A Generalized Poisson Model for Gene Expression Profiling using RNA Sequence Data

A Thesis Paper
Submitted to the Graduate School
in Partial Fulfillment of the Requirements
for the Degree
Master of Science
by
Paromita S Nitu
Dr. Munni Begum - Advisor

Ball State University
Muncie, Indiana
May, 2016
# Contents

1 Acknowledgment 3  

2 Abstract 4  

3 Introduction 5  

4 Literature review 6  

5 Methodology 9  

5.1 A Generalized Poisson Model 9  

5.2 Parameter Estimation of Generalized Poisson Model 10  

5.3 Methods to Identify Differentially Expressed Genes 13  

5.4 Comparison with other models using R-packages 15  

5.4.1 Review of methods used to detect overdispersion in genes, edgeR 16  

5.4.2 Review of methods to detect DE genes, DESeq 17  

5.5 Summary comparison of the edgeR and DESeq 20  

6 Data 21  

6.1 Sequence Count Data 21  

6.2 RNA-Seq count data of Arabidopsis by TCC 23  

7 Results 24  

7.1 Exploratory Data Analysis using Principle Component Analysis (PCA) 24  

7.2 Identification of DE genes by Generalized Poisson Model 26
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.3</td>
<td>Comparison study of differentially expressed genes by GP model, edgeR and DESeq</td>
<td>28</td>
</tr>
<tr>
<td>8</td>
<td>Conclusion and Discussion</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>Bibliography</td>
<td>33</td>
</tr>
<tr>
<td>10</td>
<td>Appendix</td>
<td>34</td>
</tr>
</tbody>
</table>
1 Acknowledgment

I would like to take this opportunity to express my sincere gratitude to my thesis supervisor Dr. Munni Begum for her continuous support, guidance, patience, motivation and immense knowledge. During thesis period, I had to go though a major physical constraint and she facilitated all possible help at her best. Without her consideration and support, it would be quite impossible for me to execute this task. Her insights and words of encouragement have inspired me to complete my thesis work. I also owe deepest gratitude to my thesis committee members Dr. Michael Karls and Dr. Yayuan Xiao for their efforts and contributions to this work.
2 Abstract

An RNA-Seq experiment is a useful tool in characterizing and quantifying transcriptomes into read counts and identifying differentially expressed (DE) genes under different treatment conditions. However, analyzing RNA-Seq data in the quest of differentially expressed genes is not straightforward. Focusing on the experiment of interest, different approaches can be made in identifying DE genes. Here, we propose to use a two parameter generalized poisson (GP) model to address the non-uniformity of read counts than the traditional poisson model and apply it to Arabidopsis pilot survey data by TCC(http://bioconductor.org/packages/release/bioc/html/TCC.html). A comparison study has also been performed with built in R-packages edgeR and DESeq with their default settings to understand the performance of GP model. Here, 28 new differentially expressed genes have been identified by GP model more than edgeR and DESeq for Arabidopsis data and these genes can be a potential source of information in treating bacterial infection to the experimenters. Therefore, the approach of using GP model in real data set evident a significant performance to the in built methods of R-packages.
3 Introduction

A successful application of deep sequencing to the biological specimen generates RNA-Seq data. The established next generation sequencing platforms produces millions of reads. These procreated reads from the experiment are typically being mapped to reference genomes. Hence, the observed number of mapped reads within genes is used as an approximation of tag abundance.

A common assumption in quantifying transcript abundance in RNA-Seq experiment is that the fragments are uniformly sampled from the transcriptome of an organism and hence the basis for Poisson distributional assumption for read counts. However, there is ample evidence of non-uniformity in fragment distributions in RNA-Seq experiments and hence a Poisson distribution for read counts may not be a tenable one. Major criticisms about the Poisson distribution for modeling read counts are that it has only one parameter that represents both mean and variance and the assumption of equal mean and variance is violated often times. To address this issue a simple yet powerful statistical tool that can be used, is the generalized Poisson model with two parameters to address uniformity and non-uniformity of distribution distinctively.

An application to RNA-Seq data could be an interesting study as the data holds concerns to be handled with generalized Poisson model. This model can fit a RNA-Seq data in a more suitable way than Poisson model with one parameter which often fails to address the huge variation in data set on a Negative Binomial model (NBM) when the data contains excessive proportion of zeros.
In estimating the parameters, that is the rate parameter $\theta$ indicating the mean number of gene counts mapped in a certain region of a reference genome or transcriptome and the dispersion parameter (DP) $\lambda$, the maximum likelihood estimation (MLE) method is applied.

Comparing the mean parameter for each gene obtained from sample under two different conditions, RNA-Seq data will allow us to identify differentially expressed genes. A proper identification of differentially expressed genes also mirrors the goodness of fitting a generalized Poisson model to the RNA-Seq data. So, the advantage of generalized Poisson model with an extra parameter to profile gene expression can be assessed precisely by using RNA-Seq data.

In this study, we will apply the generalized Poisson model (GPM) to an RNA-seq data to estimate the dispersion parameter $\lambda$ by MLE and develop a test procedure for testing differentially expressed genes. A pairwise test for the average abundance of transcripts of sample under treatment and sample under control, $\theta_c$ and $\theta_t$ is conducted to identify differentially expressed genes. In addition a comparison of the test of differentially expressed genes will be performed by using GPM and NBM.

4 Literature review

In this section, we review commonly used methodologies for testing differentially expressed genes under two conditions. The moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2 paper by author Michael I Love$^{1,2,3}$, Wolfgang Huber$^1$, Simon Anders$^{1,*}$ talks about the sta-
statistical package DESeq in R program which uses shrinkage estimation for
dispersion and fold changes to improve stability and interpretability of the
estimates from the data with small replicate numbers, discreteness, large
range, presence of outliers. DESeq method also detects and corrects disper-
sion estimate that are too low through modeling of the dispersion on the
average expression strength over all samples. The original paper of DESeq
by Robinson and Smyth [2] proposed a method to figure out digital gene
expression when the replication number is small by introducing a weighted
conditional likelihood estimator.

In [4], Robinson, McCarthy, Smyth talks about a R-package involving an
empirical Baysian method incorporated with general Poisson to detect DE
genes in digital gene expression data which can account biological variability
even with a single or two replicate sample in edgeR: a Bioconductor package
for differential expression analysis of digital gene expression data.

TCC; an R package for comparing tag count data with robust normaliza-
multi-step normalization(TBT) consisting three steps of data normalization
using trimmed mean method(TMM), this is followed by DEG identification
using empirical Baysian methods and comparing tag counts for two group
RNA-Seq data with replication as it helps detecting DE genes in shorter
period of time using statistical methods available in R packages.

In 2010, the authors of “A two-parameter generalized Poisson model
to improve the analysis of RNA-seq data”, Sundeep Srivastava and Liang
Chen* found that a Generalized Poisson (GP) model with another param-
eter to address the variability fits the RNA-Seq data much better than the
traditional Poisson model. Thus goodness-of-fit studies also support the usefulness of using GP distribution rather than Poisson model. They also show that the estimation of gene expression levels and identification of differentially expressed genes, can be done more appropriately by two parameter GP model than by the usual Poisson distribution.

In the same year, Anders and Huber [8] proposed an alternative way of addressing the huge variation with small number of replication in data set of interest in “Differential expression analysis for sequence count data”. They considered statistical testing to decide whether a given gene has significant difference in observed read counts than what can be expected. Although, these read counts usually follow multinomial distribution, approximated by Poison distribution (PD). Since, using PD is too restrictive due to the equality assumption of mean and variance; it is only able to predict small variation. To address this dispersion problem, the authors of this paper proposed to use Negative Binomial Distribution (NBD) which has two distinct parameters to uniquely determine the mean $\mu$ and variance $\sigma^2$ linked up by local regression. In this paper, they also present an implementation using DESeq, R/Bioconductor package.

A study to compare the effectiveness of GP Regression model and NB Regression model has been conducted by Melliana, Setyorini, Eko, Rosi and Purhadi in “The Comparison of Generalized Poisson Regression and Negative Binomial Regression Methods in Overcoming Overdispersion” [17] by using cervical cancer data. They considered estimating the rate parameter $\theta$ and dispersion parameter $\lambda$ of the GP model by Newton-Raphson iteration. In this study they compared Akaike Information Criterion (AIC) values ob-
tain from these two models and concluded that NB model is comparatively better as it came up with a smaller AIC value.

Authors Naim and Begum proposed to study differentially expressed genes in RNA-Seq data by using standard poisson and NB model considering two sets of genes: overdiapersed and non-overdiapersed in “A Two-Step Intregrated Approach To Detect Differentially Expressed Genes In RNA-Seq Data” [6], which is currently under review with the Journal of Bioinformatics and Computational Biology. They studied the performance of this new method combined with existing R-packages in both simulated and real life data and observed better performance by splitting the data into two parts.

5 Methodology

5.1 A Generalized Poisson Model

Let, \( x \sim \) count data and the sum of \( x_i \):

\[
S = \sum_{i=1}^{n} x_i
\]

in the total read count.

Then, \( x \sim \) Generalized Poisson Model (GPD) with rate parameter \( \theta \) and dispersion parameter \( \lambda \).

\[
P(X = x) = \begin{cases} 
\frac{\theta(\theta + x\lambda)^{x-1}\exp^{-\theta-x\lambda}/x!}{}, & \text{if } x = 1, 2, 3, ... \\
0, & \text{otherwise}
\end{cases}
\]
Here, $x = 0, 1, 2, \ldots, n$ for $x > q$ if $\lambda < 0$ where $\theta > 0$ and $\max(-1, -\theta/q) \leq \lambda \leq 1$ and $q \geq 4$, is the largest positive integer, so that $\theta + q\lambda > 0$ when $\lambda < 0$.

We impose a lower limit to $\lambda$ and chose $q \geq 4$ to ensure that, there are at least 5 classes with non-zero probabilities and truncation errors \[\sum_{x_i=0}^{\infty} P(X = x)\] is a little less than 1] do not affect practical application.

The likelihood function of Generalized Poisson model is:

$$L(\theta, \lambda) = \theta^n \prod_{x_i=0}^{n} (\theta + x_i \lambda)^{x_i - 1} \exp\left(-n\theta - \lambda \sum_{x_i=0}^{n} \frac{x_i}{x_i!}\right)$$

$$= \exp\left[-n\theta - \lambda \sum_{x_i=0}^{n} x_i + n \log \theta + \sum_{x_i=0}^{n} (x_i - 1) \log(\theta + x_i \lambda) - \sum_{x_i=0}^{n} \log x_i\right]\]$$

### 5.2 Parameter Estimation of Generalized Poisson Model

The rate parameter $\theta$ and the dispersion parameter $\lambda$ of GPD have linear functional form. So, to find out the mean and the variance of GDP, we use a special method.

$$S(k, \theta, \lambda) = \sum_{x=0}^{\infty} \frac{\theta(x + x \lambda)^{x+k-1}}{x!} \exp^{-\theta - x \lambda} / x! \quad (1)$$

with $k = 0, 1, 2, 3, \ldots$

For $k = 0$, it reduces to GPD. We can rewrite (1) as-
\[ S(k, \theta, \lambda) = \sum_{x=0}^{\infty} \theta(\theta + x\lambda)^{x+k-2} \exp\left(-\theta - x\lambda\right) + \sum_{x=1}^{\infty} \lambda(\theta + x\lambda)^{x+k-2} \exp\left(-\theta - x\lambda\right) \frac{1}{(x-1)!} \]

which gives a recurrent relation;

\[ S(k, \theta, \lambda) = \theta(S(k-1, \theta, \lambda)) + \lambda(S(k + \lambda, \lambda)) \]  \hspace{1cm} (3)

By the repeated use of above recurrence, we get-

\[ S(k, \theta, \lambda) = \sum_{x=0}^{\infty} \lambda^x(\theta + x\lambda)S(k-1, \theta + x\lambda, \lambda); \]  \hspace{1cm} (4)

with \( k = 0, 1, 2, 3, \ldots \)

From the definition, \( \theta(S(k, \theta, \lambda)) = 1 \)

For \( k = 1 \)

\[ S(1, \theta, \lambda) = \sum_{x=0}^{\infty} \lambda^x = \frac{1}{1 - \lambda} \]  \hspace{1cm} (5)

For \( k = 2 \), (4)and (5);

\[ S(2, \theta, \lambda) = \sum_{x=0}^{\infty} \lambda^x(\theta + x\lambda)(1 - \lambda)^{-1} \]
\[ = \theta(1 - \lambda)^{-2} + \lambda(1 - \lambda)^{-3} \]
So, the mean is

\[ \mu_1' = \theta \sum_{x=1}^{\infty} \frac{x(\theta + x\lambda)^{(x-1)}}{x!} \exp^{-\theta-x\lambda} \]

\[ = \theta \sum_{x=0}^{\infty} \frac{x(\theta + \lambda + x\lambda)^{x}}{x!} \exp^{-\theta-\lambda-x\lambda} \]

\[ = \theta S(1, \theta + \lambda, \lambda), \text{by (1)} \]

\[ = \theta(1 - \lambda)^{-1}, \text{by (5)} \]

Here, the second moment is

\[ \mu_2' = \theta[S(2, \theta + 2\lambda, \lambda) + S(1, \theta + \lambda, \lambda)] \]

From equation (1) and (2), we can write-

\[ \mu_2' = \frac{\theta}{(1 - \lambda)^{3}} + \frac{\theta^2}{(1 - \lambda)^{2}} \]

\[ = \frac{\theta + \theta^2(1 - \lambda)}{(1 - \lambda)^{3}} \]

Therefore,

\[ \sigma^2 = \mu_2 \]

\[ = \mu_2' - (\mu_1')^2 \]

\[ = \frac{\theta + \theta^2 - \lambda\theta^2}{(1 - \lambda)^{3}} + \frac{\theta^2}{(1 - \lambda)^{2}} \]

\[ = \frac{\theta + \theta^2 - \lambda\theta^2 - \theta^2 + \lambda\theta^2}{(1 - \lambda)^{3}} \]

\[ = \theta(1 - \lambda)^{-3} \]

13
The MLE of the parameters is

$$\hat{\theta} = \bar{x}(1 - \lambda), \text{where } \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The MLE $\hat{\theta}$ of $\theta$ can be obtained by solving the following equation using the Newton-Raphson method:

$$\sum_{i=1}^{n} \frac{x_i(1 - x_i)}{\bar{x} + (x_i - \bar{x})\lambda} - n\bar{x} = 0$$

5.3 Methods to Identify Differentially Expressed Genes

We can use a likelihood ratio test to detect those genes that are differentially expressed. Let us consider, $X$ be the read counts of genes in sample 1 with three replication and in this case, the treatment condition is treated. Similarly $Y$ be the read counts for sample 2 with treatment condition control. To estimate MLE's,

We have,

$$P(X = x) = \theta_1(\theta_1 + x\lambda_1) \exp^{-\theta_1 - x\lambda_1} / x!,$$

$$P(Y = y) = \theta_2(\theta_2 + y\lambda_2) \exp^{-\theta_2 - y\lambda_2} / y!,$$

Where $(\theta_1, \lambda_1)$ and $(\theta_2, \lambda_2)$ can be estimated by the maximum likelihood approach. Under the null hypothesis of equality of the rate parameters stating no evidence of differential expression under different treatment con-
ditions, we can rewrite the GP models as-

\[ P(X = x) = \theta (\theta + x\lambda_1)^{x-1} \exp^{-\theta - x\lambda_1} / x!, \]
\[ P(Y = y) = w\theta (\theta + y\lambda_2)^{y-1} \exp^{-w\theta - y\lambda_2} / y!, \]

Here, \( w \) is the scaling factor for comparing two sample under two different treatment conditions and we can obtain it as-

\[ w = \frac{s_2}{s_1} \text{ where, } s_2 = \sum_{g=1}^{G} \hat{\theta}_{1,g} l_g, \hat{\theta}_1 \text{ and } s_1 = \sum_{g=1}^{G} \hat{\theta}_{2,g} l_g. \]

Here, \( \hat{\theta}_1 \) is the MLE of \( \theta \) for gene \( g \) of sample 1 and \( \hat{\theta}_2 \) is the MLE of \( \theta \) for gene \( g \) of sample 2 and \( l_g \) is the length of \( g \) which we can get from the normalization performance.

For individual dispersion parameter \( \hat{\lambda}_1 \) and \( \hat{\lambda}_2 \), we can solve the following equation by Newton-Raphson method and obtain \( \hat{\lambda}_1 \) and \( \hat{\lambda}_2 \).

\[ \frac{2n}{\theta} - (n + nw) + \sum_{i=1}^{n} \frac{(x_i - 1)}{\theta} + x_i \hat{\lambda}_1 + \sum_{i=1}^{n} \frac{(y_i - 1)w}{w\theta + y_i\lambda_2} = 0 \]

The log-likelihood ratio test statistics can be calculated as:

\[ T = -2\ln\left( \frac{L(\hat{\theta}, \hat{\lambda}_1, \hat{\lambda}_2 | x, y)}{L(\hat{\theta}_1, \hat{\theta}_2, \lambda_1, \lambda_2 | x, y)} \right) \]

Assuming the null model is true, test statistics \( T \) is approximately chi-
square distributed with one degree of freedom. From the unrestricted model of Poisson model we can compare the gene expressions of these samples to detect differentially expressed genes. Here,

\[
P(X = x) = \frac{\theta_1^x \exp^{-x}}{x!} \text{ and } P(Y = y) = \frac{\theta_2^y \exp^{-y}}{y!}
\]

Thus, the log-likelihood ratio test statistics can be calculated as:

\[
T = -2 \ln\left( \frac{L(\hat{\theta}|x, y)}{L(\hat{\theta}_1, \hat{\theta}_2, |x, y)} \right)
\]

Which follows a chi-square distribution with one degree of freedom when null hypothesis is true. In the Generalized Poisson model, the log-likelihood ratio is used to identify differentially expressed gene with normalization factor \(w\) to estimate parameters involving biological replication of the samples.

### 5.4 Comparison with other models using R-packages

In this study of identifying differentially expressed genes, we are using built-in R-packages from Biocoductor namely edgeR, DESeq with the motivation of negative binomial model and normalization method TCC with robust strategies in three steps using trimmed mean method. The author of edgeR and DESeq considered different approaches to estimate the dispersion parameter \(\lambda\) and developed environment to detect differentially expressed genes by either using an exact test method or generalized linear model approach.
In comparison of Negative Binomial and Poisson model, we set-up the parameters for RNA-seq data. Let, \( y \) be the gene counts of the \( j^{th} \) replication of the \( i^{th} \) treatment and it can be written as-

\[
Y_{i,j} \sim NB(\mu_i, \varphi)
\]

where, \( \mu_i \) is the rate parameter of \( i^{th} \) gene and \( \varphi \) is the dispersion parameter.

The conditional distribution of \( y \) on \( \lambda \) is a Poisson distribution where \( y \) belongs to a negative binomial distribution and \( \lambda \) follows gamma distribution. Which can be written as follows-

\[
Y|\lambda \sim \text{Poisson}(\lambda) \text{ and } \lambda \sim G(\varphi^{-1}, \mu^{-1}\varphi^{-1})
\]

We got the marginal distribution function of \( y \) by integrating \( \lambda \), we get-

\[
P(Y = y) = \left((y + \varphi^{-1} - 1) \right) (\varphi^{-1} - 1) \left(\frac{1}{1 + \mu \varphi}\right)^{\varphi^{-1}} \left(\frac{\mu}{\mu + \varphi^{-1}}\right)^{y}
\]

5.4.1 Review of methods used to detect overdispersion in genes, edgeR

The well-established Bioconductor package edgeR has been, proposed by Robinson and Smyth[2] to detect DE genes in RNA-Seq data. Here, a com-
mon dispersion is assumed for all the genes in order to obtain a better estimate of dispersion parameter. This can be altered as it may not be the case all the time and often prove to be too restrictive on the model, and this provides us the opportunity of farther development of this method and allows a study for estimating individual estimates of dispersion parameters for each gene. This extension also allows us to individual estimation and the squeezing of each estimate towards a common dispersion by using the weighted likelihood approach. The built in model for over-dispersed RNA-Seq data on edgeR follows a NB distribution with the count model defined as follows:

\[ Y_{tr,g} \sim NB(M_{tr} p_{tg}, \varphi_g) \]

In this model, \( M_{tr} \) is the size of the library for replicate \( r \) in treatment group \( t \), \( \varphi_g \) is the dispersion parameter, and \( p_{tg} \) is the gene-wise relative abundance for treatment group \( t \). The mean of this NB distribution is \( M_{tr} p_{tg} \) and the variance is \( M_{tr} p_{tg} (1 + M_{tr} p_{tg} \varphi_g) \). TMM normalization is used in this package. The authors [3] considered \( p_{tg} \) as their parameter of interest for DE analysis. When \( \varphi_g = 0 \), the NB model reduces to a Poisson model.

5.4.2 Review of methods to detect DE genes, DESeq

In contrast to edgeR, Anders and Huber proposed DESeq by using a mean-variance relationship of the expression levels and expressed the variances as functions of the mean.
This relationship can be modeled as a local polynomial function $\sigma^2$ and it is:

$$\hat{\sigma}^2_{tg} = s_{tr}\hat{\mu}_{tg} + s_{tr}^2\mu_{tg}^2\left(a_0 + \frac{a_1}{\mu_{tg}}\right)$$

With $a_0, a_1 > 0$ and $s_{tr}$ is defined to be the size factor for normalizing the read counts. Anders and Huber chose the following NB distribution for the read counts:

$$Y_{tr,g} \sim NB(M_{tg}, \sigma^2)$$

Where, $\mu_{tg}$ and $\sigma^2$ are the mean and variance of gene $g$ in the $t$ treatment group respectively. For NB distribution, the variance is a function of mean and dispersion parameter which can be written as:

$$\sigma^2_{tg} = \mu_{tg}(1 + \mu_{tg}\varphi_g)$$

By pooling the replicates towards the treatment groups, we can estimate the gene-wise variance.
\[ \hat{\sigma}_g^2 = \frac{\sum_t \sum_r (\frac{y_{trg}}{s_{tr}} - \hat{\mu}_g)^2}{\sum_t r_t - 1} \]

\[ \hat{\mu}_g = \frac{\sum_t \sum_r \frac{y_{trg}}{s_{tr}}^2}{\sum_t r_t} \]

Then using these estimates of mean and variance, \( \varphi_g \) can be estimated.

Fisher’s exact test is used in DESeq for testing the null hypothesis of the equality of mean read counts in the two treatment groups.
5.5 Summary comparison of the edgeR and DESeq

Table a. Two R packages for differential expression analysis of RNA-Seq data under two treatment conditions.

<table>
<thead>
<tr>
<th>Package</th>
<th>Distribution of counts</th>
<th>Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>edgeR¹⁹,²⁰,¹⁷</td>
<td>[ Y_{gt} \sim NB(M_{tr}p_{gt}, \varphi_g) ]</td>
<td>Exact or generalized linear model (GLM) based test where the parameter of interest is the relative abundance ( p_{gt} ). A weighted conditional likelihood approach is carried out to estimate the per gene dispersion parameter ( \varphi_g ). Here, ( M_{tr} ) denotes the effective library size.</td>
</tr>
<tr>
<td>DESeq²¹³</td>
<td>[ Y_{gt} \sim NB(\mu_{gt}, \varphi_g) ]</td>
<td>Exact or GLM based test for testing equality of mean read counts ( (\mu_{gt}) ) in different conditions. Dispersion estimated parameter ( \varphi_g ) is by the maximum likelihood and then gene-wise dispersion is shrunk through an empirical Bayes approach.</td>
</tr>
</tbody>
</table>
6 Data

6.1 Sequence Count Data

High-throughput sequencing (HTS) technologies are widely used to analyze differential expression on the scientific question of interest. Short sequence reads produced from a biological specimen by these technologies are mapped with reference genome to identify the match to provide counts. Most RNA-Seq experiments process purified RNA and tend to convert it into a cDNA. There are established processes such as Illumina GA/HiSeq, SOLiD, Roche 454 to conduct the sequencing process. RNA-Seq experiments generates millions of short reads from one end or from the both ends of cDNA, a number of steps are involved to turn the raw sequence reads into discrete counts data. In the first stage, the reads are mapped to the genome or transcriptome. Based on the experiment interest, the reads are mapped in gene-level, exon-level or transcripton level to get the read counts. The next step is to apply different methods of normalization to the summarized read counts data. The final stage is to identify the differentially expressed genes in the count data using appropriate statistical tests. Based on ranked p-values of associated genes one can make decisions in favor of DE genes.
A flow-chart showing a typical RNA-Seq pipeline from the paper "Form RNA-Seq reads to differential expression results" by Oshtach, Robinson and Young for differentially expressed genes is as follows:

1. **Millions of short reads**
2. **Burrows wheeler transform Bowtie, BWA, SOAP2**
   - Hash tables PerM, SHRiMP, BFAST, ELAND
   - Unmapped reads
   - Map to Junction library created from annotation
   - Map to ‘de novo’ junction library SplitSeek, Tophat, SOAPals
3. **Reads aligned to reference**
   - Summarization
     - By coding sequence
       - By exon
       - By gene span
     - Junction reads
   - Table of counts
4. **Normalization**
   - Between sample TMM, upper quartile
   - Within sample RPKM, quantile
   - Poisson test DEGseg
   - Negative binomial test edgeR, baySeq, DEseq
5. **List of differentially expressed genes**
6.2 RNA-Seq count data of Arabidopsis by TCC

Here, we are using an RNA-Seq data set, obtained from a pilot study on how the Arabidopsis is response to the infection by bacteria (https://www.bioconductor.org/packages/2.13/bioc/manuals/TCC/man/TCC.pdf). They performed an RNA-Seq experiment to get the read counts from three independent biological samples where each of these samples have read counts from two treatment groups. The count matrix of this experiment contains the frequency counts of RNA-Seq reads mapped with reference genes in the database. Thus, we have a count matrix with 6 columns correspondence to three under treatment and other three are under control of independent biological samples.

This data set was primarily imported from the NBPSeq package of R. The experimenter challenged leaves of Arabidopsis with the defense-eliciting δhrcC mutant of Pseudomonas syringae pathovar tomato DC3000. A performance of infiltrated leaves of Arabidopsis with 10mM MgCl2 has been done as a mock inoculation. During the experiment RNA was isolated 7 hours after inoculation, enriched for mRNA and prepared for RNA-Seq. A sequence of one replicate per channel on the Illumina Genome Analyzer (http://www.illumina.com). The length of the RNA-Seq reads can vary in length depending on user preference and the sequencing instrument. The data is derived from a 36-cycle sequencing reaction and trimmed to 25mers. An in-house computational pipeline has been used to process, align, and as-
sign RNA-Seq reads to genes according to a reference data for Arabidopsis.

To conduct this study, the data set has been pulled on from the R-package TCC where TCC developed a multi-step normalization method for two group RNA-Seq data set with replicates. Here, the TCC count matrix is normalized using robust normalization methods with the trimmed mean method.

The first six rows of the RNA-Seq count matrix of the pilot survey of Arabidopsis infection treatment by bacteria:

<table>
<thead>
<tr>
<th>gene_id</th>
<th>G1_rep1</th>
<th>G1_rep2</th>
<th>G1_rep3</th>
<th>G2_rep1</th>
<th>G2_rep2</th>
<th>G2_rep3</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene_1</td>
<td>34</td>
<td>45</td>
<td>122</td>
<td>16</td>
<td>14</td>
<td>29</td>
</tr>
<tr>
<td>gene_2</td>
<td>358</td>
<td>388</td>
<td>22</td>
<td>36</td>
<td>25</td>
<td>68</td>
</tr>
<tr>
<td>gene_3</td>
<td>1144</td>
<td>919</td>
<td>990</td>
<td>374</td>
<td>480</td>
<td>239</td>
</tr>
<tr>
<td>gene_4</td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>18</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>gene_5</td>
<td>98</td>
<td>48</td>
<td>17</td>
<td>1</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>gene_6</td>
<td>296</td>
<td>282</td>
<td>216</td>
<td>86</td>
<td>62</td>
<td>69</td>
</tr>
</tbody>
</table>

7 Results

7.1 Exploratory Data Analysis using Principle Component Analysis (PCA)

Unsupervised learning methods are used to explore data to have a better understanding of the data for further analysis. Principal component method has a great importance to do such type of analysis when there is a huge num-
Performing PCA provides a great opportunity to deal with fewer components which explains information as much as possible. Exploratory data analysis is must for a data with no background information. In this study, we are looking at the PCA to understand, whether it is enough to study this 6 replicate samples into two groups under treatment and control and at the same time, to see a potential difference between these two groups.

Principal component analysis allows us to summarize a large set of correlated components into a smaller number of variables which explains most part of the variability. In other words, PCA visualizes a low dimensional representation of the data that holds as much information as possible. It is a statistical procedure to convert a set of observations into a set of values which are linearly uncorrelated. Here, from the PCA, we can clearly see that, it is sufficient to study these 6 replicates placed into two groups instead of studying them separately.
From above fit we can clearly see that, there are two components that can properly explain the variation. So instead of considering all sample individually, we can talk about two groups of samples to understand effect of treatment and control on the biological reasoning of bacterial infection of Arabidopsis.

### 7.2 Identification of DE genes by Generalized Poisson Model

In this study, we applied the generalized poisson model to the read counts of the Arabidopsis data to identify differentially expressed genes in interest of controlling bacterial infection under a provided treatment by the experimenter.

Here, the rate parameter $\theta$ and the dispersion parameter $\lambda$ have been estimated for each gene separately, by the application of a recurrence relation to the maximum likelihood approach. First, we estimated the overall
dispersion parameter $\lambda$ under the null hypothesis that no effect of treatment difference and estimated the overall rate parameter $\theta$. Since, we have only six observations per genes, estimating individual dispersion parameter under these two treatment condition does not seem to be very worthy. In such situation, we estimated both the treatment and the control rate parameter $\theta_1$ and $\theta_2$ for each gene separately.

Since, the additive property of the poisson distribution allows us to pool the read counts across the lanes with similar sequencing, a likelihood ratio test has been performed to evaluate differentially expressed genes considering the test for rate parameters $\theta_1$ and $\theta_2$.

The Arabidopsis pilot survey data has 1000 read counts. The application of generalized poisson model is able to detect 131 differentially expressed genes. A bonferroni correction has been performed by considering the level of significance as 0.00005 to control the false discovery rate of DE genes.

A complete list of differentially expressed genes identified by GP model is listed in the appendix.
7.3 Comparison study of differentially expressed genes by GP model, edgeR and DESeq

A comparison between GP model and R-package edgeR, in identifying DE genes, is given below:

![Venn diagram of DE genes by GPD and edgeR package](image)

From Figure 4, we can see that, generalized poisson model is able to detect 28 different DE genes than what we can get by using built in R-package edgeR.

Figure 3: Venn diagram of DE genes by GPD and edgeR package

From Figure 4, we can see that, generalized poisson model is able to detect 28 different DE genes than what we can get by using built in R-package edgeR.
A comparison between GP model and R-package DESeq, in identifying DE genes, is given below:

Figure 4: Venn diagram of DE gens by GP distribution and DESeq package
The above diagram portrays the efficiency of using the generalized poisson model in identifying DE genes of Arabidopsis. The GP distribution detects 131 genes that are differentially expressed. In comparison with DESeq package, it is able to identify 40 new DE genes. These genes could be an excellent source of information in terms of treating Arabidopsis from bacteria.
A comparison among GP model, R-package edgeR and DESeq, in identifying DE genes, is given below:

Figure 5: Venn diagram of DE gens by GP distribution, edgeR and DESeq package
From the above venn diagram, we see a significant benefit of using generalized poisson model to identify the potential deferentially expressed genes. GP model alone can identify 131 DE genes and it identifies 40 new genes that are differentially expressed in comparison with DESeq and 28 new genes than edgeR. Studding these genes can open up new reasoning to look at.
8 Conclusion and Discussion

In this thesis, we primarily focused on applying generalized poisson model to address the dispersion issue of the count data from RNA-Seq experiment. Although, there exist a good number of approaches to handle non-uniformity, mostly different normalization methods and methods involving negative binomial distribution and so on, we can use an extension of traditional poisson distribution with an extra parameter to encounter variability.

Here, we applied the generalized poisson model to the Arabidopsis pilot survey data under two treatment conditions to unfold genes those are differentially expressed due to different treatment facilities. This proposed method shows significant efficiency in identifying DE genes. The generalized poisson model, itself is able to identify a dynamic number of genes as DE. In comparison with existing methods, GP model shows potential efforts. Thus using GP model can be proved to be more powerful over existing R-packages with a simulation study.

Being able to incorporate different normalization method is a very useful aspect of generalized poisson model. However, to apply GP model to our Arabidopsis data, we have made an implicit assumption that these observed gene counts are independent of counts of other genes. A violation of this assumption lead to false discovery of differentially expressed genes. Thus an identification of DE gene does not guaranty a reflection of potential treatment result.

To take account of false discovery, we used bonferroni correction by dividing the nominal level of significance by the total number of test conducted.
A more sophisticated approach, such as Benjamin Hochberg false discovery rate, can be executed to get a better performance. In this research, we only considered two treatment conditions. A future study can be done to incorporate more than two conditions to identify differentially expressed genes by applying generalized poisson model.
9 Bibliography


[14] Zhao Yang$^{*1}$, Jams W. Hardin$^2$, Cheryl L. A Ady$^2$, Quang H. Vuong$^3$. 


10 Appendix

Principal Components Analysis

testrawdataandextractingPCs
fromthecorrelationmatrix

\[ \text{fit} = \text{princomp}(\text{hypoData}, \text{cor} = \text{TRUE}) \]

summary(fit)printvarianceaccountedfor
loadings(fit)plloadings
plot(fit, type = "lines")screeplot
apply(hypoData, 2, var)

\[ \text{pr.out} = \text{prcomp}(\text{hypoData}, \text{scale} = \text{T}) \]

biplot(pr.out, scale = 0)

pr.outsdev

\[ \text{pr.var} = \text{pr.outsdev}^2 \]

\[ \text{pve} = \text{pr.var} / \text{sum(pr.var)} \]

plot(pve, xlab = "PrincipalComponent", ylab = "Proportion of Variance Explained", ylim = c(0, 1), type = 'b')plot(cumsum(pve), xlab = "PrincipalComponent", ylab = "Cumulative Proportion of Variance Explained", ylim = c(0, 1), type = 'b')

\[ \text{fitscores the principal components} \]

biplot(fit)

DE genes with built-in TCC normalization by edgeR:
data(hypoData)
head(hypoData)
group = c(1, 1, 1, 2, 2, 2)
tcc = new("TCC", hypoData, group)

35
\textit{tcc\_edgeR} = \text{calcNormFactors}(tcc, \text{norm.method} = \text{"tmm"},
\text{test.method} = \text{"edger"}, \text{iteration} = 3,
\text{FDR} = 0.1, \text{floorPDEG} = 0.05)
\textit{tcc\_edgeR} = \text{estimateDE}(tcc\_edgeR,
\text{test.method} = \text{"edger"}, \text{FDR} = 0.1)
\textit{result\_edgeR} = \text{getResult}(tcc\_edgeR, \text{sort} = \text{TRUE})
\text{E\_res} = \text{result\_edgeR}[1:50,]
\textbf{DE genes with built-in TCC normalization by DESeq:}
data(hypoData)
\text{head(hypoData)}
group= c(1, 1, 1, 2, 2, 2)
tcc= \text{new}("\text{TCC"}, \text{hypoData}, \text{group})
\textit{tcc\_DESeq} = \text{calcNormFactors}(tcc, \text{norm.method} = \text{"tmm"},
\text{test.method} = \text{"deseq"}, \text{iteration} = 3,
\text{FDR} = 0.1, \text{floorPDEG} = 0.05) \textit{tcc\_DESeq} = \text{estimateDE}(tcc\_DESeq,
\text{test.method} = \text{"deseq"},
\text{FDR} = 0.1)
\textit{result\_DESeq} = \text{getResult}(tcc\_DESeq, \text{sort} = \text{TRUE})
\text{D\_res}=\text{result\_DESeq}[1:50,]
\textbf{DE genes of hypoData by R-package edgeR:}
\text{source("http://bioconductor.org/biocLite.R")}
\text{biocLite("edgeR")}
\text{n}
\text{citation("edgeR")}
\text{library(edgeR)}
\text{mydata=hypoData}
\text{write.table(mydata,}
\text{"C:/Users/Nitu/Desktop/mydata.txt", sep="\t")}
\text{x=read.delim("C:/Users/Nitu/Desktop/mydata.txt")}
\text{group=factor(c(1,1,1,2,2,2))}
\textbf{To perform quasi likelihood F-test:}
y=\text{DGEList(counts=x,group=group)}
y=\text{calcNormFactors(y)}
\text{design=model.matrix( group)}
y=\text{estimateDisp(y,design)}
\text{fit=glmQLFit(y,design)}
\text{qlf=glmQLFTest(fit,coef=2)}
\text{topTags(qlf)}
tp1 = topTags(qlf, n=50)
\textbf{To perform likelihood ratio test:}
\text{fit1=glmFit(y,design)}
\text{lrt=glmLRT(fit1,coef=2)}
\text{topTags(lrt)}
tp2 = topTags(lrt, n=50)  
Testing for DE genes: et = exactTest(y) 
tp3 = topTags(et, n=185) 

Venn Diagram Comparison of DE genes by different methods

\[ \text{THRESHOLD} = 0.05 \]

\[ V_{tcc\_edgeR} = \text{subset}(\text{result}_{\text{edgeR}}, \text{p.value} < \text{THRESHOLD}) \]

result was generated from TCC with edgeR

\[ V_{tcc\_DESeq} = \text{subset}(\text{result}_{\text{DESeq}}, \text{p.value} < \text{THRESHOLD}) \]

result was generated from TCC with DESeq

\[ V_{tcc\_baySeq} = \text{subset}(\text{result}_{\text{baySeq}}, \text{p.value} < \text{THRESHOLD}) \]

result was generated from TCC with baySeq

\[ V_{edgeR} = \text{subset}(et, \text{et}\_PValue < \text{THRESHOLD}) \]

et was generated from edgeR

library(VennDiagram)

pdf("venn_diagram.pdf")

venn.plot_ED = venn.diagram(list(V_{tcc\_edgeR}$gene_id, 
V_{tcc\_DESeq}$gene_id), NULL, fill=c("blue", "green"), 
alpha=c(0.5,0.5), cex = 2, cat.fontface=4, 
category.names=c("edgeR","DESeq"))
grid.draw(venn.plot_ED)
dev.off()

venn.plot_EB = venn.diagram(list(V_{tcc\_edgeR}$gene_id, V_{tcc\_baySeq}$gene_id) , NULL, fill=c("blue", "red"), alpha=c(0.5,0.5), 
cex = 2, cat.fontface=4, category.names=c("edgeR","baySeq"))
grid.draw(venn.plot_EB)

venn.plot_DB = venn.diagram(list(V_{tcc\_baySeq}$gene_id, 
V_{tcc\_DESeq}$gene_id), NULL, fill=c("red", "green"), 
alpha=c(0.5,0.5), cex = 2, cat.fontface=4, 
category.names=c("baySeq","DESeq"))
grid.draw(venn.plot_DB)

venn.plot_top = venn.diagram(list(E$_r$es$gene_id, 
D$_r$es$gene_id), NULL, fill=c("pink", "yellow"), 
alpha=c(0.5,0.5), cex = 2, cat.fontface=4, 
category.names=c("top\_edgeR","top\_DESeq"))
grid.draw(venn.plot_top)

venn.plot_top = venn.diagram(list(E$_r$es$gene_id,
B$res$gene_id), NULL, fill=c("pink", "gray"),
alpha=c(0.5,0.5), cex = 2, cat.fontface=4,
category.names=c("top.edgeR", "top.baySeq")
grid.draw(venn.plot_top)

venn.plot_top = venn.diagram(list(B$res$gene_id,
D$res$gene_id), NULL, fill=c("gray", "yellow"),
alpha=c(0.5,0.5), cex = 2, cat.fontface=4,
category.names=c("top.baySeq", "top.DESeq")
grid.draw(venn.plot_top)

venn.plot = venn.diagram(list(V_tcc$edgeR$gene_id,
V_tcc$DESeq$gene_id), NULL, fill=c("blue", "green"),
alpha=c(0.5,0.5), cex = 2, cat.fontface=4,
category.names=c("edgeR", "DESeq")
grid.draw(venn.plot)
dev.off()
Table 1: Complete list of differentially expressed genes by GP model

| gene_id | $L(\hat{\theta}|X,Y)$ | $L(\hat{\theta}_1, \hat{\theta}_2|X,Y)$ | $T = \frac{L(\hat{\theta}|X,Y)}{L(\hat{\theta}_1, \hat{\theta}_2|X,Y)}$ | $\chi^2_1$ | Decision |
|---------|----------------------|----------------------|----------------------|----------------------|--------|
| gene_1  | 2.19E+27             | 1.09E-140            | 2.02E+167            | 16.44811             | Significant |
| gene_2  | 4.21E-24             | 5.98E-101            | 7.04E+76             | 16.44811             | Significant |
| gene_3  | 3.97E-17             | 1.25E-211            | 3.19E+194            | 16.44811             | Significant |
| gene_5  | 1.39E+12             | 1.19E-84             | 1.17E+96             | 16.44811             | Significant |
| gene_6  | 1.22E-10             | 1.63E-24             | 7.48E+13             | 16.44811             | Significant |
| gene_8  | 8.51E-20             | 1.17E-61             | 7.27E+41             | 16.44811             | Significant |
| gene_9  | 28999094884          | 1.40E-124            | 2.07E+134            | 16.44811             | Significant |
| gene_11 | 6.53E-16             | 3.74E-46             | 1.74E+30             | 16.44811             | Significant |
| gene_14 | 5.56E-20             | 1.28E-62             | 4.33E+42             | 16.44811             | Significant |
| gene_20 | 1.26E-24             | 3.95E-140            | 3.20E+115            | 16.44811             | Significant |
| gene_23 | 2.92E+06             | 4.64E-32             | 6.30E+37             | 16.44811             | Significant |
| gene_24 | 1.63E-19             | 1.17E-196            | 1.40E+177            | 16.44811             | Significant |
| gene_25 | 71.20321331          | 1.68E+286            | 4.24E+286            | 16.44811             | Significant |
| gene_29 | 6.37E-07             | 6.33E-88             | 1.01E+81             | 16.44811             | Significant |
| gene_31 | 6.37E-07             | 6.33E-88             | 1.01E+81             | 16.44811             | Significant |
| gene_34 | 1.44E+07             | 8.61E-290            | 1.67E+296            | 16.44811             | Significant |
| gene_35 | 3.47E+06             | 4.49E-135            | 7.72E+140            | 16.44811             | Significant |
| gene_38 | 2.40E+08             | 3.33E-122            | 7.22E+205            | 16.44811             | Significant |
| gene_39 | 2.68E-19             | 3.88E-198            | 6.91E+178            | 16.44811             | Significant |
| gene_40 | 1.32E-24             | 1.54E-134            | 8.56E+109            | 16.44811             | Significant |
| gene_46 | 3.67E+01             | 1.47E-00             | 2.50E+01             | 16.44811             | Significant |
| gene_49 | 1.54E+05             | 8.64E-298            | 1.78E+302            | 16.44811             | Significant |
| gene_50 | 1.67E-21             | 1.02E-178            | 1.64E+157            | 16.44811             | Significant |
| gene_52 | 2.58E-17             | 1.98E-49             | 1.30E+32             | 16.44811             | Significant |
| gene_54 | 1.17E-14             | 4.99E-224            | 2.34E+209            | 16.44811             | Significant |
| gene_55 | 6.34E-18             | 1.25E-52             | 5.08E+34             | 16.44811             | Significant |
| gene_56 | 1.04E+95             | 3.38E-188            | 3.08E+282            | 16.44811             | Significant |
| gene_58 | 2.34E-10             | 3.13E-236            | 7.47E+225            | 16.44811             | Significant |
| gene_60 | 2.83E-13             | 3.01E-33             | 9.41E+19             | 16.44811             | Significant |
| gene_62 | 4.68E-11             | 5.84E-35             | 8.01E+23             | 16.44811             | Significant |
| gene_63 | 6.37E-07             | 6.33E-88             | 1.01E+81             | 16.44811             | Significant |
| gene_64 | 6.37E-07             | 6.33E-88             | 1.01E+81             | 16.44811             | Significant |
| gene_67 | 2.30E-01             | 5.47E-87             | 4.20E+85             | 16.44811             | Significant |
| gene_68 | 1.63E-17             | 9.48E-209            | 1.72E+191            | 16.44811             | Significant |
| gene_71 | 2.33E+72             | 7.49E-135            | 3.11E+206            | 16.44811             | Significant |
| gene_73 | 1.38E-24             | 3.20E-118            | 4.32E+93             | 16.44811             | Significant |
| gene_74 | 6.55E-22             | 2.75E-77             | 2.38E+55             | 16.44811             | Significant |
| gene_75 | 1.01E-09             | 1.61E-24739          | 6.24E+237            | 16.44811             | Significant |
| gene_76 | 1.54E-24             | 3.10E-128            | 4.96E+103            | 16.44811             | Significant |
Table 2: Complete list of differentially expressed genes by GP model

| gene_id | \(L(\hat{\theta}|X,Y)\) | \(L(\hat{\theta}_1, \hat{\theta}_2|X,Y)\) | \(T = \frac{L(\hat{\theta}|X,Y)}{L(\hat{\theta}_1, \hat{\theta}_2|X,Y)}\) | \(\chi^2\) | Decision |
|---------|-----------------|-----------------|-----------------|-------------|----------|
| gene_77 | 6.71E-07        | 3.47E-45        | 1.93E+38        | 16.44811    | Significant |
| gene_78 | 1.44E-21        | 8.40E-75        | 1.71E+53        | 16.44811    | Significant |
| gene_79 | 1.44E-21        | 8.40E-75        | 1.71E+53        | 16.44811    | Significant |
| gene_80 | 2.59E-19        | 4.30E-186       | 6.03E+166       | 16.44811    | Significant |
| gene_82 | 7.30E-13        | 8.14E-32        | 8.97E+18        | 16.44811    | Significant |
| gene_83 | 3.35E+36        | 8.61E-103       | 4.92E+138       | 16.44811    | Significant |
| gene_85 | 2.07E+01        | 1.16E-62        | 1.79E+63        | 16.44811    | Significant |
| gene_88 | 1.45E-04        | 4.44E-253       | 3.28E+248       | 16.44811    | Significant |
| gene_89 | 2.07E+01        | 1.16E-62        | 1.79E+63        | 16.44811    | Significant |
| gene_91 | 1.44E-21        | 8.40E-75        | 1.71E+53        | 16.44811    | Significant |
| gene_99 | 2.50E-24        | 1.10E-123       | 2.28E+99        | 16.44811    | Significant |
| gene_101| 2.17E+21        | 4.00E-90        | 5.42E+110       | 16.44811    | Significant |
| gene_105| 6.79E-11        | 1.21E-61        | 5.62E+50        | 16.44811    | Significant |
| gene_106| 6.79E-11        | 1.21E-61        | 5.62E+50        | 16.44811    | Significant |
| gene_107| 6.53E-16        | 3.74E-46        | 1.74E+30        | 16.44811    | Significant |
| gene_113| 3.56E-23        | 2.75E-88        | 1.30E+65        | 16.44811    | Significant |
| gene_114| 7.66E-20        | 1.84E-176       | 4.17E+156       | 16.44811    | Significant |
| gene_115| 1.40E-19        | 1.87E-195       | 7.51E+175       | 16.44811    | Significant |
| gene_117| 1.88E+85        | 1.80E-188       | 1.04E+273       | 16.44811    | Significant |
| gene_119| 6.98E-24        | 1.06E-97        | 6.61E+73        | 16.44811    | Significant |
| gene_122| 4.93E-21        | 4.11E-185       | 1.20E+164       | 16.44811    | Significant |
| gene_123| 5.66E-20        | 1.17E-179       | 4.83E+159       | 16.44811    | Significant |
| gene_124| 1.32E-15        | 2.40E-201       | 5.49E+185       | 16.44811    | Significant |
| gene_125| 9.34E-10        | 2.13E-189       | 4.38E+179       | 16.44811    | Significant |
| gene_127| 3.56E-23        | 2.75E-88        | 1.30E+65        | 16.44811    | Significant |
| gene_131| 8.82E+45        | 2.86E-33        | 3.08E+78        | 16.44811    | Significant |
| gene_133| 1.85E-24        | 1.88E-118       | 9.85E+93        | 16.44811    | Significant |
| gene_136| 1.54E+05        | 8.64E-298       | 1.78E+302       | 16.44811    | Significant |
| gene_138| 5.01E-12        | 2.22E-29        | 2.25E+17        | 16.44811    | Significant |
| gene_139| 8.93E+43        | 2.57E-259       | 3.48E+302       | 16.44811    | Significant |
| gene_140| 0.002706508     | 6.26E-259       | 4.32E+255       | 16.44811    | Significant |
| gene_142| 4.86E-09        | 6.27E-135       | 7.75E+125       | 16.44811    | Significant |
| gene_143| 1.19E+02        | 1.33E-276       | 8.96E+277       | 16.44811    | Significant |
| gene_144| 2.17E-19        | 5.90E-61        | 3.68E+41        | 16.44811    | Significant |
| gene_149| 2.71E-03        | 6.26E-259       | 4.32E+255       | 16.44811    | Significant |
| gene_id  | \(L(\theta|X,Y)\) | \(L(\hat{\theta}_1, \hat{\theta}_2|X,Y)\) | \(T = \frac{L(\theta|X,Y)}{L(\hat{\theta}_1, \hat{\theta}_2|X,Y)}\) | \(\chi_1^2\) | Decision |
|----------|-----------------|-----------------|-----------------|---------|-----------|
gene_152  | 2.44E-24        | 2.42E-141       | 1.00E+117       | 16.44811 | Significant |
gene_154  | 2.86E-07        | 2.03E-257       | 1.41E+250       | 16.44811 | Significant |
gene_156  | 5.68E-21        | 1.30E-69        | 4.38E+48        | 16.44811 | Significant |
gene_157  | 2.17E-19        | 5.90E-61        | 3.68E+41        | 16.44811 | Significant |
gene_161  | 4.14E-09        | 5.69E-27        | 7.27E+17        | 16.44811 | Significant |
gene_162  | 1.67E-21        | 1.02E-178       | 1.64E+157       | 16.44811 | Significant |
gene_164  | 2.58E-17        | 1.98E-49        | 1.30E+32        | 16.44811 | Significant |
gene_168  | 8.82E+45        | 2.86E-33        | 3.08E+78        | 16.44811 | Significant |
gene_171  | 1.32E-15        | 2.40E-201       | 5.49E+185       | 16.44811 | Significant |
gene_173  | 1.67E-21        | 1.02E-178       | 1.64E+157       | 16.44811 | Significant |
gene_175  | 4.43E-15        | 7.26E-40        | 6.10E+24        | 16.44811 | Significant |
gene_176  | 1.40E-09        | 2.18E-196       | 6.41E+186       | 16.44811 | Significant |
gene_177  | 1.40E-09        | 2.18E-196       | 6.41E+186       | 16.44811 | Significant |
gene_178  | 1.30E+34        | 1.96E-126       | 6.62E+159       | 16.44811 | Significant |
gene_181  | 4.13E-05        | 7.83E-265       | 5.28E+259       | 16.44811 | Significant |
gene_182  | 2.10E+54        | 1.14E-242       | 1.85E+296       | 16.44811 | Significant |
gene_184  | 1.32E-15        | 2.40E-201       | 5.49E+185       | 16.44811 | Significant |
gene_187  | 1.22E+52        | 2.47E-174       | 4.94E+225       | 16.44811 | Significant |
gene_189  | 154211.1399     | 8.64E-298       | 1.78E+302       | 16.44811 | Significant |
gene_190  | 1.67E-21        | 1.02E-178       | 1.64E+157       | 16.44811 | Significant |
gene_191  | 2.58E-17        | 1.98E-49        | 1.30E+32        | 16.44811 | Significant |
gene_195  | 1.40E-19        | 1.87E-195       | 7.51E+175       | 16.44811 | Significant |
gene_196  | 1.88E+85        | 1.80E-188       | 1.04E+273       | 16.44811 | Significant |
gene_199  | 8.11E-10        | 1.62E-96        | 4.99E+86        | 16.44811 | Significant |
gene_204  | 2.73E-24        | 5.62E-134       | 4.86E+109       | 16.44811 | Significant |
gene_207  | 7.12E-06        | 4.98E-19        | 1.43E+13        | 16.44811 | Significant |
gene_272  | 0.70566675      | 3.49E-280       | 2.02E+279       | 16.44811 | Significant |
gene_273  | 7.31E-24        | 1.47E-97        | 4.96E+73        | 16.44811 | Significant |
gene_274  | 0.70566675      | 3.49E-280       | 2.02E+279       | 16.44811 | Significant |
gene_275  | 7.31E-24        | 1.47E-97        | 4.96E+73        | 16.44811 | Significant |
gene_277  | 0.852325596     | 3.84E-280       | 2.22E+279       | 16.44811 | Significant |
gene_279  | 53781.15757     | 3.57E-296       | 1.51E+300       | 16.44811 | Significant |
gene_281  | 2.91E-24        | 1.19E-105       | 2.45E+81        | 16.44811 | Significant |
gene_284  | 1.69E-09        | 1.34E-247       | 1.26E+238       | 16.44811 | Significant |
gene_290  | 1.49E-17        | 3.53E-63        | 4.24E+45        | 16.44811 | Significant |
gene_297  | 2.86E-07        | 2.03E-257       | 1.41E+250       | 16.44811 | Significant |
Table 4: Complete list of differentially expressed genes by GP model

| gene_id  | $L_1(\theta|X,Y)$ | $L_2(\theta_1,\theta_2|X,Y)$ | $T = \frac{L(\theta|X,Y)}{L(\theta_1,\theta_2|X,Y)}$ | $\chi^2$ | Decision |
|----------|------------------|-----------------|-----------------|--------|----------|
| gene_313 | 2.44E-24         | 2.42E-141       | 1.00E+117       | 16.44811 | Significant |
| gene_297 | 2.86E-07         | 2.03E-257       | 1.41E+250       | 16.44811 | Significant |
| gene_301 | 1.89E-12         | 1.35E-235       | 1.40E+223       | 16.44811 | Significant |
| gene_302 | 5.68E-21         | 1.30E-69        | 4.38E+48        | 16.44811 | Significant |
| gene_306 | 2.17E-19         | 5.90E-61        | 3.68E+41        | 16.44811 | Significant |
| gene_306 | 2.17E-19         | 5.90E-61        | 3.68E+41        | 16.44811 | Significant |
| gene_317 | 1.74E-24         | 2.37E-131       | 7.33E+106       | 16.44811 | Significant |
| gene_320 | 6.51E-18         | 3.08E-207       | 2.11E+189       | 16.44811 | Significant |
| gene_393 | 6.11E-16         | 7.03E-218       | 8.69E+201       | 16.44811 | Significant |
| gene_414 | 4.08E-17         | 1.22E-73        | 3.34E+56        | 16.44811 | Significant |
| gene_599 | 2.75E-17         | 2.32E-49        | 1.19E+32        | 16.44811 | Significant |
| gene_646 | 2.69E-09         | 3.79E-116       | 7.10E+106       | 16.44811 | Significant |
| gene_661 | 8.37E-17         | 5.53E-47        | 1.51E+30        | 16.44811 | Significant |
| gene_706 | 1.21E-24         | 2.85E-139       | 4.25E+114       | 16.44811 | Significant |
| gene_720 | 3.79E-23         | 4.42E-89        | 8.57E+65        | 16.44811 | Significant |
| gene_760 | 5.04E-24         | 1.60E-115       | 3.14E+91        | 16.44811 | Significant |
| gene_790 | 1.11E-16         | 9.45E-47        | 1.17E+30        | 16.44811 | Significant |