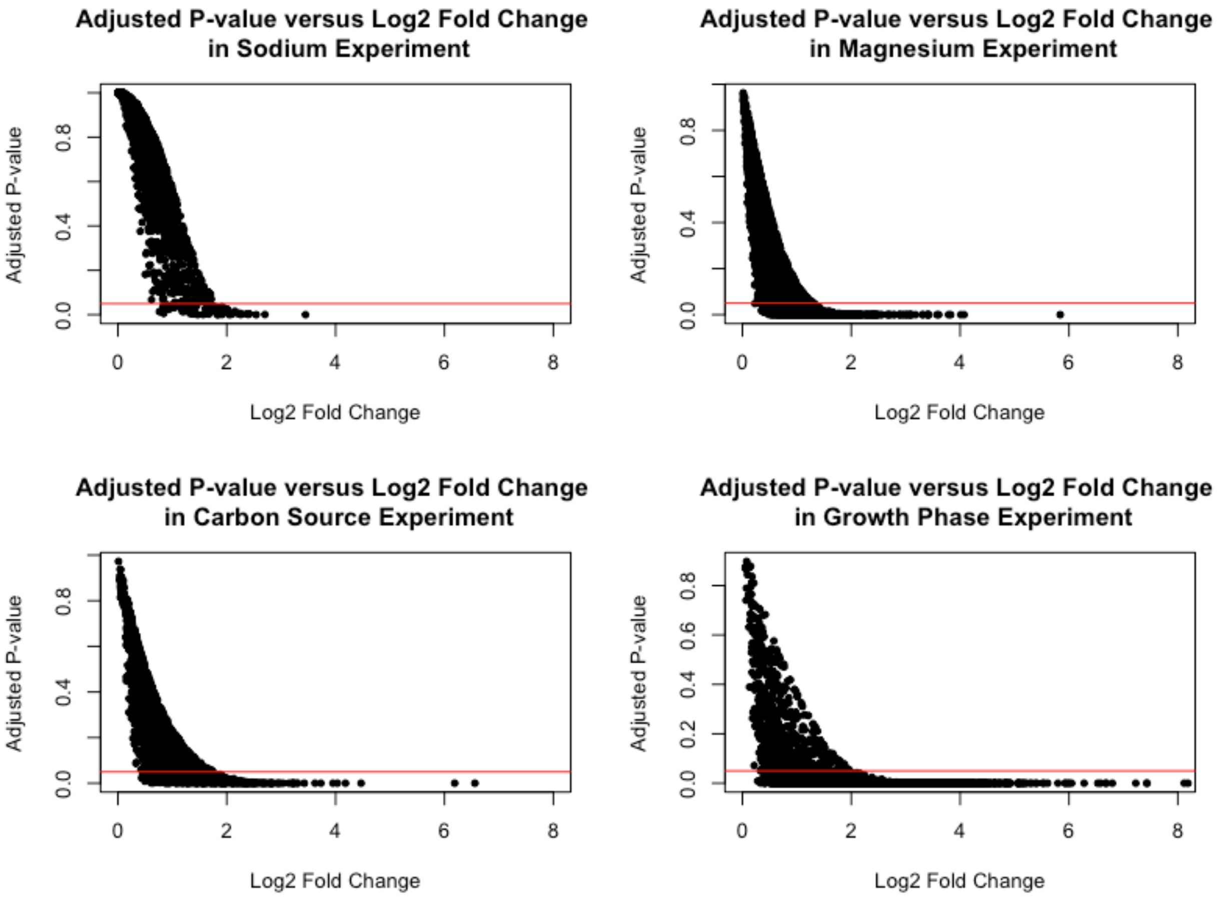

Figure 3.1 Relationship between the DESeq2 Adjusted p-values and Log2 Fold Change with Respect to 4 Categorical Environmental Factors. The DESeq2 adjusted p-value decreases as log2 fold change increases, but this correlation is weaker with intermediate log2 fold change. The red line in each panel pinpoint the position of significance level. The data points below red line represent genes with significant difference in gene expression level.

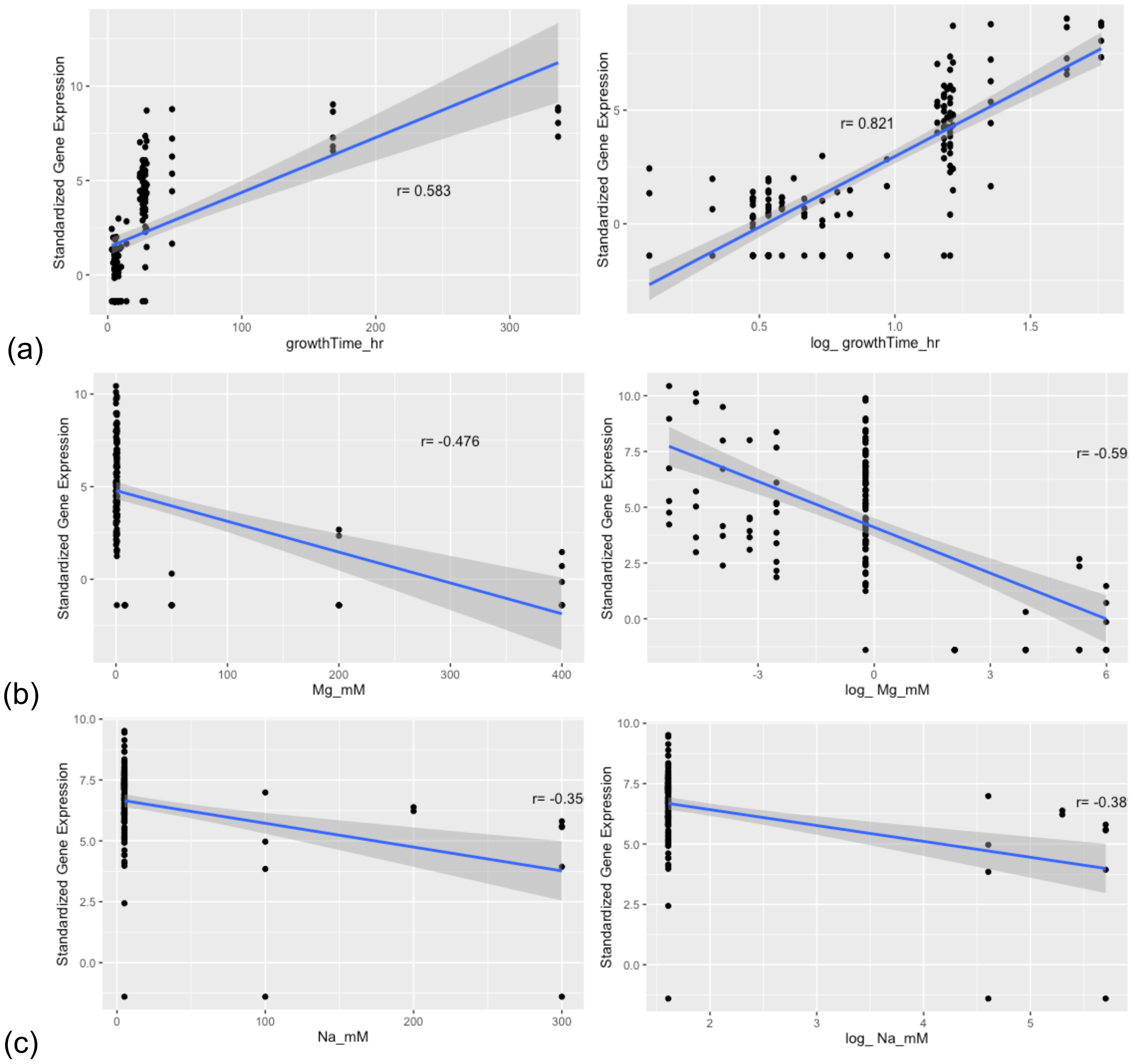

Figure 3.2 The Effect of Logarithmic Transformation of the Continuous Factors on Pearson's Test. (a) Gene ycfR and growth time (hr). (b) Gene asr and Mg2+ (mM). (c) Gene nagE and Na+ (mM). The gene in each panel was selected from the genes identified to be significant in both transformed and untransformed dataset. In each panel, the plot on the left shows the result based on original numeric factor and the plot on the right shows the result based on the logarithmic transformed factor.

Variable	Method	# Significant Genes	% Significant Genes
Growth Time (hr)	Pearson's Test with Log Transformation	2851	0.666
Growth Time (hr)	Pearson's Test	1087	0.254
Mg2+ (mM)	Pearson's Test with Log Transformation	625	0.146
Mg2+ (mM)	Pearson's Test	204	0.048
Na+ (mM)	Pearson's Test with Log Transformation	7	0.002
Na+ (mM)	Pearson's Test	10	0.002

(a)

	Sodium Experiment	
	Pearson's Test with Log Transformation	
Pearson's Test	Non-significant	Significant
Non-significant	4262 (0.996)	1 (0.0002)
Significant	4 (0.0009)	6 (0.0014)
	Magnesium Experiment	
	Pearson's Test with Log Transformation	
Pearson's Test	Non-significant	Significant
Non-significant	3602 (0.842)	467 (0.109)
Significant	47 (0.011)	157 (0.036)
	Growth Time Experiment	
	Pearson's Test with Log Transformation	
Pearson's Test	Non-significant	Significant
Non-significant	1105 (0.258)	2081 (0.486)
Significant	318 (0.074)	769 (0.180)

(b)

Table 3.1 Logarithmic Transformation of the Continuous Factors Increased the Power of Pearson's Test. (a) Pearson's tests using transformed factors identified more significant genes. (b) Pearson's tests with transformation captured most of the genes identified by tests without transformation. The overlap between Pearson's test with or without transformation is shown as the highlighted cell. Percentages over the total number of genes (4279 genes) are shown in the parenthesis of each cell.

3.2 The Logarithmic Transformation of the Continuous Factors Increased the Robustness of Pearson's Correlation Tests by Reducing the Effect of High-leverage Data Points

As shown in Figure 3.2, the untransformed numeric factors take values ranging from level of 10^{-2} to several hundreds, with most of data points clustered around 0. This unbalanced data distribution results in the high-leverage data points at the far end of the left-side plots in each panel. The concept of high-leverage data points indicates the correlation tests are more vulnerable to outliers and measurement errors, and leads to a less robust model. Consequently, taking logarithmic transformation of the numeric factors largely alleviates the problem of data imbalance and increases the strength of linearity (as shown in plots on the right-hand side of Figure 3.2).

The logarithmic transformation also increased the power of Pearson's correlation tests in most cases, due to the increased linearity. According to the statistics in Table 3.1, the Pearson's tests with transformed dataset capture most of the significant genes identified from the untransformed dataset in all three experiments, and meanwhile they identify a lot more significant genes in the Magnesium Experiment and Growth Time Experiment. The exception of Sodium Experiment may be resulted from the highly unbalanced data structure and the lack of linearity among the log sodium salt level and the standardized gene expression levels.

On the other hand, the results from Spearman's correlation tests are not affected by the logarithmic transformation. The reason lies in the fact that Spearman's correlation test is non-parametric and based on rank only. Consequently, the monotonic logarithmic transformation only changes the value, but not the rank. The results from the Spearman's tests and Pearson's tests were compared in the next section.

3.3 Pearson’s Test and Spearman’s Test Perform Similarly in Most of the Experiments

The Spearman’s correlation test is well suited for scenarios when the assumption of linearity and normality is absent, but it will also lead to less power than the Pearson’s test if the assumptions are met. According to the results in table 3.2, the Spearman’s tests perform very similarly with Pearson’s tests with both continuous factors (a) and categorical factors (b), given the same pre-processing steps. In most experiments, Pearson’s test and Spearman’s test identified similar amount of significant genes. Moreover, over 90% of those genes were identified in both tests (result not shown).

The only exception comes from the Sodium Experiments, in both numeric and categorical forms. Spearman’s tests failed to identify any gene with significant changes in expression level after adjusted for multiple testing. On the other hand, Pearson’s tests identified up to 10 genes, which is still fewer than what we expected for bacterial genomics response to sodium stress. A possible explanation is the large variation resulted from the small number of data associated with certain levels (4 data at 100mM, 2 data at 200mM and 5 data at 300mM). This highly unbalanced data distribution leads to high model variance and hence more uncertainty. However, we are still able to identify some genes that aligns with previous research result, which will be discussed in later sections.

Variable	Method	# significant Genes	% significant Genes
Continuous Experimental Factors			
Growth Time (hr)	Spearman's Test	2858	0.668
Growth Time (hr)	Pearson's Test with Log Transformation	2851	0.666
Mg^{2+} (mM)	Spearman's Test	735	0.172
Mg^{2+} (mM)	Pearson's Test with Log Transformation	625	0.146
Na^{+} (mM)	Spearman's Test	0	0.000
Na^{+} (mM)	Pearson's Test with Log Transformation	7	0.002

(a)

Categorical Experimental Factors			
Factor Coded by Consecutive Integer			
Carbon Source	Spearman's Test	216	0.050
Carbon Source	Pearson's Test	190	0.044
Growth Phase	Spearman's Test	3886	0.908
Growth Phase	Pearson's Test	3938	0.920
Mg2+ Levels	Spearman's Test	1966	0.459
Mg2+ Levels	Pearson's Test	2068	0.483
Factor Coded as Binary Vector			
Carbon Source	Pearson's Test	89	0.021
Carbon Source	Spearman's Test	153	0.036
Growth Phase	Pearson's Test	3983	0.931
Growth Phase	Spearman's Test	4031	0.942
Mg2+ Levels	Pearson's Test	2136	0.499
Mg2+ Levels	Spearman's Test	2074	0.485
Na+ Levels	Spearman's Test	0	0.000
Na+ Levels	Pearson's Test	10	0.002

(b)

Table 3.2 Pearson’s Test and Spearman’s Test Perform Similarly Given the Same Pre-processing Steps. (a) Pearson’s and Spearman’s tests identified similar number of genes for each continuous factor (Note: log transformation was always used as discussed in section 3.2). (b) Pearson’s and Spearman’s tests identified similar number of genes for each categorical factor given the same coding strategy.

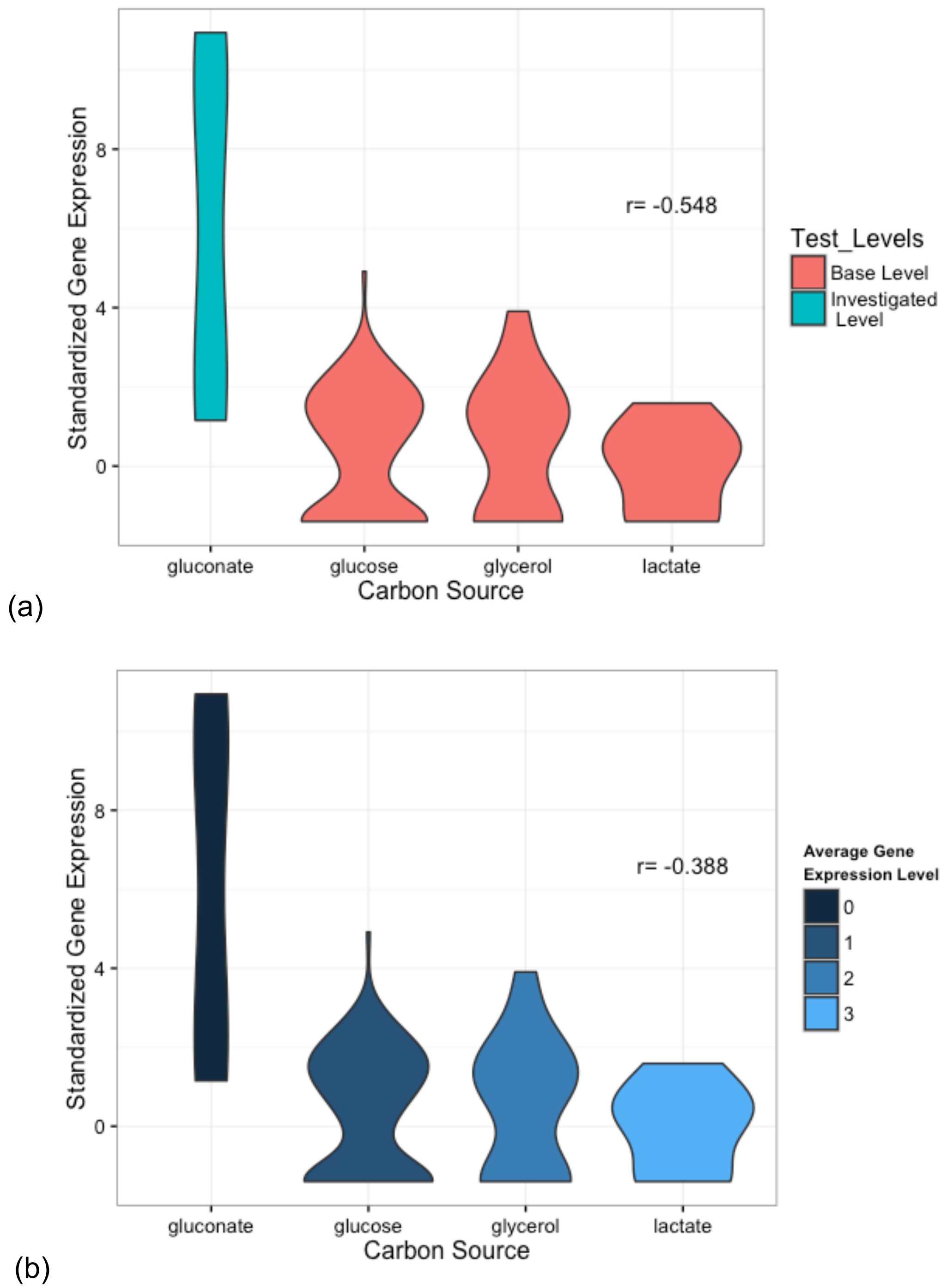

Figure 3.3 Illustration of the Differences between the Coding Schemes of Categorical Factors. (a) Binary-coded factors (one level is coded as 1 and all the other levels are coded as 0). (b) Integer-coded factors (the four levels of carbon source are coded as consecutive integers 1 to 4). Both panel used data from gene idnD.

Magnesium Experiment				
Pearson's Test (Integer-coded)	Pearson's Test (Binary-coded)	Spearman's test (Integer-coded)	Spearman's Test (Binary-coded)	
			Non-significant	Significant
Non-Significant	Non-Significant	Non-Significant	1629 (0.381)	24 (0.006)
		Significant	17 (0.004)	36 (0.008)
	Significant	Non-Significant	278 (0.065)	82 (0.019)
		Significant	9 (0.002)	131 (0.031)
Significant	Non-Significant	Non-Significant	6 (0.001)	0 (0)
		Significant	4 (0.0009)	1 (0.0002)
	Significant	Non-Significant	209 (0.049)	80 (0.019)
		Significant	48 (0.011)	1719 (0.402)

Carbon Source Experiment				
Pearson's Test (Integer-coded)	Pearson's Test (Binary-coded)	Spearman's test (Integer-coded)	Spearman's Test (Binary-coded)	
			Non-significant	Significant
Non-Significant	Non-Significant	Non-Significant	3983 (0.931)	15 (0.004)
		Significant	52 (0.012)	27 (0.006)
	Significant	Non-Significant	4 (0.0009)	2 (0.0005)
		Significant	0 (0)	0 (0)
Significant	Non-Significant	Non-Significant	29 (0.007)	2 (0.0005)
		Significant	40 (0.009)	36 (0.008)
	Significant	Non-Significant	11 (0.003)	11 (0.003)
		Significant	1 (0.0002)	60 (0.014)

Growth Phase Experiment				
Pearson's Test (Integer-coded)	Pearson's Test (Binary-coded)	Spearman's test (Integer-coded)	Spearman's Test (Binary-coded)	
			Non-significant	Significant
Non-Significant	Non-Significant	Non-Significant	164 (0.038)	19 (0.004)
		Significant	7 (0.002)	68 (0.016)
	Significant	Non-Significant	2 (0.0005)	36 (0.008)
		Significant	0 (0)	40 (0.009)
Significant	Non-Significant	Non-Significant	6 (0.001)	8 (0.002)
		Significant	4 (0.0009)	15 (0.004)
	Significant	Non-Significant	48 (0.011)	105 (0.025)
		Significant	12 (0.003)	3739 (0.874)

Table 3.3 The Effect of Coding Scheme Varies across Experiment. Genes identified in growth phase experiment have high overlap between the two coding schemes, while this is not the case for carbon source experiment or magnesium experiment. Percentages over the total number of genes (4279 genes) are shown in the parenthesis of each cell.

3.4 The Effect of Different Coding Strategy of the Categorical Factors on Correlation Tests Result

The two coding strategy of the categorical factors (coded by consecutive integers or coded by binary vectors) are expected to identify two distinct groups of genes with different expression patterns. The binary-coding strategy aims to identify genes differentially expressed in one of the factor levels, while the integer-coding strategy aims to identify genes showing a gradient increase or decrease as factor changes. To demonstrate the difference between these two coding strategies, we used the expression levels of gene idnD as an example (Figure 3.3). idnD is identified as a differentially expressed gene through Pearson's correlation test based on both coding strategy. However, the binary-coding method (Panel A) focused on the difference between one level (gluconate) and all other levels and identified the most different carbon source, while the integer-coding method (Panel B) suggests a gradual decrease in gene expression from gluconate to lactate.

Although the expectations are different for the two coding strategies, surprisingly they end up identifying similar groups of genes in the Growth Phase experiment, with more 90% overlap (represented in the highlighted cells in table 3.3). On the other hand, the results in the Magnesium experiment and the Carbon Source experiment showed 65% and 20% overlap in the identified genes respectively.

3.5 Correlation Tests using Categorical Factors Identified More Genes than their Continuous Equivalents

The continuous numeric form of the environmental factors retains more information than the categorical equivalent, and hence are expected to be more accurate in distinguish the difference in gene expression. However, the results suggest that categorical factors generally have more power in the correlation tests, and also identified more significant genes, regardless of method. According to our results, tests using Growth Phase identified 3886~4031 genes, significantly larger than tests using Growth Time (1087~2858 genes). Likewise, Mg^{2+} Levels factor identified 1966~2136 genes, compared to the Mg^{2+} (mM) factor (204~735 genes).

3.6 Both Correlation Tests and DESeq2 Method Support and Complement Previous Studies

Groups of genes responding to external Mg^{2+}/Na^{+} fluctuation or change of growth phase have been identified in numerous past studies through genetic, biochemical and molecular biology methods.

With respect to the effect of osmotic stress, Weber et al. identified a group of 22 genes that are induced by high concentration of NaCl within the first 60min of induction (Weber et al., 2006). The correlation tests identified 7 of them before multiple testing adjustments, and 1 of them remains to be significant after the adjustment. Considering the small number of sodium-responsive genes (<10) that correlation tests identified, the chance of capturing 1 out of the 22 is much larger than randomly selection (~0.0023). On the other hand, DESeq2's negative binomial generalized linear model identified 71 significant genes, with 9 of them in common with the group of 22 genes. Meanwhile, some genes identified by the correlation method but not in the previous study are supported by other researches. Two transport and binding proteins proV and proW, and a protease hycI were found up-regulated with a global false discovery rate 12% (Weber and Jung, 2002; Weber et al., 2005).

Researches investigating magnesium-stimulated transcription also provide a list of genes known to be regulated by Mg^{2+} signal in bacterial cells (Minagawa et al., 2003). Out of the set of 13 genes, 9 of them were identified by the correlation tests with categorical factor Mg^{2+} level, and 7 of them were identified by DESeq2 package model.

The pool of genes identified as differentially expressed at different growth phase or growth time was also supported by literature. Among the list of 121 genes, 110 were identified by the correlation tests, and 105 were identified by the DESeq2 package model (Weber et al., 2005). Moreover, genes with regulatory roles with DNA functions, as well as genes known to be expressed phase-specific were all identified in our statistical tests (Ali Azam et al., 1999; Chuang et al., 1993).

In sum, the correlation tests outperformed DESeq2 package model for both magnesium dataset and growth phase dataset, but generated less powerful result for sodium dataset. Possible explanations for the lack of significance in the sodium data correlation tests might be the limited number of samples at each treatment levels and the lack of consistent control factors, and both will lead to a larger variance and less powerful test result. Moreover, the auto-filtration performed by DESeq2 package prior to the statistical tests might help get rid of noises existed in the dataset.

3.7 Trial Studies by Bayesian Regression Methods and Machine Learning algorithms

The performance indicator adopted by the Bayesian Regression Model is likelihood of data, from which a conventional and default cutting point for significance is absent. Thus, we compared the top 20 genes identified from the correlation tests and the top 20 genes identified from the Bayesian regression models, and found that a heavy overlap between the genes being selected (data not shown).

On the other hand, the result achieved from the clustering algorithm is not well aligned with the correlation tests, although the top 20 genes picked do suggest differential expression according to visual inspection. One major assumption of the Gaussian mixture model is that samples from each treatment group are i.i.d. normally distributed. However, the interference from confounding factors within each treatment group may violate this crucial normality assumption and consequently lead to defective results. To avoid such drawbacks from the model-based clustering algorithms, density-based clustering algorithms like DBSCAN might be helpful in improving prediction accuracy.

Chapter 4 Concluding Remarks and Future Directions

Considering the widespread distribution and extensive application of E. coli, its genetic regulatory network is of interest academically, clinically and industrially. The adaptive response of E. coli cells to the various environmental conditions roots from their ability to sense distinct external signals and induce or repress groups of metabolic-related and transportation-related genes correspondingly. Our study investigated the genomics-wide response of E. coli cells to four environmental conditions: carbon source, growth phase (time), Na^+ and Mg^{2+} concentrations. The expression levels of 4279 genes were measured and analyzed by correlation tests, Bayesian regression model, and clustering algorithm. The results were compared between models, as well as to the existing R package DESeq2 and past literatures.

Pearson's correlation test and Spearman's correlation test showed evidence of a global effect on *E. coli* genomics by Mg^{2+} stress, switch of growth phase and change of carbon source. Hundreds of genes were proved to be significantly differentially expressed in each of the above datasets by both types of test. Moreover, Based on the large percentage of overlap between result from Pearson's tests and Spearman's tests, the choice of test does not seem to influence the result much.

However, the logarithmic transformation of the continuous factors and the coding scheme of the categorical factors do have an impact. Transformation of the numeric factor reduces data range and improves the linearity between variables, so that it helps capture more genes with significant changes in expression levels. On the other hand, the different coding strategies of categorical factors were aimed to identify genes with different changing behaviors, but they turned out finding similar genes from the growth phase experiment, and results with less overlap in magnesium and carbon source dataset.

The correlation tests performed not as well for the sodium dataset, with only 10 genes found with significant induction or repression. Despite of the lack of power in most tests, the identification of the 10 genes is supported by past studies using biochemical and molecular biology experiments, including the transport proteins proV, proW and proX, and a protease hycI (Weber et al., 2006, 2005). Similar evidence was also found for magnesium-induced genes, as the well-characterized phoP/phoQ two-component system were both identified in our correlation test result (Minagawa et al., 2003).

The lack of default threshold is a problem for both Bayesian regression model and the Gaussian mixture clustering algorithm. Without the threshold, the definition of significance will be vague and quite arbitrary. Nonetheless, the Bayesian regression models identifies almost the same group of top-ranked genes as the correlation tests, supporting the validity of the correlation tests from another aspect. However, the Gaussian-based clustering algorithm does not agree with the correlation tests all the time, possibly due to the violation of normality in part of the dataset. The implementation of model-free density-based clustering algorithm like DBSCAN might avoid such problem and be more helpful in distinguishing treatment levels with respect to our data structure.

In sum, our correlation tests were proved to be effective in identifying feature genes in most of the datasets with certain pre-processing steps, while some additional effort will be worth trying for the machine learning algorithms' implementation.

Appendix R code for Correlation Tests and Clustering

```
calculate_log2 <- function(selectLevel){
        coldata = subset(metaRNA, subset = dataSet %in%
colnames(UnfilteredData), select = c('dataSet',selectLevel ))
        colnames(coldata)[2] = 'condition'
        countData = DESeqDataSetFromMatrix(countData = UnfilteredData,
colData = coldata, design = ~ condition)
        RESULT = NULL; LOG2CHANGE=NULL
        nlevel = length(unique(countData$condition))
        orders = combn(1:nlevel, m=2, simplify=FALSE)
        for (x in orders){
                countData$condition =
factor(coldata$condition,levels(coldata$condition)[c(x[1], (1:nlevel)[-x], x[2])])
                countData = DESeq(countData)
                result = results(countData)
                result = data.frame(gene_id = rownames(UnfilteredData),
log2FoldChange = abs(result$log2FoldChange),pvalue = result$pvalue)
                RESULT = rbind(RESULT, result)
                LOG2CHANGE = cbind(LOG2CHANGE,
abs(result$log2FoldChange))
        }
        log2foldchange = data.frame(gene_id = rownames(UnfilteredData),
log2FoldChange = apply(LOG2CHANGE, 1, max))
        finalResult = merge(RESULT, log2foldchange,
by=c('gene_id','log2FoldChange' ))
        finalResult1 = finalResult %>% group_by(gene_id) %>%
dplyr::summarize(log2FoldChange = max(log2FoldChange), pvalue =
min(pvalue))
        finalResult1$padj = p.adjust(finalResult1$pvalue, method='fdr')
        finalResult1 = merge(finalResult1, geneName, by='gene_id')
        write.csv(finalResult1, file = paste('../XY/DESeq2',selectLevel,'.csv'))
        return(finalResult1)
  }


##switch level  ##function for discrete variables ##number coded
switch_level <- function(df, nlevel, cortest){
        orders = permn(1:nlevel, fun = function(x){
                df$newlevel = as.numeric(as.character(mapvalues(df[,3], from=
levels(df[,3]), to = x)))
                test = cor.test(df[,4], df[,2], method = cortest)
                return(c(pvalue = test$p.value, rvalue = test$estimate))
                })
        orders = matrix(unlist(orders), ncol=2, byrow=T)
        pvalue = min(orders[,1])
        return(list(p.value = pvalue, estimate = max(orders[orders[,1]
==pvalue ,2])))
}


##function for discrete variables ##binary coded
binary_switch <- function(df, nlevel, cortest){
          STAT = NULL
          for (i in 1:nlevel){
                    df$numfac = as.numeric(df[,3])
                    df$numfac[df$numfac != i] = 0
                    df$numfac[df$numfac == i] = 1
                    test = cor.test(df[,4], df[,2], method = cortest)
            STAT = rbind(STAT, c(pvalue = test$p.value, rvalue = test$estimate,
level1 = i))
          }
        pvalue = min(STAT[,1])
        return(list(p.value = pvalue, estimate = max(STAT[STAT[,1]==pvalue,2]),
           level1 = STAT[STAT[,1]==pvalue,3] ))
}

metaRNA = read.csv('../../initialPaper01r/metaRNA.csv')    ##149 SAMPLE
geneName =
unique(read.csv('../../generateDictionary/nameDictionary_RNA_barrick.csv'))

##main function
compute_correlation <- function(dataframe,
                         selectLevel,
                         x.log = F,
                         binary = F,
                         cortest = 'pearson'){
          metaRNA_selectLevel = dataframe[, colnames(dataframe) %in%
c('dataSet', selectLevel)]

  TESTSTAT = NULL
  nlevel = length(unique(metaRNA_selectLevel[,2]))
  for (each_gene in 1:nrow(finalData)){
    one_gene = merge(data.frame(count = finalData[each_gene,], dataSet =
colnames(finalData)),
                 metaRNA_selectLevel, by = 'dataSet')      ##total 143
    if (selectLevel %in% c('carbonSource','Na_mM_Levels',
'growthPhase','Mg_mM_Levels')) {
      if (cortest == 'glm'){
        test = summary(lm(one_gene[,2] ~ one_gene[,3]))
        test = list(p.value = test$coefficients[2,4], estimate = sqrt(test$r.squared))
      }
```

```
   else if (binary){
    test = binary_switch(one_gene, nlevel, cortest)
   }
   else{test = switch_level(one_gene, nlevel, cortest)}
  }
  else{
   if (x.log == T) one_gene[,3] = log(one_gene[,3])
   test = cor.test(one_gene[,3], one_gene[,2], method = cortest)
  }
  TESTSTAT = rbind(TESTSTAT, c(pvalue = test$p.value, rvalue =
test$estimate))
 }
     return(TESTSTAT)
}

##Bayesisn Regression Model
MCMC_continuous <- function(iteration, one_gene, testdata, method = 'test'){
 #observations
 X = one_gene[,2]
 Y = log(one_gene[,3])
 n = length(X)

 #initial values
 u0 = 0; tau02 = 1    #beta0 prior hyperparameter
 u1 = 0; tau12 = 1    #beta1 prior hyperparameter
 a = 1; b = 1         #sigma2 prior hyperparameter
 beta0 = 0; beta1 = 0; sigma2 = 1

 PARAMETERS = c(beta0, beta1, sigma2)
 for (i in 1:iteration){
  ###update posteriors

  #update sigma2
  sigma2 = 1/rgamma(1, a+n/2, rate = b+ sum((Y-beta0-beta1*X)^2)/2)

  #update beta0, beta1
  C0n = 1/(1/tau02 +n/sigma2)
  m0n = (u0/tau02 + sum(Y-beta1*X)/sigma2) *C0n
  beta0 = rnorm(1, mean = m0n, sd = sqrt(C0n))

  C1n = 1/(1/tau12 + sum(X^2)/sigma2)
  m1n = (u1/tau12 + sum(X*(Y-beta0))/sigma2) * C1n
  beta1 = rnorm(1, mean = m1n, sd = sqrt(C1n))

  PARAMETERS = rbind(PARAMETERS, c(beta0, beta1, sigma2))
 }
```

```
  ##plot convergence parameters
  #pdf(file=paste('../XY/Convergence_Parameter_',selectLevel,'.pdf', sep=''))
  #par(mfrow=c(ncol(PARAMETERS),1))
  #for (i in 1:ncol(PARAMETERS)){plot(PARAMETERS[,i], type = 'l')}
  #dev.off()

  exp_parameter = colMeans(PARAMETERS)
  if (method =='grid'){
    ##Grid evaluation of likelihood
    lowbound = max(min(one_gene$count)- 2*exp_parameter[3], -1.4)
    highbound = max(one_gene$count)+2*exp_parameter[3]
    sample_points = seq(lowbound, highbound, length = 100)
    likelihood = dnorm(sample_points, mean = sample_points, sd =
sqrt(exp_parameter[3]))
  }
  else{
    ##testing the trained model
    likelihood = dnorm(log(testdata[,3]), mean = exp_parameter[1] +
testdata[,2]*exp_parameter[2], sd = sqrt(exp_parameter[3]), log = T)
  }
  return(list(likelihood = sum(likelihood), par = exp_parameter))
}

##Gaussian mixture model
MCMC_discrete <- function(iteration, one_gene){

  ##Generate summary statistics
  one_gene_stat = one_gene %>% group_by(one_gene[,3]) %>%
dplyr::summarize(avg = mean(count), std = sd(count), n = n())

  ##Set initial values for parameters and hyperparameter
  nlevel = nrow(one_gene_stat)
  mu = numeric(length = nlevel)   ####one_gene_stat$meanMg
  #constant variance/ non-constant variance
  #sigma2 = 0.01
  sigma2 = numeric(length = nlevel) + 0.01
  alpha = numeric(length = nlevel)       ###(one_gene_stat$nMg)/10
  C0 = 1
  n = sum(one_gene_stat$n)

  new_label = data.frame(y = one_gene$count, z = as.numeric(one_gene[,3]))

  PARAMETERS = c(mu, alpha, sigma2)
  Z = new_label$z
  for (i in 1:iteration){
    new_stat = new_label %>% group_by(z) %>% dplyr::summarize(avg =
mean(y), std = sd(y), n = n())

    ##posterior distributions
    C1 = 1/(1/C0 + new_stat$n/sigma2)
    mu0n = (0/C0 + new_stat$n * new_stat$avg/sigma2)*C1
    #update mu
    mu = rnorm(nlevel, mean = mu0n, sd = sqrt(C1))

    #update omega
    omega = rdirichlet(1, alpha+ new_stat$n)

    ##update sigma2
    sigma2 = 1/rgamma(nlevel, shape = new_stat$n/2 +1, rate =
new_stat$n*(new_stat$std^2)/2)
    sigma2[is.nan(sigma2)] = 0.0000001

    ##update z
    for (i in 1:n){
      probs = dnorm(new_label$y[i], mean = mu, sd = sqrt(sigma2))
      #new_label$z[i] = base::sample(1:nlevel, 1, prob = probs)
      new_label$z[i] = which(probs == max(probs))
    }

    ##post-processing to corre
    Z = cbind(Z, new_label$z)
    PARAMETERS = rbind(PARAMETERS, c(mu, omega, sigma2))
  }
  Z = Z[, (iteration-200):iteration]
  z = apply(Z, 1, FUN = function(x){names(which.max(table(x)))})
  counttable = table(one_gene[,3], z)
  counttable = counttable/rowSums(counttable)

  rowmax = unlist(apply(counttable,1, FUN = function(x) which(x ==max(x))[1]))
  probs = sum(counttable[cbind(1:nlevel, rowmax)])
  if (length(unique(rowmax)) ==1) {probs = 0}

  return(probs)
}
```

References

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Ali Azam, T., Iwata, A., Nishimura, A., Ueda, S., and Ishihama, A. (1999). Growth phase-dependent variation in protein composition of the Escherichia coli nucleoid. J. Bacteriol. *181*, 6361–6370.

Brückner, R., and Titgemeyer, F. (2002). Carbon catabolite repression in bacteria: choice of the carbon source and autoregulatory limitation of sugar utilization. FEMS Microbiol. Lett. *209*, 141–148.

Chuang, S.E., Daniels, D.L., and Blattner, F.R. (1993). Global regulation of gene expression in Escherichia coli. J. Bacteriol. *175*, 2026–2036.

Crasnier, M. (1996). Cyclic AMP and catabolite repression. Res. Microbiol. *147*, 479–482.

Eguchi, Y., Okada, T., Minagawa, S., Oshima, T., Mori, H., Yamamoto, K., Ishihama, A., and Utsumi, R. (2004). Signal transduction cascade between EvgA/EvgS and PhoP/PhoQ two-component systems of Escherichia coli. J. Bacteriol. *186*, 3006–3014.

François Jacob, J.M. (1961). Jacob F, Monod J.. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3: 318-356. J. Mol. Biol. *3*, 318–356.

Hengge-aronis, R. (1996). Regulatory gene expression during entry into stationary phase. ResearchGate *2*.

Higgins, C.F., Dorman, C.J., Stirling, D.A., Waddell, L., Booth, I.R., May, G., and Bremer, E. (1988). A physiological role for DNA supercoiling in the osmotic regulation of gene expression in S. typhimurium and E. coli. Cell *52*, 569–584.

Jenkins, D.E., Chaisson, S.A., and Matin, A. (1990). Starvation-induced cross protection against osmotic challenge in Escherichia coli. J. Bacteriol. *172*, 2779–2781.

Kato, A., Tanabe, H., and Utsumi, R. (1999). Molecular characterization of the PhoP-PhoQ two-component system in Escherichia coli K-12: identification of extracellular Mg2+-responsive promoters. J. Bacteriol. *181*, 5516–5520.

Kolter, R., Siegele, D.A., and Tormo, A. (1993). The stationary phase of the bacterial life cycle. Annu. Rev. Microbiol. *47*, 855–874.

Liu, M., Durfee, T., Cabrera, J.E., Zhao, K., Jin, D.J., and Blattner, F.R. (2005). Global transcriptional programs reveal a carbon source foraging strategy by Escherichia coli. J. Biol. Chem. *280*, 15921–15927.

Minagawa, S., Ogasawara, H., Kato, A., Yamamoto, K., Eguchi, Y., Oshima, T., Mori, H., Ishihama, A., and Utsumi, R. (2003). Identification and molecular characterization of the Mg2+ stimulon of Escherichia coli. J. Bacteriol. *185*, 3696–3702.

Purvis, J.E., Yomano, L.P., and Ingram, L.O. (2005). Enhanced trehalose production improves growth of Escherichia coli under osmotic stress. Appl. Environ. Microbiol. *71*, 3761–3769.

Reed, J.L., Vo, T.D., Schilling, C.H., and Palsson, B.O. (2003). An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol. *4*, R54.

Saier, M.H. (1998). Multiple mechanisms controlling carbon metabolism in bacteria. Biotechnol. Bioeng. *58*, 170–174.

Siegele, D.A., and Kolter, R. (1992). Life after log. J. Bacteriol. *174*, 345–348.

Soncini, F.C., and Groisman, E.A. (1996). Two-component regulatory systems can interact to process multiple environmental signals. J. Bacteriol. *178*, 6796–6801.

Trentini, F., Ji, Y., Iwamoto, T., Qi, Y., Pusztai, L., and Müller, P. (2013). Bayesian Mixture Models for Assessment of Gene Differential Behaviour and Prediction of pCR through the Integration of Copy Number and Gene Expression Data. PLOS ONE *8*, e68071.

Weber, A., and Jung, K. (2002). Profiling early osmostress-dependent gene expression in Escherichia coli using DNA macroarrays. J. Bacteriol. *184*, 5502–5507.

Weber, A., Kögl, S.A., and Jung, K. (2006). Time-dependent proteome alterations under osmotic stress during aerobic and anaerobic growth in Escherichia coli. J. Bacteriol. *188*, 7165–7175.

Weber, H., Polen, T., Heuveling, J., Wendisch, V.F., and Hengge, R. (2005). Genome-Wide Analysis of the General Stress Response Network in Escherichia coli: σS-Dependent Genes, Promoters, and Sigma Factor Selectivity. J. Bacteriol. *187*, 1591–1603.