Implementation, Adaptation and Evaluation of Statistical Analysis Techniques for Next Generation Sequencing Data

Rachael Fulton

A thesis submitted for the degree of

Master of Science

Department of Statistics

October 2009

Abstract

Deep sequencing is a new high-throughput sequencing technology intended to lower the cost of DNA sequencing further than what was previously thought possible using standard methods. Analysis of sequencing data such as SAGE (serial analysis of gene expression) and microarray data has been a popular area of research in recent years. The increasing development of these different technologies and the variety of the data produced has stressed the need for efficient analysis techniques.

Various methods for the analysis of sequencing data have been developed in recent years: both SAGE data, which is discrete; and microarray data, which is continuous. These include simple analysis techniques, hierarchical clustering techniques (both Bayesian and Frequentist) and various methods for finding differential expression between groups of samples. These methods range from simple comparison techniques to more complicated computational methods, which attempt to isolate the more subtle dissimilarities in the data.

Various analysis techniques are used in this thesis for the analysis of unpublished deep sequencing data. This analysis was approached in three sections. The first was looking at clustering techniques previously developed for SAGE data, Poisson C / Poisson L algorithm and a Bayesian hierarchical clustering algorithm and evaluating and adapting these techniques for use on the deep sequencing data. The second was looking at methods to find differentially expressed tags in the dataset. These differentially expressed tags are of interest, as it is believed that finding tags which are significantly up or down regulated

across groups of samples could potentially be useful in the treatment of certain diseases. Finally due to the lack of published data, a simulation study was constructed using various models to simulate the data and assess the techniques mentioned above on data with pre-defined sample groupings and differentially expressed tags. The main goals of the simulation study were the validation of the analysis techniques previously discussed and estimation of false positive rates for this type of large, sparse dataset.

The Bayesian algorithm yielded surprising results, producing no hierarchy, suggesting no evidence of clustering. However, promising results were obtained for the adapted Poisson C / Poisson L algorithm applied using various models to fit the data and measures of similarity. Further investigation is needed to confirm whether it is suitable for the clustering of deep sequencing data in general, especially where the situation of three or more groups of interest occurs.

From the results of the differential expression analysis it can be deduced that the over-dispersed log linear method for the analysis of differential expression, particularly when compared to simple test such as the 2-sample t-tests and the Wilcoxon signed rank test is the most reliable. This deduction is made based upon the results of the overlapping with other methods and the more reasonable number of differentially expressed tags detected, in contrast to those detected using the adapted log ratio method. However none of this can be confirmed, as no information was known about the tags in either dataset.

The success of the Poisson C / Poisson L algorithm on both the Poisson and Truncated Poisson simulated datasets suggests that the method of simulation is acceptable for the assessment of clustering algorithms developed for use on sequencing data. However,

evaluation of the differential expression analysis performed on the simulated data indicates that further work is needed on the method of simulation to increase its reliability.

The algorithms presented can be adapted for use on any form of discrete data. From the work done here, there is there is evidence that the adapted Poisson C / Poisson L algorithm is a promising technique for the analysis of deep sequencing data.

Acknowledgements

I would like to show my gratitude to my supervisor Dr Raya Khanin whose guidance and support enabled me to develop an understanding of the subject. The thesis would not have been possible had MSKCC New York not supplied the unpublished datasets for analysis.

I also owe my deepest gratitude to John McClure as he has made his support available in many ways. I am indebted to him for the help and guidance he provided during the absence of my primary supervisor. Finally, I offer my regards to my second supervisor Dr Harper Gilmour and the rest of the University of Glasgow Department of Statistics who have supported me in many ways during the completion of my project.

Contents

Table of Figures

Chapter 1

Introduction

1.1 Introduction

Several methods of analysis for data produced by deep sequencing are presented, evaluated and discussed in this thesis. Deep sequencing is a novel, high-throughput sequencing technology intended to lower the cost of DNA sequencing further than what was previously thought probable using standard methods. Analysis of sequencing data such as SAGE (Serial Analysis of Gene Expression) and microarray data has been a popular area of research in recent years. The increasing development of these different technologies and the variety of the data produced has stressed the need for efficient analysis techniques.

Various methods for the analysis of sequencing data have been developed in recent years: many have been developed for both SAGE data, which is discrete; and microarray data, which is continuous. These include simple analysis techniques, clustering techniques (both Bayesian and Frequentist) and various methods for finding differential expression between groups of samples. These methods range from simple comparison techniques to more complicated computational methods, which attempt to isolate the more subtle dissimilarities in the data.

In this thesis various analysis techniques for clustering and differential expression, previously developed for the analysis of sequencing data will be evaluated and in some

cases adapted for the use on the data provided; next-generation sequencing data produced by deep sequencing. In an attempt to predict false positives that may occur in the data a simulation study was constructed and each of the analysis techniques tested on the simulated dataset.

1.2 Background

1.2.1 What is DNA sequencing?

The basic structure of DNA is built up of a large collection of nucleotide bases A (adenine), C (cytosine), G (guanine) and T (thymine) joined together (shown in Figure 1a). A fifth base, called uracil (U), usually takes the place of thymine in RNA molecules. However uracil is not usually found in DNA, occurring only as a breakdown product of cytosine. DNA sequencing is a collective expression for the methods used to isolate the order of these bases. This is important as it determines the genetic information that is contained on a single strand of DNA i.e. the order of the nucleotide bases present in the DNA strand. From this scientists can then determine which individual genes appear in this specific DNA strand as each gene has a unique order of nucleotide bases. Molecules such as microRNAs amd coding segments of DNA called exons also have a unique sequence of these nucleotide bases which can be identified using DNA sequencing methods. The data investigated in this thesis is microRNA sequencing data, below in Figure 1b, is a diagram of an individual microRNA molecule, illustrating the individual sequence of nucleotide bases. Mutations in these sequences can also be identified which may cause disease or genetic disorders. [1]

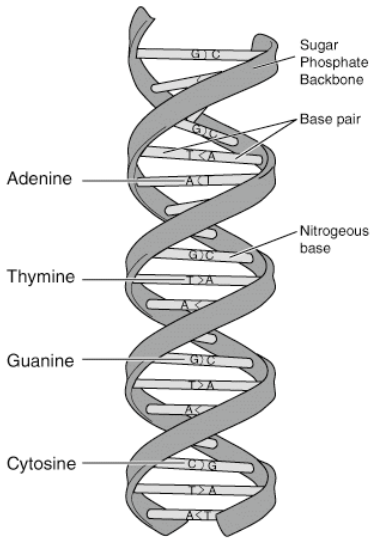


Figure 1a: An example of an double strand of DNA. Each colour red, green yellow and blue represent a specific nucleotide base.

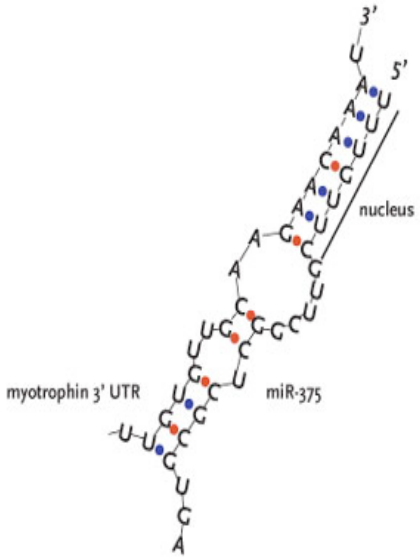


Figure 1b: An example of a miRNA molecule, which consists of a specific sequence of nucleotide bases. Hundreds, sometimes thousands of these miRNAs can be obtained when DNA is sequenced.

Hutchison [2] discusses the need for sequencing by highlighting the genetic nature of all disease. 'All disease has a genetic basis, whether in genes inherited by the affected individual, environmentally induced genetic changes that produce a cancer, or the genes of a pathogen and their interaction with those of the infected individual.'[2]

Gene sequencing is beginning to have a significant influence in medicine on the diagnosis and treatment of diseases. 'Genome sequences have provided potential targets for drug therapy' [3] as well as candidates for certain vaccines. The aim is to eventually provide genotype based treatments which, potentially will be more effective than current treatments. Metzker [4] discusses the various uses of gene sequencing in relation to health and disease, with applications ranging from comparative genomics and evolution to epidemiology and applied medicine.

Although scientists had developed methods for protein sequencing, in order to sequence DNA effectively many obstacles needed to be overcome [2]. These included:

- Chemical properties of two or more individual DNA molecules being similar between two or more different molecules.
- Compared to previously examined protein sequences, DNA sequences have a much larger chain length.
- Due to the low number of nucleotide bases in DNA (four) this made sequencing more difficult for DNA than for protein.
- No base specific DNA assays were known

Methods of DNA sequencing, such as the Sanger sequencing method[5], developed in the late 1970's, tried to overcome these problems. Initially they were not powerful enough to isolate compete gene sequences. However the Sanger sequencing method (sequencing by synthesis) has provided a basis for all DNA sequencing technology since its development. This method, conducted *in vivo (*i.e conducted within a living organism), employs DNA synthesis on a single stranded template while integrating chain terminators at random ('Chain termination is the process whereby the last amino acid is added to a polypeptide, also known as stop codons. [6].'). This generates a variety of fragment sizes corresponding to the locations of the terminators [7]. This method however is not ideal as certain properties of DNA do not replicate well.

Using older methods, sequencing of an individual gene could take months and could prove very costly. In the last decade many new methods have surfaced which have revolutionised the way sequencing is carried out. These methods are high-throughput and enable sequencing to be conducted in parallel making the sequencing process

significantly faster and much less costly. These methods are also performed *in vitro* (in an artificial environment) which bypass the replication issues encountered when using the *In vivo* Sanger method.

The most recently developed methods are known as deep sequencing which is achieved using methods such as 454 sequencing and Solexa. Both of these methods adopt a sequencing by synthesis approach. 'Sequencing by synthesis involves extracting an individual strand of the DNA to be sequenced and synthesising its complimentary strand enzymatically'[8] The main advantage of deep sequencing other than the speed and cost is that it allows small regions of DNA to be amplified vastly and mutations can then be detected at much higher sensitivity levels than previous methods such as Sanger which has massive implications in medical research[9]. Other methods such as single molecule sequencing and sequencing by hybridisation and ligation exist but will not be discussed in this thesis. Shown in Figure 1c is an outline of past and present sequencing technologies, taken from a review by Hall [6].

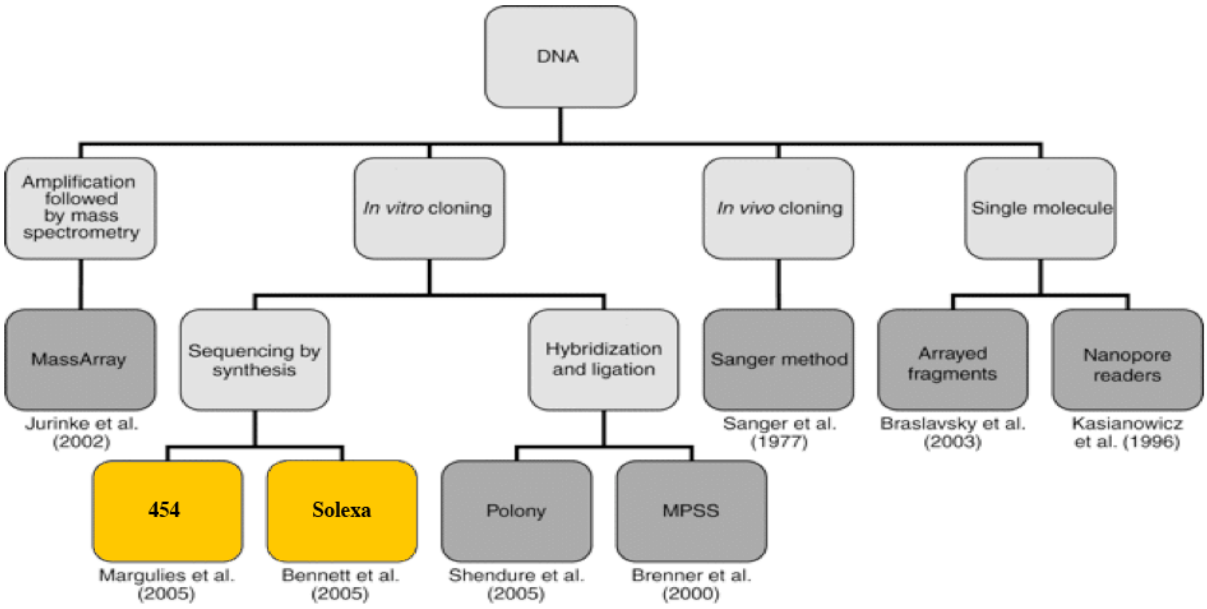


Figure 1c: A brief look at past and present sequencing technologies. The two of these techniques that are methods of deep sequencing are highlighted in yellow.

1.2.2 Why do we need DNA sequencing?

Next generation sequencing can be used in many applications to reduce the cost of sequencing and providing quicker means to approach vital biological discoveries. These discoveries are leading to advancements in cancer, AIDS and many other areas of medical research. One of the most publicised fields these technologies are currently used in is personalised medicine. Companies such as Roche applied science use 454 sequencing for human exome sequencing. [10]

'Since exons are the most functionally relevant part of the gene, sequencing of these can lead to the discovery of much of the functional variation responsible for major diseases such as cancer and Alzheimer's. This technology can also begin to shed light on why certain diseases occur more frequently in specific populations or subset of individuals.'[10]

In this thesis the sequencing identification of microRNAs from DNA sequencing data will be the topic of interest, as the data provided was microRNA-sequencing data from various cancerous tissue samples. MicroRNAs (tags) are short RNA molecules 19-25 nucleotides in length [11] so a given tag can be represented by a sequence of nucleotide bases. These tags play an important role in gene regulation. They act as a regulator of gene expression by pairing to a section of one or more messenger RNAs (mRNAs). Many studies have been carried out that suggest the importance of tags as analytical tools in the study of conditions such as cancer and heart disease. [12][13][14]

Sequencing can be used to identify and classify microRNAs, which is a growing area of research. This identification and classification is vital in the research of many viruses and diseases as these microRNAs regulate numerous processes such as cell replication and cell death. In various diseases these microRNAs can play a vital role in treatment development, as specific microRNAs that are differentially expressed or have a high level of expression in certain samples may regulate processes specific to a particular disease.[15]

In order to study the effect that microRNAs have on gene regulation the expression level of each tag in a sample needs to be found. Older sequencing methods, although useful for detecting novel tags, were very slow and costly. Using sequencing methods such as 454 and Solexa, the speed is increased and the cost lowered. This then provides a clearer perception of the tag itself. Using these methods of deep sequencing tags that express low differences between samples can be detected and tag expression can then be extensively profiled and any changes in expression can be clearly identified.[11]

1.2.3 Other methods available

Other methods to study tag expression level previously used include microarray and SAGE (Serial Analysis of Gene Expression). Microarray studies are conducted using competitive hybridisation and are primarily used to identify differentially expressed tags between two different groups or samples. This technology can prove to be very expensive, so the experiments are performed with very few replicates. This can lead to false positives and false negatives. However, using a larger sample size can increase the detection level of expression in the analysis and can decrease the error. However this can waste resources and time[16]. The data collected from a microarray experiment is continuous as it is a measurement of florescence[17]. Microarray experiments are restricted to detect known tags and only those that are printed on the array.

SAGE also known as serial analysis of gene expression can also be used to assess expression levels of tags. SAGE is a sampling by sequencing method, which is single clone sequencing using multiple transcripts and multiple tags. Each SAGE experiment represents multiple transcripts.[18] Due to the sequencing nature of the experiment SAGE can potentially detect lower levels of expression and can also detect novel tags. The data produced by SAGE experiments is presented in the form of counts, this data can provide information on all the tags in the given sample.[17]

The data provided by deep sequencing experiments is similar to SAGE in that it is in the form of counts and can provide information on all the tags in the given sample. However the data is considerably sparser than published SAGE datasets (i.e. a large number of zero counts) and also the range of the data is much higher. Deep sequencing can also look into much lower levels of expression than SAGE, providing deeper insight into the sample.

1.3 Aims and Objectives

The aims of this thesis are to explore various techniques previously adopted for the analysis of sequencing data and in some cases adapt these techniques, in others evaluate their performance for use in the analysis of next generation sequencing data.

This will be approached in three separate sections:

- Two clustering methods developed for the analysis of SAGE data have been implemented on the data provided and assessed for use on this new type of data. Adaptations to one of these algorithms will also be implemented and discussed.
- Using the results from the clustering algorithms mentioned above, various methods of differential expression analysis were discussed and implemented on the data provided.
- A simulation study is proposed and presented to evaluate each of the techniques explored in the sections above.

1.4 Data

Several techniques for the analysis of next generation sequencing data are discussed in this thesis. These were tested on deep sequencing data, and many of these methods can potentially be adapted to any type of discrete data. The work presented here was carried out by attempting to model the data using various probability distributions. These include: the standard Poisson distribution traditionally used for count data, the Truncated Poisson distribution used in an attempt to account for the high zero count nature of the data and finally the negative binomial distribution, again used in an attempt to take into account the high nature of the zero counts in the data.

The computational biology research centre at the Memorial Sloan-Kettering Cancer Centre (MSKCC) in New York provided two datasets. Each dataset consists of a number of libraries. These libraries are tissue samples from cancerous and non-cancerous subjects. Each library (also known as sample) contains an individual count for a given number of microRNAs (also referred to as tags), which have appeared during sequencing.

The first is a large dataset consisting of 55 samples each sample containing a count of over 500 tags. This count represents the number of times the sequence related to this specific tag appears in the given sample. The information on which group (i.e. cancerous and non cancerous) each sample belongs to was given a-priori in this dataset. The information given stated that this dataset consisted of 3 groups (or clusters) of samples, this information is used in Chapter 2 and Chapter 4 to assess the reliability of the clustering algorithm. No information was given about the clustering of the tags in the dataset for example it would be useful to know what tags are expected to appear together in a sample when DNA is sequenced.

The second dataset is of similar format to the first but has over twice as many individual tags (1186) and less than half the number of samples (26). The main difference of this dataset is that no information about groupings was given a-priori so any analysis performed on this dataset is speculative. It is not known whether these datasets were produced using Solexa or 454 sequencing as no other information was given other than to say it was produced using methods of deep sequencing. It is important to note that due to the dataset being unpublished no information was given about tag grouping in either of the two datasets so the results of the analysis presented in Chapter 5 cannot be confirmed.

Although in this thesis tags are referred to as different microRNAs, these tags could also be genes, exons or pieces of DNA whose functional role has not been discovered yet. An example of the data structure is given below. Each cell of the table represents the count of the given tag in the given sample.

Table 1: An example of the outline of sequencing data.

	Sample1	Sample2	Sample3	Sample4
Tag1	0	3456	65	9
Tag2	765	43	1002	8
Tag3	0	1	2	0

Chapter 2

Literature and methods

In this chapter a brief explanation of various methods and models previously developed for analysis of similar data types and found in literature will be given. Further investigation and adaptation of these methods will be investigated later in this thesis.

Notation is consistent throughout-miRNAs are referred to as tags and libraries of miRNAs are referred to as samples. The technical notation is denoted as follows:

- Samples range from $t = 1,...,T$ where T is the total number of samples in the given dataset.
- Tags range from $i = 1,...,N$, where N is the total number of individual tags in the dataset.
- $y_i(t)$ denoted the observed count of $tag\ i$ in sample t.
- θ_t denotes the total count of all tags in sample t, and $\theta(i)$ denotes the total count of $tag\ i$ over all samples.
- $\lambda_i(t)$ denotes the proportion of $tag\ i$ in sample t.
- $\mu_i(t) = \lambda_i(t)\theta_t$.

2.1 Different Models

Due to the count nature of this data many different models can be applied to it for analysis. The three that were used in this thesis were Poisson, Truncated Poisson and the Negative binomial.

2.1.1 Poisson

When looking at count data the Poisson distribution is a logical choice. If the data is Poisson distributed it is assumed that the count of each tag i in each sample t, $y_i(t)$ follows $Po(\theta_t \lambda_i(t))$. Where θ_t denotes the total count of all tags in sample t, and $\lambda_i(t)$ is the proportion of tag i in sample t. The probability mass function for this distribution is given by *(1)*.

$$p(y_i(t) \mid \theta_t \lambda_i(t)) = \frac{\exp\left(-\theta_t \lambda_i(t)\right)\left(\theta_t \lambda_i(t)\right)^{y_i(t)}}{y_i(t)!} \qquad (1)$$

However, this distribution does not always take into account the nature of the large zero counts in the data and other distributions need to be investigated.

2.1.2 Truncated Poisson

The truncated Poisson distribution is inherently Poisson in nature but with the desired limit removed. In this case the zero-truncated Poisson will be the only truncation of interest. This distribution is useful because by removing the zero counts the data can be analysed differently.

As for the Poisson, it is assumed each count $y_i(t)$ follows $TrPo(\theta_t\lambda_i(t))$, where θ_t and $\lambda_i(t)$ are the same as mentioned above. The probability mass function for this distribution is given by *(2)*.

$$p(y_i(t) \mid \theta_t\lambda_i(t)) = \frac{\exp\left(-\theta_t\lambda_i(t)\right)\left(\theta_t\lambda_i(t)\right)^{y_i(t)}}{(1-e^{-\theta_t\lambda_i})y_i(t)!} \qquad (2)$$

However, due to the zero counts being removed, a way to estimate $\lambda_i(t)$ needs to be found. David et al [19] suggest using the truncated sample mean $\bar{y}(t)$ for each tag and calculating $\hat{\lambda}(t)$ using *(3)*.

$$\bar{y}(t) = \hat{\lambda}(t)\theta_t(1-e^{\hat{\lambda}(t)}\theta_t) \qquad (3)$$

Although it seems non-trivial to get an estimate for $\hat{\lambda}_t$ from this equation, methods such as the Newton's method of root finding can be employed here.

This method does not effectively take into account the nature of the data as it removes all of the zero counts. Due to the large number of zero counts present in the data, it is a distinct possibility that by removing these counts the analysis could be incorrect.

2.1.3 The Negative Binomial Distribution

The negative binomial distribution is often used to model biological count data as, although it is an extension of the Poisson distribution, it takes into account that often the observed variance can be much greater than the mean. Robinson et al [20] explore using the Negative Binomial distribution to model SAGE data. Various methods for estimating the dispersion parameter are also suggested.

If it is assumed that a given tag i over all samples $t = 1,..,T$, y_i is Negative Binomially distributed. So $y_i(t) \sim NegBin(\theta(t)\lambda_i(t),\phi)$ where ϕ is the estimated dispersion. The probability mass function of the negative binomial is given by *(4)*.

$$p(y_i(t) \mid \theta_t\lambda_i(t)) = \frac{\Gamma(y_i(t) + \phi^{-1})}{\Gamma\left(\phi^{-1}\right)\Gamma\left(y_i(t)+1\right)}\left(\frac{1}{1+\theta_t\lambda_i(t)\phi}\right)^{\phi^{-1}}\left(\frac{\left(\theta_t\lambda_i(t)\right)^{y_i(t)}}{\phi^{-1}+\theta_t\lambda_i(t)}\right) \qquad (4)$$

In most cases all tags would have a common dispersion, Robinson et al [20] suggest a Pseudo-Likelihood *(5)* and Quasi-Likelihood *(6)* approach for dispersion estimation, which can both be used to calculate both common and tag-specific dispersion estimates.

The pseudo likelihood (PL) method *(5)* estimates variance function parameters of the GLM using a distribution free goodness of fit statistic.

$$\sum_{t=1}^{T}\frac{(y_i(t)-\hat{\theta}_t\hat{\lambda}_i(t))^2}{\hat{\theta}_t\hat{\lambda}_i(t)\left(1+\phi_{PseudoLik}\hat{\theta}_t\hat{\lambda}_i(t)\right)} = n-1 \qquad (5)$$

The quasi-likelihood (QL) method *(6)* estimates dispersion in a similar way to *(5)*. This method replaces the Pearson statistic with a deviance statistic.

$$2\sum_{t=1}^{T} y_i(t)\log\left[\frac{y_i(t)}{\hat{\theta}_t\hat{\lambda}_i(t)}\right] - \left(y_i(t)+\phi^{-1}{}_{QuasiLik}\right)\log\left[\frac{y_i(t)+\phi^{-1}{}_{QuasiLik}}{\hat{\theta}_t\hat{\lambda}_i(t)+\phi^{-1}{}_{QuasiLik}}\right] = n-1 \qquad (6)$$

Both the pseudo and quasi likelihood equations above can be used to estimate a tag-specific dispersion. Robinson et al [20] also introduce maximum likelihood and quantile adjustment methods for dispersion estimation but they will not be studied here.

2.2 Clustering

'Clustering is the grouping of similar objects' [21]. The aim of clustering analysis is to allocate the objects of interest into mutually exclusive clusters. Cluster analysis can provide valuable insight into patterns and important groupings in the data. Two methods of clustering are explored and evaluated in this thesis. In this section methods previously developed are explained and will be discussed further in Chapter 4.

2.2.1 Poisson C / Poisson L algorithm

2.2.1.1 Likelihood and Chi-Square distance measures

In order to effectively cluster any kind of data, an appropriate similarity measure has to be chosen that takes into account the nature of the specific data. The Poisson C/ Poisson L algorithm, proposed by Cai et al [22] is a K-means [21] based clustering algorithm. This method was developed for SAGE data and introduces two new similarity measures - likelihood and chi-square.

The assumption is made that the distribution of each individual tag in an individual sample is Poisson. Let $y_i(t)$ be the count of tag i in sample t, then $y_i(t) \sim Po(\theta(i)\lambda_i(t))$. where $\theta(i)$ is the expected sum of counts of tag i over all samples; and $\lambda_i(t)$ is the proportion of tag i in sample t. Number of samples considered is $t = 1,..,T$. Using $\theta(i)\lambda_i(t)$, the count of each tag is redistributed according to the cluster profile determined beforehand (λ) but keeps the sum of counts across all samples constant. [22]

The joint Likelihood function for a cluster consisting of m tags is given by *(7)* where Y_i denotes the vector of the counts of tag i across all samples $t = 1,..T$:

$$L(\lambda\theta \mid y) \propto f(Y_1,..,Y_m \mid \lambda,\theta(1),..,\theta(m)) = \prod_{i=1}^{m}\prod_{t=1}^{T} \frac{\exp\left(-\theta(i)\lambda_i(t)\right)\left(\theta(i)\lambda_i(t)\right)^{y_i(t)}}{y_i(t)!} \qquad (7)$$

The maximum likelihood estimates of each θ and λ can then be calculated using *(8)*.

$$\hat{\theta}(i) = \sum_{t=1}^{T} y_i(t), \quad \hat{\lambda}_i(t) = \sum_{i=1}^{m} \frac{y_i(t)}{\hat{\theta}(i)} \qquad (8)$$

Using this, a cluster centre $\lambda = \left(\lambda(1),\lambda(2),..,\lambda(m)\right)$ can be calculated for all tags in the cluster. The expected total count for tag i, $\theta(i)$, and the proportion of each tag i in sample t, $\lambda_i(t)$ can be estimated using *(8)*.

Both the likelihood function *(7)* and the chi square statistic *(9)* are used to calculate the similarity of an individual tag to a cluster centre.

$$S = \sum_{i}\sum_{t=1}^{T}\left(y_i(t) - \hat{\lambda}(t)\hat{\theta}(i)\right)^2 \Big/ \hat{\lambda}(t)\hat{\theta}(i) \qquad (9)$$

The algorithm works based on a k-means principle, as follows:

1. The number of clusters K is selected a-priori.
2. $\hat{\theta}(i)$ is calculated *(8)* for each individual tag and each tag is randomly assigned to a cluster.
3. Cluster centres λ_r^k are calculated from *(8)*. Initialisation $r = 0$.
4. Now each tag is individually assigned to the cluster, which minimises the chi-square statistic *(10)* or to the cluster in which the individual likelihood of the tag *(11)* is minimised depending on whether the method chosen is the chi square statistic or the likelihood of the individual tag.

$$S_{i,k} = \sum_{t=1}^{T} \left(y_i(t) - \hat{\lambda}_r^k \hat{\theta}(i) \right)^2 \Big/ \hat{\lambda}_r^k \hat{\theta}(i) \qquad (10)$$

$$L_{i,k} = -\log f(Y_i(t) \mid \hat{\lambda}_r^k \hat{\theta}(i)) \qquad (11)$$

5. New cluster centres λ_r^k can then be calculated from *(8)* using the reassignment of the tags.
6. This is repeated until the algorithm converges

2.2.1.2 New data transformations

Kim et al [23] propose an adaptation to the Poisson C/Poisson L algorithm by replacing the likelihood and Chi-square as similarity measures with a new similarity measure denoted 'TransChisq.'

This data transformation [23] is a more robust alternative to the likelihood function and chi square statistic, it is proposed. It is said to highlight the expression shape, and consider the common differences of the original vectors of tag counts. Given the expression profile of an individual tag, $Y_i = (y_i(1),...,y_i(T))$ the transformed vector Z_i is of dimension $T(T-1)/2$, where the number of samples is $t = 1,...,T$, the components are in the form of $y_i(t_1) - y_i(t_2)$, where $(t_1,t_2) = (1,2),(2,3),....,(T-1,T)$.

If the Poisson model is used, as in the Poisson C / Poisson L algorithm the expected value of the transformed data becomes *(12)* and variance of the data becomes *(13)*.

$$E(y_i(t_1) - y_i(t_2)) = \left(\lambda_i(t_1) - \lambda_i(t_2) \right) \theta(i) \qquad (12)$$

$$Var(y_i(t_1) - y_i(t_2)) = \left(\lambda_i(t_1) + \lambda_i(t_2) \right) \theta(i) \qquad (13)$$

So, using these, the following statistic can now be used as a measure of similarity for a cluster consisting of m tags:

$$S_{trans} = \sum_{i} \sum_{t_1 t_2} \left((y_i(t_1) - y_i(t_2)) - E(y_i(t_1) - y_i(t_2)) \right)^2 / Var(y_i(t_1) - y_i(t_2)) \qquad (14)$$

The maximum likelihood estimates $\hat{\lambda}_t$ and $\hat{\theta}(i)$ can be calculated using *(8)* as in 2.2.1.1. The algorithm is approached as is shown in 2.2.1.1, when step 5 is reached the distance measure is replaced with *(15)* and the algorithm continues on to step 6 as in 2.2.1.1.

$$S_{trans.j,k} = \sum_{t_1 t_2} \left((y_i(t_1) - y_i(t_2)) - E(y_i(t_1) - y_i(t_2)) \right)^2 / Var(y_i(t_1) - y_i(t_2)) \qquad (15)$$

Although the Poisson C/ Poisson L algorithm has been proven adequate for SAGE data using the likelihood, Chi Square and the TransChiSquare similarity measures, it does not appear to take into account the high dimensionality and the sparseness of the deep sequencing datasets, as it does not cluster the samples in dataset 1 correctly. New adaptations to this method need to be considered which will be discussed and implemented in Chapter 4.

2.2.2 Bayesian Method

Berninger et al [24] suggested a Bayesian method, which can then be used for hierarchical clustering. It was observed [24] that in frequency distributions of tag counts in two individual samples, few tags were highly expressed occurring in copies of greater than one hundred. The majority of tags occur in only a small number of copies with a high number of tags with a zero count in each sample. Due to this style of frequency distribution a great deal of sampling noise is observed. To account for this, Berninger et al

[24] suggest a Bayesian probability framework as a similarity measure to identify significant changes in tag expression between samples.

Denote the true, but unknown count of $tag\ i$ in the first and second samples as $p_i(1)$ and $p_i(2)$ respectively. The observed tag counts in each of the two samples are denoted as $y_i(1)$ and $y_i(2)$ respectively, these can be considered multinomial samples from the distributions $\{p_i(1)\}$ and $\{p_i(2)\}$. If the true frequencies are known the probability of the data is given by *(16)*.

$$p\left(\{y_i(1)\},\{y_i(2)\} \mid \{p_i(1)\},\{p_i(2)\}\right) \propto \prod_i \left[p_i(1)^{y_i(1)} p_i(2)^{y_i(2)} \right] \qquad (16)$$

Two models, model I and model S, are assumed for calculating the probability of the observed counts $\{y_i(1)\}$ and $\{y_i(2)\}$. Model I assumes that the true frequencies $\{p_i(1)\}$ and $\{p_i(2)\}$ are unknown and independent of one another. To calculate the marginal likelihood of this model L_1, a Dirichlet prior of the form *(17)* $(x = 1,2)$ is assigned to the unknown frequency distributions.

$$p\left(\{p_i(x)\}\right) = \Gamma(N\alpha) \prod_i \frac{p_i(1)^{\alpha-1}}{\Gamma(\alpha)} \qquad (17)$$

where N is the number of tags and α is the pseudo count of the Dirichlet prior which is not tag specific and is set to 0.05. This is then integrated over all distributions where $\sum_i p_i(1) = \sum_i p_i(2) = 1$. The integral can be performed analytically using *(18)*.

$$L_1 = \frac{\Gamma(N\alpha)^2}{\Gamma(\theta_1 + N\alpha)\Gamma(\theta_2 + N\alpha)} \prod_i \frac{\Gamma(y_i(1) + \alpha)\Gamma(y_i(2) + \alpha)}{\Gamma(\alpha)^2} \qquad (18)$$

Model S assumes that the true counts of tag i in the two samples are equal, i.e. $p_i(1) = p_i(2) \forall i$ again a prior of the form *(17)* is assigned to the frequency distributions. The likelihood of this model is calculated analytically using *(19)*

$$L_S = \frac{\Gamma(N\alpha)}{\Gamma(\theta_1 + \theta_2 + N\alpha)} \prod_i \frac{\Gamma(y_i(1) + y_i(2) + \alpha)}{\Gamma(\alpha)} \qquad (19)$$

'The posterior probability for model S is then given by $L_I / (L_I + L_S)$. From this probability a measure to define the similarity between the expression profiles of two samples can be defined below'. [24]

$$d = \log\left((L_I + L_S)/L_S\right) \qquad (20)$$

The given similarity measure *(20)* can then be used for hierarchical clustering, in a k-means method similar to the Poisson C / Poisson L algorithm in 2.2.1. This algorithm shows a complex approach to clustering, which is computationally expensive. Different priors can be used and clustering methods other than k- means can be adapted from this, this will be investigated further later in this thesis.

2.3 Differential expression

Differential expression refers to finding which tags are significantly differently expressed between two or more samples or groups of samples. A tag is flagged as differentially expressed between two individual samples or two groups of samples if the selected testing method gives a p-value of less than 0.05. In this section several existing methods for finding differentially expressed tags are outlined and are reviewed in Chapter 5.

2.3.1 Statistical analysis of transcript profiles

Audic and Claverie [25] suggest a probability distribution that governs the occurrence of the same tag appearing in two different samples. They state that 'differentially expressed genes can be detected from the variations in the counts of their cognate sequence tags.' [25] It is proposed that this is a general result applicable to a wide variety of experimental applications.

$p(x)$ denotes the probability of observing x occurrences of a given tag in a sequence sample where θ_1 denotes the sample size. For each tag representing a small percentage of the sample and the sample size $N \geq 1000$, $p(x)$ closely follows a $Po(\mu)$ distribution as in *(21)*.

$$p(x) = \frac{e^{-\mu}\mu^{x}}{x!} \qquad (21)$$

If y occurrences of the same tag are observed in another sample of size θ_2, what is the probability of these y values? It is proposed that a solution can be constructed using x as a maximum likelihood estimate of μ and computing the probability of y occurrences given a Poisson distribution of mean $\mu = x$ *(22)*.

$$p(x) = \frac{e^{-x}x^{y}}{y!} \qquad (22)$$

Equation *(22)* is not the correct formula is it does not take into account the fluctuations of x around the unknown mean μ. To do this, *(22)* needs to be integrated over all possible values of μ and becomes the integral *(23)*. Equation *(23)* gives the probability of

observing a count y of the tag of interest in the second sample, given that x tags were observed in the first sample and x followed a Poisson distribution with mean μ_1.[25]

$$p(y \mid x) = \int_0^\infty \int_0^\infty p(d_1 = \mu_1 \mid x) p(y \mid d_2 = \mu_2) \delta\left(\mu_2 - \frac{\theta_2}{\theta_1}\mu_1\right) d\mu_1 d\mu_2 \qquad (23)$$

where μ_1 and μ_2 are forced in the same ratio as the sample sizes, θ_1 and θ_2 , so $\mu_2 = \frac{\theta_2}{\theta_1}\mu_1$. The term $p(d_1 = \mu_1 \mid x)$ is the probability that the true count of a given tag is μ_1 given that x occurrences of the same tag have been observed in a different sample. The other term $p(y \mid d_2 = \mu_2)$ is the probability of y tags given a Poisson distribution of mean μ_2, so $p(y \mid d_2 = \mu_2) = \frac{e^{-\mu_2}\mu_2{}^y}{y!}$. The next step to simplifying *(23)* is completed by applying Bayes theorem to $p(d_1 = \mu_1 \mid x)$ and defining the prior distribution $p(d_1 = \mu_1)$ by attributing an equal a priori probability to all the μ_1 values in the $[0,\infty]$ range. This leads to $p(y \mid d_2 = \mu_2) = \frac{e^{-\mu_1}\mu_1{}^x}{x!}$, applying these to *(23)* gives *(24)*:

$$p(y \mid x) = \frac{1}{x!y!}\left(\frac{\theta_2}{\theta_1}\right)^y \int_0^\infty d\mu_1 e^{-\mu_1\left(1+\frac{\theta_2}{\theta_1}\right)} \mu_1^{x+y} \qquad (24)$$

This can then be evaluated to give *(25)*.

$$p(y \mid x) = \left(\frac{\theta_2}{\theta_1}\right)^y \frac{(x+y)!}{x!y!\left(1+\frac{\theta_2}{\theta_1}\right)^{(x+y+1)}} \qquad (25)$$

This, it is proposed, is a valid statistic for calculating the differential expression of a tag in two samples. The main drawback of this particular method is that *(25)* cannot be

generalised to find differentially expressed tags in two groups of samples rather than two individual samples. It is suggested that pooling the data could be an option however; by doing this much of the information on within sample variation is lost. The use of this statistic given in *(25)* is restricted to only two samples, so it is very limited, more complex methods will now be introduced.

2.3.2 Weighted t-statistic

Simple tests such as the two-sample t-test and chi-square statistic can be applied to the proportions of the tags in each sample. However, these statistics do not provide a valid solution (discussed further in Chapter 5). Baggerly et al [26] advocate the use of a weighted t-statistic that incorporates both between sample and within sample variation in the dataset.

They consider the case of modelling a specific tag across one cluster of samples where the clusters are known a-priori. Let θ_t denote the total tag counts of sample t, $\lambda_i(t)$ denote the proportion of this particular tag and $y_i(t)$ denote the count for this tag in sample t. For the first part of the model it is assumed that the proportions follow a Beta distribution, $\lambda_i(t) \sim Beta(\alpha,\beta)$. This is a standard distribution for proportions. This distribution is not degenerate: it can have a positive variance. Only the first two moments of the distributions are taken into account in these calculations in attempt to invoke the central limit theorem and to get an approximately normal test statistic, and also for computational simplicity. If the proportions follow a beta distribution the mean and the variance are given by *(26)*.

$$E(\lambda_i(t)) = \frac{\alpha}{\alpha+\beta} \text{ and } Var(\lambda_i(t)) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad (26)$$

The next part of the model states that; given the true proportion in a sample $\lambda_i(t)$, the corresponding count $y_i(t)$ will have a binomial distribution conditional on the proportion as in *(27)*.

$$y_i(t) \mid \lambda_i(t) = Bi(\theta_t, \lambda_i(t)) \quad (27)$$

The unconditional mean and variance of $\hat{\lambda}_i(t) = y_i(t)/\theta_t$ can be calculated using the tower property of conditional expectation, $E(y_i(t)) = E(E(y_i(t) \mid \lambda_i(t)))$. This leads to

$$E(y_i(t)) = \theta_t \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad E(y_i(t)^2) = \theta_t \frac{\alpha}{\alpha+\beta} + \theta_t(\theta_t - 1)\left[\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{\alpha^2}{(\alpha+\beta)^2}\right].$$

The unconditional mean and variance are then given by *(28)*:

$$E(\hat{\lambda}_i(t) = \frac{y_i(t)}{\theta_t}) = \frac{\alpha}{\alpha+\beta} \text{ and } Var(\hat{\lambda}_i(t) = \frac{y_i(t)}{\theta_t}) = \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}\left[\frac{1}{\alpha+\beta} + \frac{1}{\theta_t}\right] \quad (28)$$

Denoted in the square bracket there are two components of the $Var(\hat{\lambda}_i(t))$ in *(28)*, both the within sample variation $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\left[\frac{1}{\theta_t}\right]$ and between sample variation $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\left[\frac{1}{\alpha\beta}\right]$ are calculated. Given the equations in *(28)* weights (w_t) are now added to see how to combine the results from different samples which gives *(29)* and *(30)*.

$$E(\sum w_t \hat{\lambda}_i(t)) = \frac{\alpha}{\alpha+\beta}\sum w_t = \frac{\alpha}{\alpha+\beta} \quad (29)$$

$$Var\left(\sum w_t \hat{\lambda}_i(t)\right) = \sum_t \frac{w_t^2 \alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}\left[\frac{1}{\alpha+\beta}+\frac{1}{\theta_t}\right] \qquad (30)$$

Provided $\sum(w_t) = 1$ the combination has the correct mean. Weights need to be chosen so as to minimise the between and within sample variation. Using the method of Lagrange multipliers, the constraint on the sum of the weights is introduced, so $\frac{\partial}{\partial w_t}\left[Var\left(\sum w_t \hat{\lambda}_i(t)\right) + \mu\left(1 - \sum w_t\right)\right] = 2w_t \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}\left[\frac{1}{\alpha+\beta}+\frac{1}{\theta_t}\right] - \mu = 0.$ The weights are then estimated by *(31)*.

$$w_t \propto \left[\frac{1}{\alpha+\beta}+\frac{1}{\theta_t}\right]^{-1} \qquad (31)$$

Given *(31)* the estimated proportion for the group (or cluster), $\hat{\lambda}$ can be written as *(32)*:

$$\hat{\lambda} = \sum(w_t \hat{\lambda}(t)) \qquad (32)$$

The variance of this proportion can then be given by:

$$\hat{V}(\hat{\lambda}) = \frac{\sum({w_t}^2 \hat{\lambda}(t)^2) - \left(\sum({w_t}^2)\right)\hat{\lambda}(t)^2}{1 - \sum({w_t}^2)}. \qquad (33)$$

An algorithm can be written to estimate α and β by manipulating the equations in *(29)* which can then be used to approximate $\hat{\lambda}$ and $\hat{V}(\hat{\lambda})$. To calculate the proportion and variance for a given tag in one cluster the algorithm steps through as follows:

1. Calculate the initial weights for each tag using $\frac{\theta_t}{\sum\theta_t}$

2. Calculate the proportion of the tag in the cluster from *(32)*.
3. Calculate the variance of the tag in the cluster using *(33)*.
4. Now $\hat{\beta}$ can be calculated by manipulating *(29)* to get *(34)*.

$$\hat{\beta} = \frac{\hat{\lambda}\left(1-\hat{\lambda}\right) - \sum\left(w_t\right)^2 - \hat{V}}{\hat{V}\left(1-\hat{\lambda}\right)^{-1} - \hat{\lambda}\left(\sum\left(w_t\right)^2/\theta_t\right)} \qquad (34)$$

5. In the same way calculate $\hat{\alpha}$ *(35)*.

$$\hat{\alpha} = \frac{\hat{\lambda}}{\left(1-\hat{\lambda}\right)}\hat{\beta} \qquad (35)$$

6. New weights can be calculated from *(31)*.
7. The algorithm returns to step 2 and continues until convergence, i.e. until the weights calculated in the previous iteration are equal to those calculated in the current iteration to 3 decimal places.

From the algorithm above an approximate value for $\hat{\lambda}$ and $\hat{V}(\hat{\lambda})$ which can then be used to calculate the t-statistic *(36)* and degrees of freedom *(37)*, where $\hat{\lambda}_A$ and $\hat{\lambda}_B$ denote the proportion of the tag of interest in cluster 1 and cluster 2 respectively.

$$t_w = \frac{\hat{\lambda}_A - \hat{\lambda}_B}{\sqrt{\hat{V}_A + \hat{V}_B}} \qquad (36)$$

$$df = \frac{\left(\hat{V}_A + \hat{V}_B\right)^2}{\frac{\hat{V}_A{}^2}{\theta_A - 1} + \frac{\hat{V}_B{}^2}{\theta_B - 1}} \qquad (37)$$

The p-values can then be calculated using the two statistics above using the pt() command in R [37] the statistical computing language. A significance level of 0.05 or 0.01 can be chosen and if the p-value is below this then the tag can be defined as differentially expressed.

Although an improvement upon the traditional t-test this test is not entirely robust. This will be discussed later in Chapter 5.

2.3.3 Log ratio method

Stekel et al [27] derived a variation of the log ratio statistic for finding differentially expressed tags. Consider the differential expression of tag i over all samples $t = 1,..,T$. The total count of each sample t is represented by θ_t and the count of a given tag in a given sample is denoted by $y_i(t)$. Two hypotheses relating to the frequency of tag i are compared using a log ratio statistic. The null and alternative hypotheses for defining differential expression of a given tag between two clusters are given by:

> H_0 (null): The tag is not differentially expressed, so the frequency of the gene is the same in all samples.
>
> H_1 (alternative): The tag is differentially expressed, so the frequency of the gene is different in at least some of the samples.

It is assumed that each tag count $y_i(t)$ follows an approximately Poisson distribution with mean $\lambda_i(t)\theta_t$. The maximum likelihood estimate for λ is found using *(7)* and *(8)*, similar to finding the distance measure in 2.2.1.1. So $\hat{\lambda}_i(t)$ is given by $\sum_{t=1}^{T} \frac{y_i(t)}{\theta_t}$. This is just the

proportion of the tag of interest among all samples. The maximum likelihood of the likelihood of the observed data under the null hypothesis is given by *(38)*.

$$L_i^{null} = \prod_{t=1}^{T} \frac{e^{-\hat{\lambda}_i(t)\theta_t}\left(\hat{\lambda}_i(t)\theta_t\right)^{y_i(t)}}{y_i(t)!} \qquad (38)$$

Under the alternative hypothesis the frequency of the tag counts can be different in each sample. Therefore, the count for each tag in each sample is approximately distributed as a Poisson variable with mean $y_i(t)$. Thus the likelihood for the observed data under the alternative hypothesis becomes *(39)*:

$$L_i^{alt} = \prod_{t=1}^{T} \frac{e^{-y_{t,i}}\left(y_{t,i}\right)^{y_{t,i}}}{y_{t,i}!} \qquad (39)$$

Performing a generalised likelihood ratio test by taking the log of the ratio of the two likelihoods compares the two hypotheses: $\log\left(L_i^{alt}/L_i^{null}\right)$. This then leads to the test statistic *(40)*.

$$R_i = \sum_{t=1}^{T} y_i(t)\log\left(\frac{y_i(t)}{\theta_t\hat{\lambda}_i(t)}\right) \qquad (40)$$

From *(40)* it can be decided whether or not to reject the null hypothesis and if differential expression exists. This statistic can also be used to estimate false positive rates in the data by generating random datasets that follow the null hypothesis, and performing the analysis on these data. This gives a basis to which the original values in a given dataset can be compared.

2.3.4 Over-dispersed logistic regression model

Baggerly et al [17] suggest a method for detecting differentially expressed tags between two groups or clusters by generalising using logistic regression with over-dispersion. This is done by fitting the vector of proportions of each tag i in each sample t, denoted $\lambda_i(t)$, as a function of the given covariates (clusters) x_t.

Now the interest shifts to the form of the relationship. If $\lambda_i(t) = \beta_0 + \beta_1 x_i(t) + \varepsilon$ the relationship is linear and fitted proportions can potentially be obtained outside of the interval $[0,1]$. This then leads to fitting a transformed version of the $\lambda_i(t)'s$ being linear in the covariates. A typical choice when proportions are concerned is the logistic transformation, $\log it(\lambda_i(t)) = \log(\lambda_i(t)/[1-\lambda_i(t)]) = \beta_0 + \beta_1 x_i(t) + \varepsilon$. What is being done here is fitting a straight line to a transformed version of the data; this is analogous to the method of least squares.

An assumption typically made for least squares is that all of the observations are weighted equally, as they are all known with equal precision. However, this is not the case here as the variance of a proportion, $V(\lambda_i(t)) = \lambda_i(t)(1-\lambda_i(t))/\theta_t$, depends both on the proportion and the size of the sample from which the proportion was derived. When the observations are known with different precision, the standard amendment is to fit a weighted version of least squares. This minimises the weighted sum of the squared differences between the observations and their fitted values, where the weights are inversely proportional to the variance of each observation. A logistic curve using weighted least squares is now fitted. The weights used are inversely proportional to these initial estimators of λ, $(y_i(t)+0.5)/(\theta_t+1)$. [17]

The predicted values of the observations are obtained from this initial fit, which then suggests new values for the variances and thus the new weights. The second step is to refit the data with these new weights. This process is then repeated until convergence. In the case where over-dispersion is observed; i.e. the sizes of the squared deviations are larger than expected if the variances are of the form $V(\lambda_i(t)) = \lambda_i(t)\left(1-\lambda_i(t)\right)/\theta_t$. Here the data is said to be exhibiting over-dispersion relative to the postulated model. The estimate of the scale of the over dispersion is then required. The case of the quasi-likelihood is being dealt with here, where the variance is then of the form $V(\lambda_i(t)) = \theta_t\lambda_i(t)\left(1-\lambda_i(t)\right)\sigma_{QL}^2$ for $\sigma_{QL}^2 > 1$. Using the quasi-likelihood model for over-dispersion, the actual parameters of the best fitting model will not change. What changes, is the presumed precision associated with these parameters; the variances are multiplied by σ_{QL}^2, and significance tests need to be adjusted accordingly. To estimate σ_{QL}^2 the distribution of the sum of the squared weighted residuals is assumed to be chi-squared with $T-p$ degrees of freedom, where T is the number of samples and p is the number of β terms being estimated. The initial estimate of σ_{QL}^2 is given by *(41)*. [17]

$$\sigma_{QL}^2 = \left[\sum_{t=1}^{T} T\left(\lambda_i(t)-\hat{\lambda}_i(t)\right)^2 \Big/ V\left(\hat{\lambda}_i(t)\right)\right] \Big/ \left(T-\lambda\right) \qquad (41)$$

'Given an estimate for σ_{QL}^2 the significances can be recomputed and the p-values calculated. If the p-value is less than 0.05 then the tag is differentially expressed.' [17]

Although this method is said to work well for SAGE data [17], issues arise when it is used to analyse deep sequencing data. The weights here are calculated considering only the

sample size θ_t. The R source code was available in [17] and used in the analysis presented in Chapter 5.

2.3.5 Over-dispersed log-linear model

Lu et al [28] suggest an adaptation to the method of Baggerly et al [17] by introducing an over-dispersed log-linear model approach to assessing differential expression of tags. This model is closely linked to the model presented in 2.3.4.

To method to derive this model is based on the Gamma-Poisson hierarchical model assumption [38][28]. It is assumed that an unobserved random variable α is distributed according to *(42)*:

$$\alpha_t = Gamma(\sigma\lambda_i(t)\theta_t, 1/\sigma) \qquad (42)$$

where $\sigma > 0$, $E(\alpha_t) = \lambda_i(t)\theta_t$ and $Var(\alpha_t) = \left(\lambda_i(t)\theta_t\right)^2\sigma$. Given the proportions $\lambda_i(t)$, the response variable r_t is assumed to follow the conditional distribution.

$$r_t \mid \lambda_t \sim Po(\alpha_t) \qquad (43)$$

Working through it is found that r_t follows a negative binomial distribution i.e. $r_t \sim NegBin(\frac{1}{\sigma}, \frac{1}{\frac{1}{\lambda_i(t)\theta_t\sigma}+1})$. The unconditional mean and variance of r_t are then found to be $E(r_t) = \frac{1}{\sigma}\lambda_i(t)\theta_t\sigma = \lambda_i(t)\theta_t$ and $Var(r_t) = \frac{1}{\sigma}\lambda_i(t)\theta_t\sigma\frac{\frac{1}{\lambda_i(t)\theta_t\sigma}+1}{\frac{1}{\lambda_i(t)\theta_t\sigma}} = \lambda_i(t)\theta_t(1+\lambda_i(t)\theta_t\sigma)$.

As σ approaches 0 the $Var(r_t)$ approaches a normal Poisson variance. The mean $\mu_t = \lambda_i(t)\theta_t$ of r_t and the clusters (or covariates) x_t are connected through a log-link function *(44)*.

$$\log(\mu_i(t)) = \log(\lambda_i(t)\theta_t) = x_t\beta \qquad (44)$$

As in [17], the estimates of β are obtained by the iteratively re-weighted least squares procedure, where the weights are $1/\left[1+\mu_i(t)\sigma\right]$. In contrast to the method proposed in [17], the weights calculated in this method depend on both $\lambda_i(t)$ and θ_t. The R source code for this method was available from the additional material in [28], and used in the analysis discussed in Chapter 5.

2.3.6 Poisson mixture model

Zuyderduyn [29] proposed a Poisson mixture model similar to the methods proposed in [17] and [28] claiming it performs well as a method for assessing differential expression. It is assumed that for the observed tag, i, the counts follow a conditional Poisson distribution *(45)*.

$$y_i(t) \mid k \sim Poisson(\mu_i(t,k) = \lambda_i(t,k)\theta_t) \qquad (45)$$

where the component $k = 1,\ldots,K$ and $\lambda_i(t,k)$ is the actual expression for component k in terms of the proportion of all expressed tags. The posterior probability that an observed tag count belongs to a component k is given by *(46)*:

$$p(k \mid y_i(t),\psi) = \frac{\pi_k f(y_i(t) \mid \mu_i(t,k))}{\sum_{j=1}^{K} \pi_j f(y_i(t) \mid \mu_i(t,j))} \qquad (46)$$

where ψ is the parameter vector containing the component means and mixing coefficients $(\pi_1,\ldots,\pi_{K-1})$. $f(y_i(t) \mid \mu_i(t))$ is the probability mass function for the Poisson distribution. Maximum likelihood estimation is used to estimate the values of ψ; the

expectation maximisation (EM) algorithm is then used to fit the model. The R source code is supplied in [29] and applied in the analysis presented in Chapter 5.

2.4 Simulating the data

2.4.1 Simulation study

Lu et al [28] investigate how to simulate SAGE data. However, they do not go into much detail about the process itself - this will be investigated more in Chapter 6. The data is simulated from various distributions: Binomial, Beta-Binomial and negative Binomial. Different tag proportions were selected and different values of dispersion were chosen for both the Beta-Binomial and negative Binomial. Various methods of detecting differential expression were tested on this generated data and false positive rates were predicted.

2.4.2 Scale free networks

Khanin and Wit [30] discuss the use of the power-law distribution to assess the scale-free nature of biological networks. The most interesting property of scale-free networks is their indifference to changes in scale, i.e. the function $f(x)$ remains unchanged upon changing the scale of x. This property is often referred to as self-similarity. A network can be called scale-free if the probability that any given node is connected with k other nodes follows a power-law $P(k) \sim k^{-\gamma}$, where γ is the power-law exponent.

This power-law exponent γ is calculated using maximum likelihood to fit the power-law distribution to the data and then a goodness of fit test is performed to determine if the data is drawn from this Power-Law distribution.

Khanin and Wit [30] provide a function in the additional material of the paper that performs the test outlined above and calculates γ. Another function also provided then uses this γ to simulate data from a Power-Law distribution. The scale-free nature of deep-sequencing data is exploited here and the power-law function is used to simulate the 'true' counts in the algorithm outlined in 6.2.

Chapter 3

Preliminary Data Analysis

3.1 Looking at the data

Various analysis techniques exist for both continuous and count data obtained from sequencing. Although useful, many of the techniques developed do not take into account the large number of zero counts and the vast range of counts that appear in deep sequencing data.

3.1.1 Dataset 1

3.1.1.1 Samples

In the first dataset (data 1), the clusters were known a-priori. Samples 1-23 were in the first cluster, 24-33 the second and 34-55 the third. Due to the multidimensional nature of the data, finding a way to look at the dataset as a whole proved difficult. Sammon mapping is a form of multidimensional scaling using a distance or similarity matrix. It creates distances between the points of interest in a lower-dimensional space (usually 2-dimensional) as similar as possible to the between-point distances in the multi-dimensional space. If there is correlation between the variables (original dimensions) then points close together in the multi-dimensional space should appear close together on the Sammon map. This technique is, however exploratory, it generally involves some

distortion of relative distances between samples and so is not definitive evidence of differences or similarities. [31]

A Sammon map of the samples was plotted using both Euclidean and Manhattan distance measures; these measures have previously been used in the analysis of sequencing data such as microarray data [32]. As the clusters were known in this dataset the information shown in this map can give an indication as to whether or not the given clusters are correct. Figure 2 and Figure 3 show a large overlap of the three clusters; clusters 1 and 2 seem more isolated from one another whereas cluster 3 overlaps both clusters 1 and 2. This could be due to the fact that the cluster 3 is very similar to clusters 1 and 2. Looking at Figure 2 and Figure 3, both show very similar results, however the Manhattan distance measure seems to have identified the three clusters more distinctly. The majority of samples in cluster 3 lie in between clusters 1 and 2 while samples 4, 29, 30, 31, 51 and 52 do not appear to belong to any cluster.

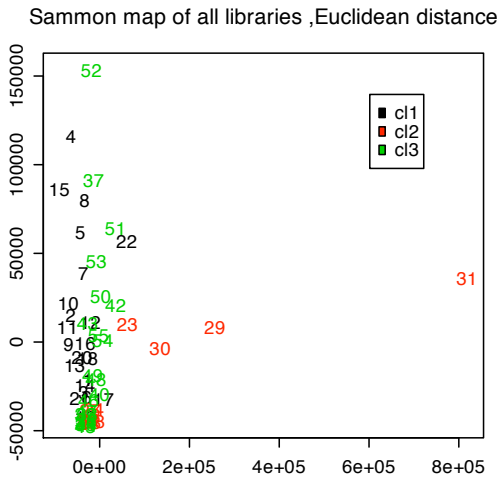


Figure 2: Sammon plot of all samples in dataset 1. Each colour represents a different cluster. Euclidean distance measure used.

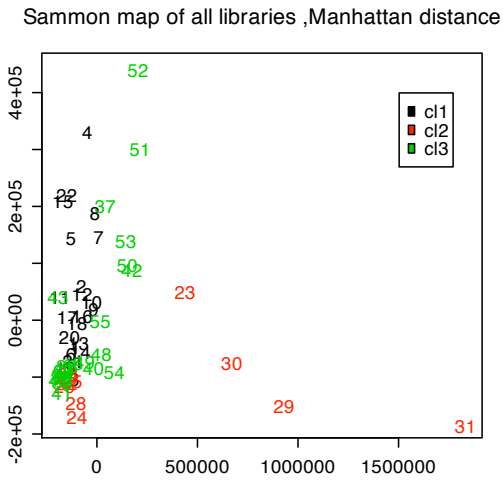


Figure 3: Sammon plot of all samples in dataset 1. Each colour represents a different cluster. Manhattan distance measure used.

In order to obtain more information each pair of clusters were mapped separately using both Euclidean and Manhattan distance measures.

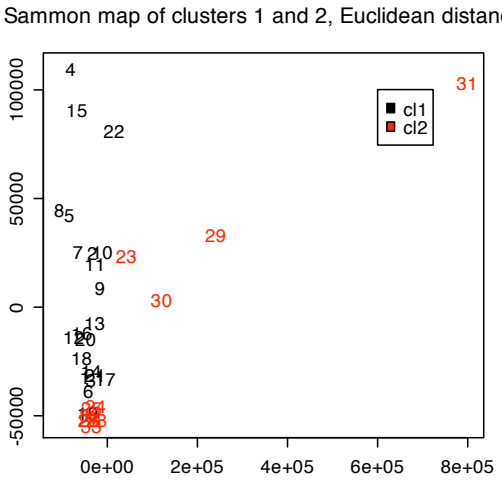


Figure 4: Sammon plot of samples in clusters 1 and 2, from dataset 1. Each colour represents a cluster. Euclidean distance measure used.

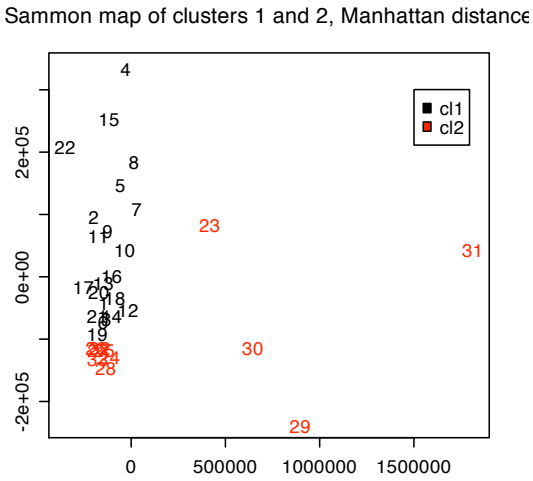


Figure 5: Sammon Plot of samples in clusters 1 and 2, from dataset 1. Each colour represents a cluster. Manhattan distance measure used.

Looking at Figure 4 and Figure 5, it appears that using both Euclidean and Manhattan distances as similarity measures, clusters 1 and 2 appear to be quite distinctly separated

with outlying samples 4, 15, 22, 29, 30 and 31. It is interesting that when clustering the samples in all of the clusters and clustering the samples in only clusters 1 and 2, sample 31 is an outlier and it appears to be distinctly different from the other samples. However, due to no biological information being known about the data it is difficult to make any conclusions as to why this may occur. The other outliers may occur because the distance measures used were not adequately sensitive. Clustering methods using different distance measures will be investigated further in Chapter 4.

In order to examine these outliers more closely scatter-plots of each of the outlying samples were plotted in Figure 6 using the *pairs()* function in R [37]. In Figure 6, each element of the plot shows two samples plotted against each other. These were plotted on a logarithmic scale so as to get a clearer picture of the data, a count of one was added so as to account for the zero counts in the data.

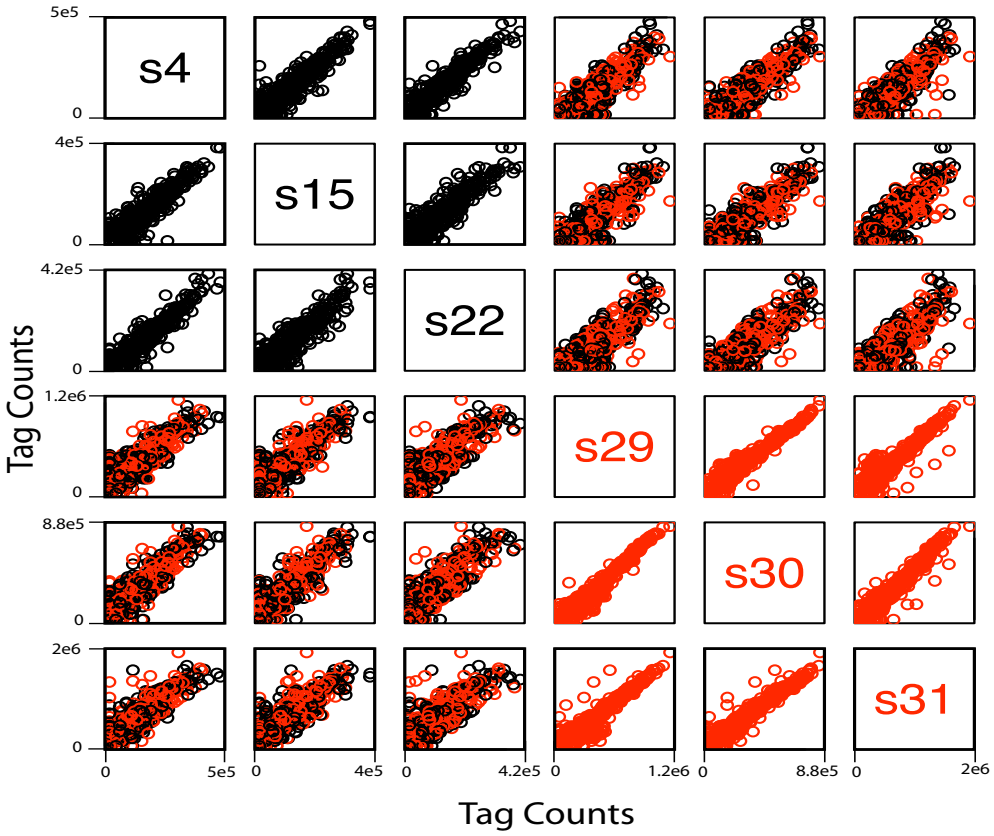
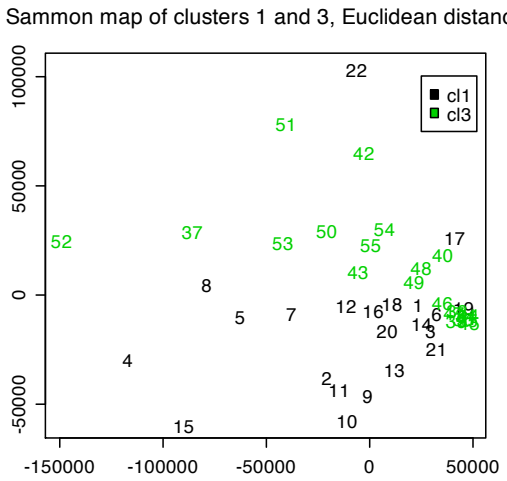


Figure 6: Pairs plot of outlying samples observed in Figure 4 and Figure 5. Each colour represents a different cluster.

When plotting different samples against each other, samples from the same cluster would be expected to group closely together, producing an almost diagonal line due to the overlapping counts. Samples from different clusters would be expected to scatter more widely. Looking at Figure 6 as expected samples 29, 30 and 31 from cluster 2 group very close together however samples 4, 15 and 22 from cluster 1 give a more scattered plot than expected. This anomaly could be due to the distance measure used. More sensitive distance measures will be investigated in Chapter 4. Looking at the plots of samples from different clusters, they are considerably more widely scattered than the same cluster sample plots, as expected.

Figure 7 and Figure 8 below illustrate the similarity of cluster 3 to both clusters 1 and 2.



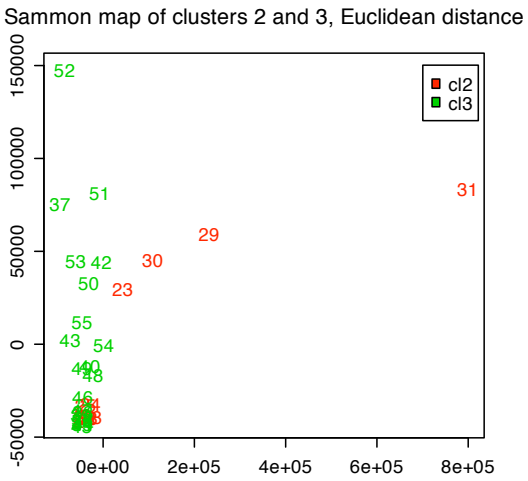


Figure 7: Sammon plot of samples in clusters 1 and 3, from dataset 1. Each colour represents different (known) clusters. Distance measure used is Euclidean.

Figure 8: Sammon plot of samples in clusters 2 and 3 from dataset 1. Each colour represents different (known) clusters. Distance measure used is Euclidean.

Looking at Figure 7, cluster 1 appears not to cluster at all and cluster 3 clusters weakly. Figure 8 illustrates the distinct similarity between clusters 2 and 3. Although there is evidence of correct clustering of some of the samples in both clusters, there is a large overlap of the two clusters. In any formal analysis this would be expected to provide no useful information. The only obvious outliers when plotting the three clusters are samples 22 (in Figure 7), 31 and 52 (in Figure 8). These are plotted below in Figure 9, which illustrates the similarity between the samples from clusters 1 and 2 (samples 22 and 31 respectively) to the sample from cluster 3 (sample 52). Looking at the scatter-plots, it is evident from the wide spread of the data that the samples 22 and 31 are not similarly distributed. However, looking at these samples plotted separately against sample 52 there is some evidence of similarity as the points group very closely together. This enforces the conclusion that cluster 3 is very similar to both clusters 1 and 2. Further analysis and investigation into this will be conducted in Chapter 4. Only Euclidean

distance was presented here, as using Manhattan distance provided very similar results, leading to the same conclusions.

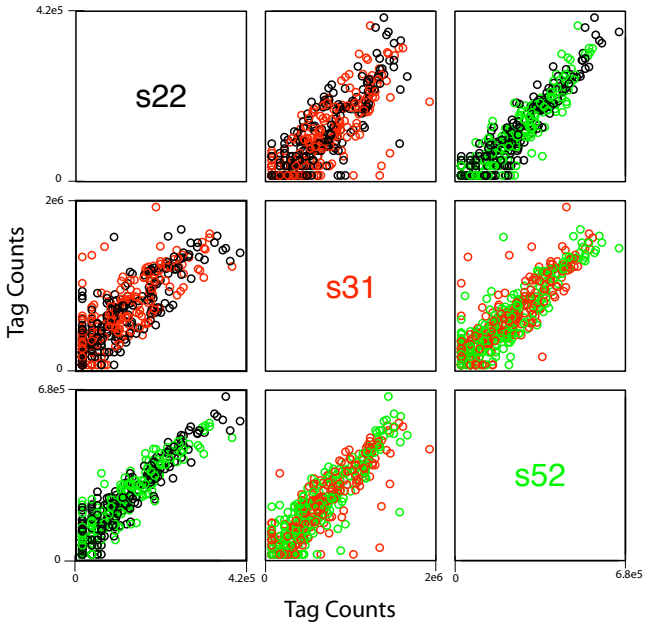
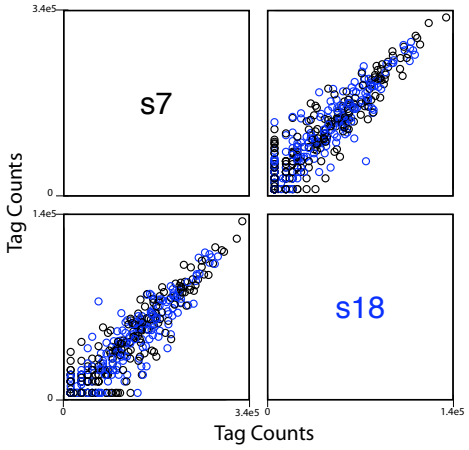


Figure 9: Pairs plot of outlying samples observed in Figure 7 and Figure 8 each colour represents a different cluster (as in Figure 7 and Figure 8).

Now the distribution of the samples has been investigated the next point of interest is correlation of the samples. The first step to accomplish this was to create a frequency matrix by dividing each element of the dataset by the sum of the column in which it was contained. A correlation matrix was then constructed using R. The most and least correlated samples were found and are plotted against each other in Figure 10 and Figure 11 below. It is expected that the two most correlated samples would be in the same cluster and the two least correlated samples would be in the different clusters. Once the correlation matrix was constructed, it was found that sample 7 and sample 18 were the most correlated and samples 17 and 31 were the least correlated. This validates the

assumption made above as the two most correlated samples are contained in cluster 1 and the two least are from two separate clusters.



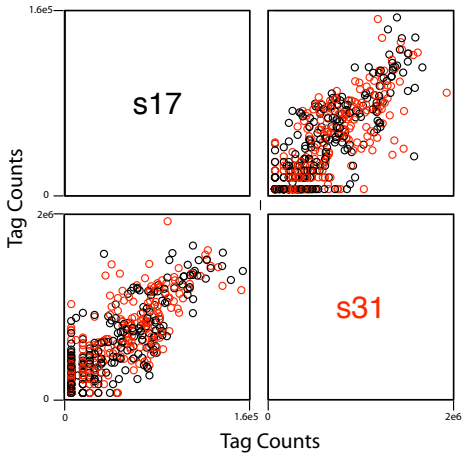


Figure 10: A pairs plot looking at the two most correlated samples in dataset 1, both from cluster 1.

Figure 11: A pairs plot looking at the two least correlated samples in dataset 1. Sample 17 is from cluster 1 and sample 31 is from cluster 2.

From the scatter plot in Figure 10 it is apparent that the two most correlated samples are reasonably similar, as the data is not widely spread. In contrast, Figure 11 shows the two least correlated samples. It is clear from the wide spread nature of the data that the two least correlated samples are considerably different.

Figure 12 and Figure 13 show the frequency distribution of the tags in the most and least correlated samples respectively. This was done to investigate whether there is any visual difference in the distribution of tags in these samples. In order to get an informative look at the data the graphs show only the tags that have a count of less than 100. This is due to the fact that the range of the counts goes so high but the majority of the tags have a count of less than 100.

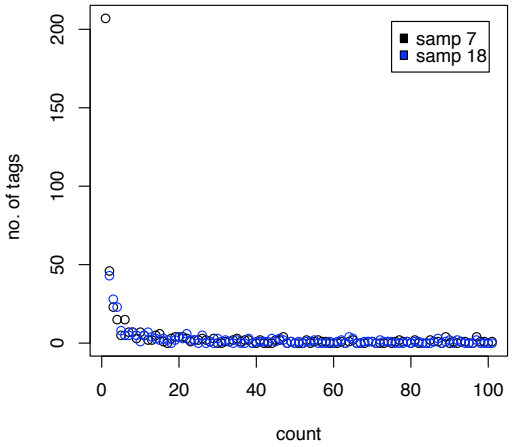


Figure 12: Frequency distribution of tag counts in the two most correlated samples of dataset 1.

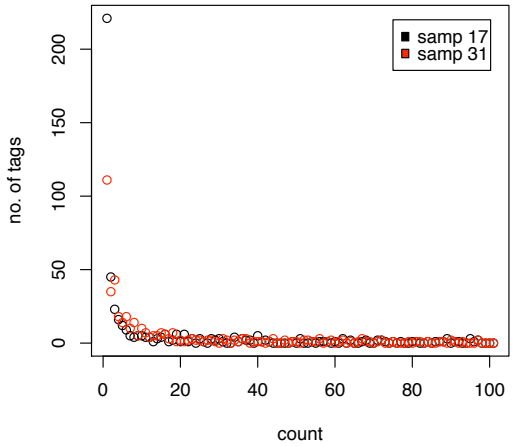


Figure 13: Frequency distribution of tag counts in the two least correlated samples of dataset 1.

From Figure 12 it is evident that the frequency distribution of the two most correlated samples are very similar. In Figure 13 there is slight evidence of a difference between the two least correlated samples. However the differences between the two samples are not large.

From both Figure 12 and Figure 13 that more than half of tags have a count of zero and the those which don't have a zero count, have a count of between zero and fifty. In order to take a closer look at the distribution of tags in the most and least correlated samples, plots of the frequency distribution with counts between one and fifty were constructed and shown below. Looking at Figure 14 it is noticeable that the two most correlated samples have a very similar frequency distribution as expected. However, looking at Figure 15, although samples 17 and 31 are the least correlated there is slight evidence of a difference in the frequency distribution of tags but this difference does not appear to be large. Analysis of how tags are expressed between samples is investigated further in Chapter 5.

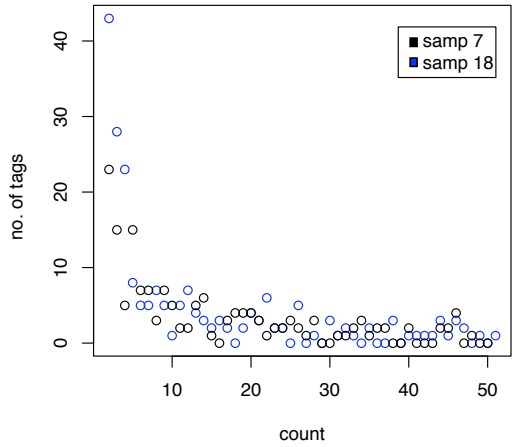


Figure 14: A closer look at the frequency distribution of tag counts in the two most correlated samples in dataset 1 (both from cluster 1). All counts between 1 and 50 are shown.

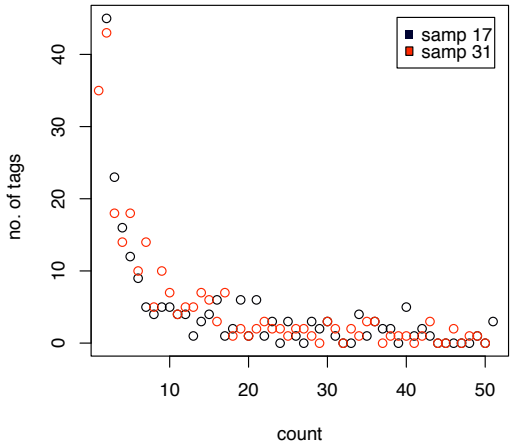


Figure 15: A closer look at the frequency distribution of tag counts of the two least correlated samples in dataset 1 (both from different clusters). All counts between 1 and 50 are shown

3.1.1.2 Tags

As no a-priori information was given about tags and due to the large number of tags, a Sammon map would be somewhat uninformative. Clustering of tags will be investigated further in Chapter 4, using more sensitive distance measures and different clustering methods.

As in 3.1.1.1, a correlation matrix was made to find the two most correlated tags. The frequency distributions were plotted in Figure 16 and Figure 17 below. Figure 16 shows a clear similarity between the two most correlated tags as would be expected. Looking at Figure 17 an unmistakable difference can be observed between the two least correlated tags. When investigating clustering of tags, tag 581 and tag 10 would be expected to be in different clusters. To get a better idea of the distribution of the most and least correlated tags over all samples, Figure 18 and Figure 19 were constructed to illustrate the count of each of these tags in each sample.

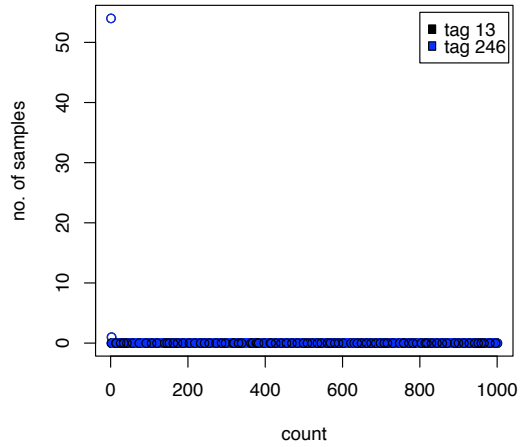


Figure 16: Frequency distribution of sample counts for the most correlated tags in dataset 1.

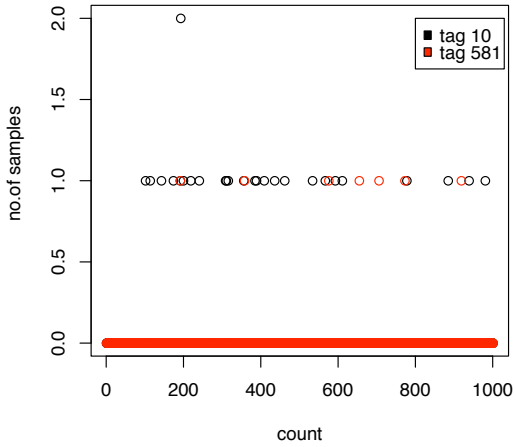


Figure 17: Frequency distribution of sample counts for the least correlated tags in dataset 1.

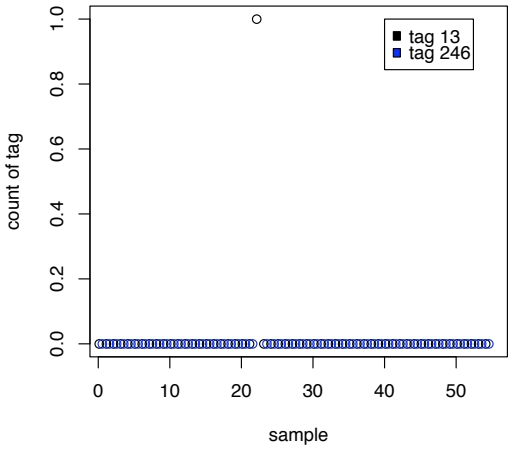


Figure 18: A plot of tag counts over all samples for the two most correlated samples.

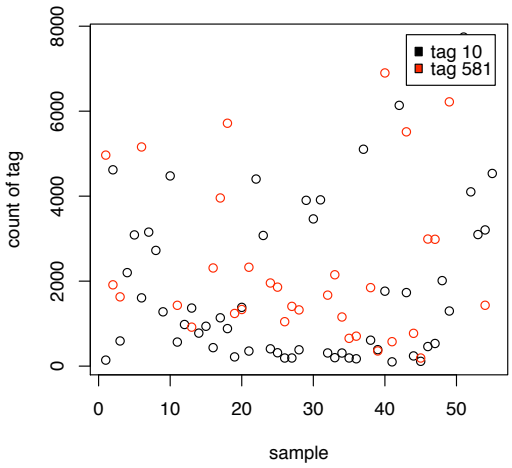


Figure 19: A plot of tag counts over all samples for the two least correlated samples.

From Figure 18 the two most correlated tags, tag 13 and tag 246, give the impression of being distributed identically, as expected. However tag 13 appears once, this count is likely to be a false positive i.e. a count recorded as one that should have been zero. This will be investigated further in the simulation study Chapter 6.

It is important to study tags to investigate the differential expression of tags between samples and to investigate false positive results in the data. Differential expression will be investigated in Chapter 4 and false positives will be investigated in Chapter 7.

3.1.2 Dataset 2

3.1.2.1 Samples

In the second dataset (dataset 2) no information is given a-priori about the dataset. In an attempt to loosely predict any clustering of the samples a Sammon map was constructed using both Euclidean and Manhattan distance measures.

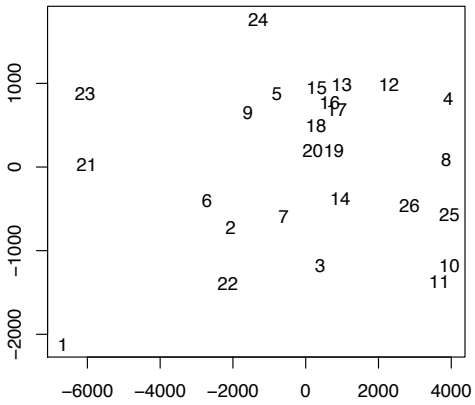


Figure 20: Sammon map of samples in dataset 2 where no clusters are known a-priori. Euclidean distance used.

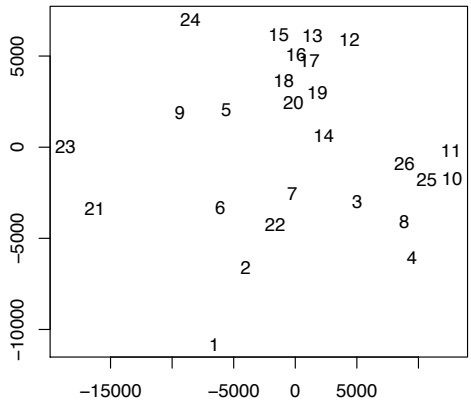


Figure 21: Sammon map of samples in dataset 2 where no clusters are known a-priori. Manhattan distance used.

Looking at both Figure 20 and Figure 21, there is some weak grouping occurring, but this is not conclusive enough to say there is any concrete evidence of clustering. Both Euclidean and Manhattan distance measures may not be sensitive enough to detect the clusters compared to other distance measures which will be investigated in Chapter 4. It is evident from both plots that samples 1,21,23 and 24 are outliers. A pairs plot was

constructed on a log scale plotting each of these outliers against each other to investigate the relationships between the outliers.

Looking at the spread of the data in each of the plots in Figure 22, it can be inferred that sample 1 is in a different cluster than samples 21,23 and 24. This inference is made on the basis that the spread of the data when samples 21,23 and 24 are plotted against sample 1 is considerably wider than when samples 21,23 and 24 are plotted against each other.

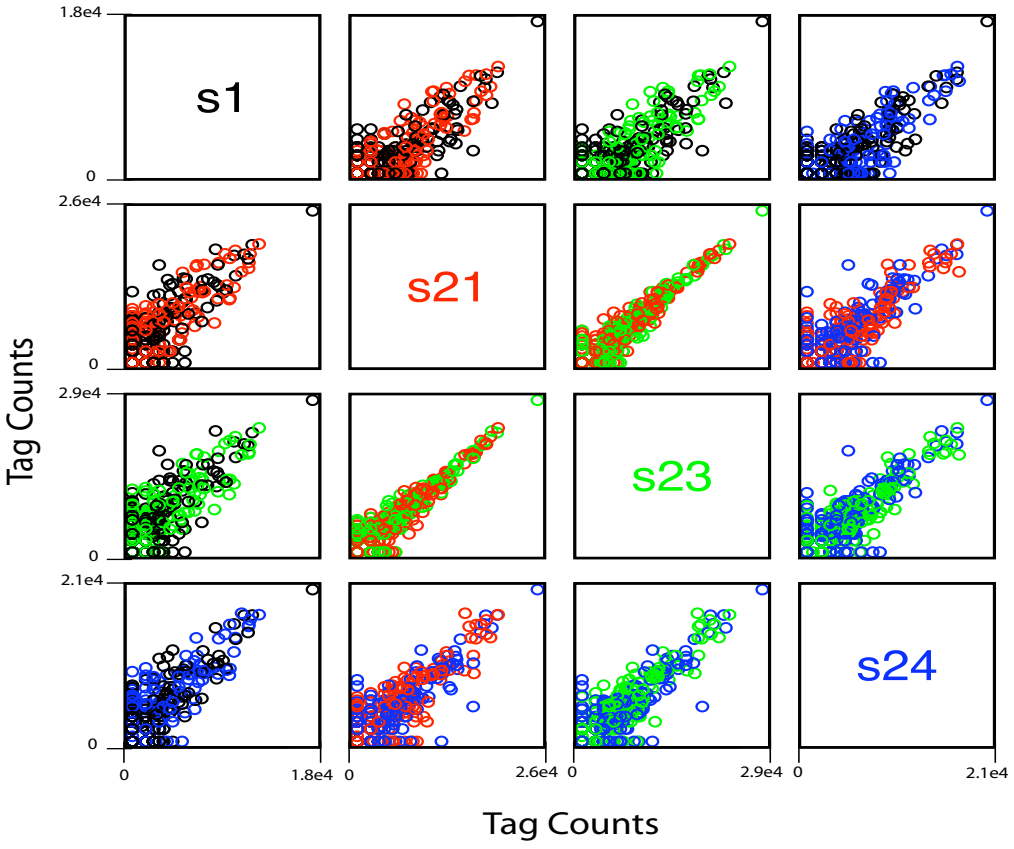


Figure 22: Pairs plot of outlying samples observed in Figure 20 and Figure 21, a different colour was used for each sample as no clusters were known.

As was done for dataset 1, a correlation matrix of samples was constructed and the two most correlated samples were found to be samples 16 and 22, while the two least correlated samples were found to be samples 11 and 24. Although no information is

known about the clustering in this dataset, it is predicted that samples 16 and 22 belong to the same cluster and samples 11 and 24 different clusters. Figure 23 and Figure 24 illustrate the two most and least correlated samples plotted against each other on a logarithmic scale.

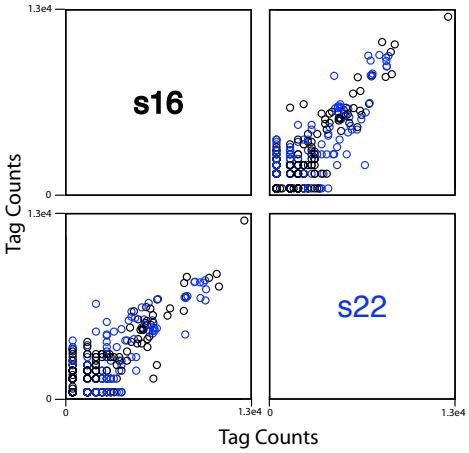


Figure 23: Pairs plot of the two most correlated samples in dataset 2.

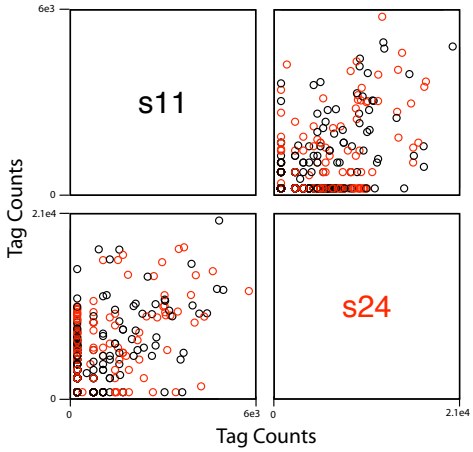


Figure 24: Pairs plot of the two least correlated samples in dataset 2.

From Figure 23 it can be observed that the data is grouped reasonably close, as would be expected of two very similar samples. In contrast, looking at Figure 24 the data is very widely spread suggesting a difference between the two least correlated samples as would be expected.

The frequency distribution of tag counts in both the two least and two most correlated samples was plotted below. As in 3.1.1.1 a count of 100 was chosen as the cut-off due to very few tags in each sample having a count greater than 100. Looking at Figure 25 and Figure 26, there is no apparent difference between the frequency distribution of tag counts in the two most and two least correlated samples. However, this could be due to the high number of tags that have a zero count and the concentration of tags between

the counts of 1 and 50. In order to make a better comparison of the samples the frequency distribution of tag counts was plotted only for counts between 1 and 50 for both the most and least correlated samples in Figure 27 and Figure 28.

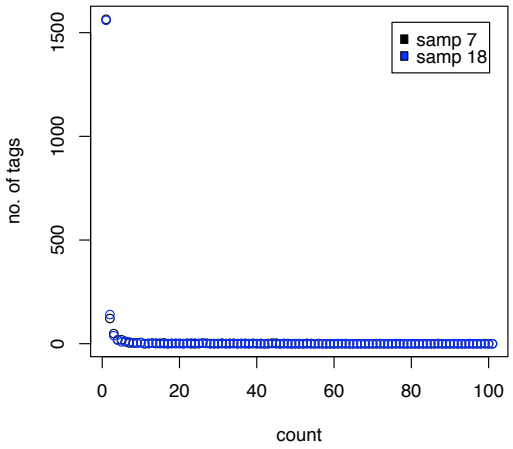


Figure 25: Frequency distribution of tag counts for the two most correlated samples in dataset 2.

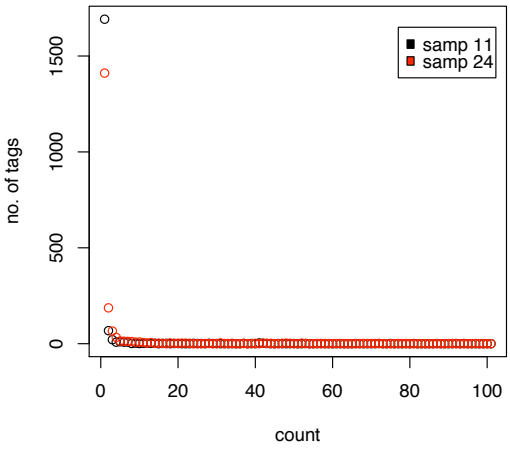


Figure 26: Frequency distribution of tag counts for the two least correlated samples in dataset 2.

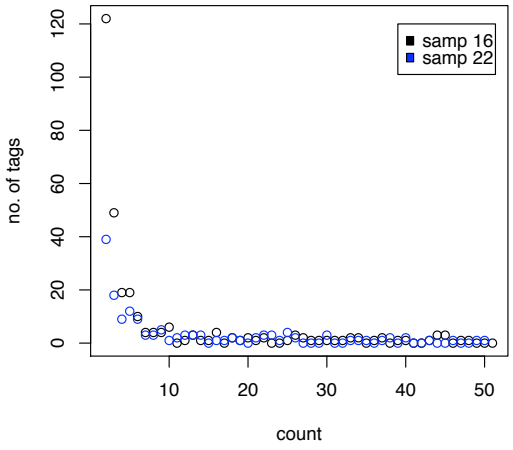


Figure 27: A closer look at the two most correlated samples in dataset 2. All counts between 1 and 50 are shown.

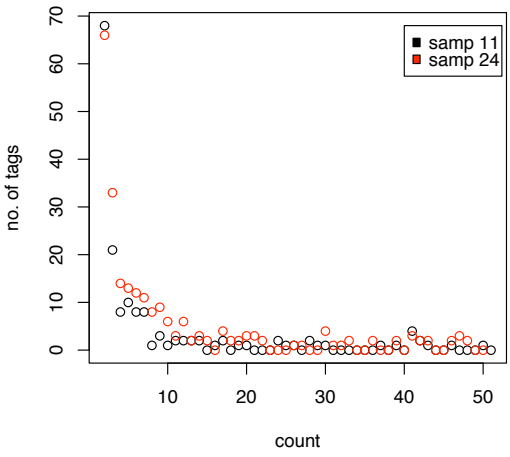


Figure 28: A closer look at the two least correlated samples in dataset 2. All counts between 1 and 50 are shown.

Looking at these plots, the similarity of samples 16 and 22 is evident as, in Figure 27, there is very little deviation of the two in the frequency distribution of tag counts. There is some evidence of difference between samples 11 and 24 as, in Figure 28, the frequency

distribution of tag counts for each sample deviate from one another. Differential expression of tags between clusters will be investigated further in Chapter 5.

3.1.2.2 Tags

As for dataset 1 no a-priori information was known about the grouping of tags. However, in this dataset there is a considerably larger variety of tags – more than triple that in dataset 1. Due to the abundance of tags and also the large number of tags that have low levels of expression, a Sammon map would prove entirely uninformative for predicting any patterns in the tag expression. This will be investigated further in Chapter 4 and Chapter 5.

As above, a correlation matrix of tags was constructed and the two most and least correlated tags were found to be tag 920 and 921 and tag 1551 and 1496 respectively. Frequency distributions of sample counts for these tags were plotted. As expected the frequency distribution of the sample counts for the most correlated tags are almost identical, whereas for the two least correlated tags the frequency distributions vary dramatically.

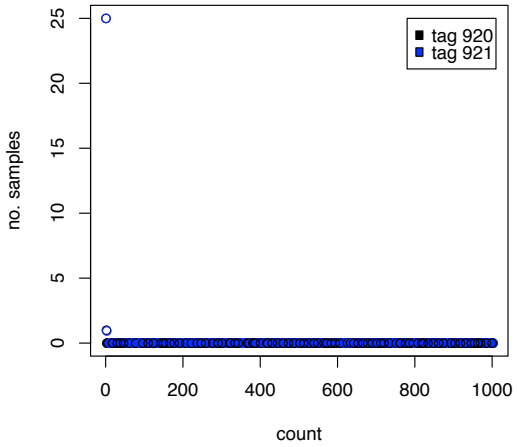


Figure 29: Frequency distribution of sample counts for the most correlated tags in dataset 2.

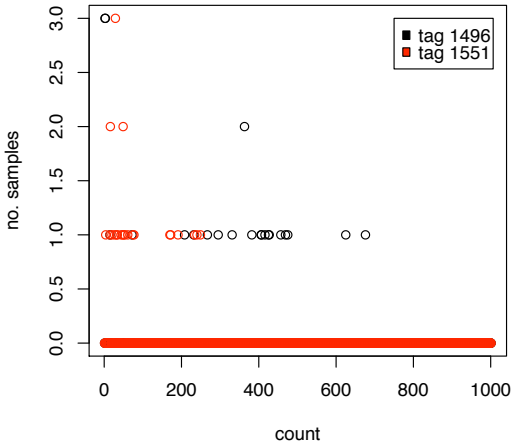


Figure 30: Frequency distribution of sample counts for the least correlated tags in dataset 2.

To get a better idea of the distribution of these tags over all samples, the count of the most and least correlated tags in each sample were plotted in Figure 31 and Figure 32. As anticipated, the two most correlated tags are identically distributed across all samples. For the two least correlated samples the counts of the tags are more scattered across all of the samples.

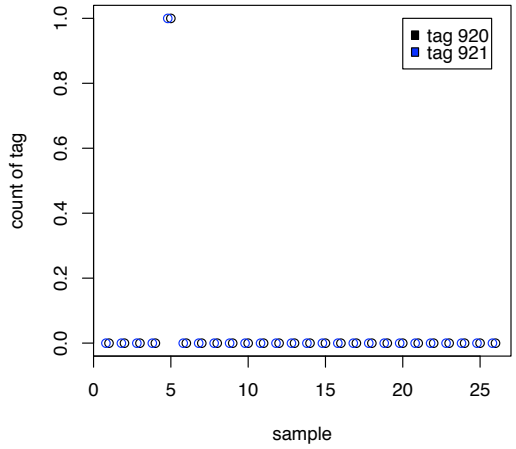


Figure 31: A plot of tag counts over all samples for the two most correlated samples in dataset 2.

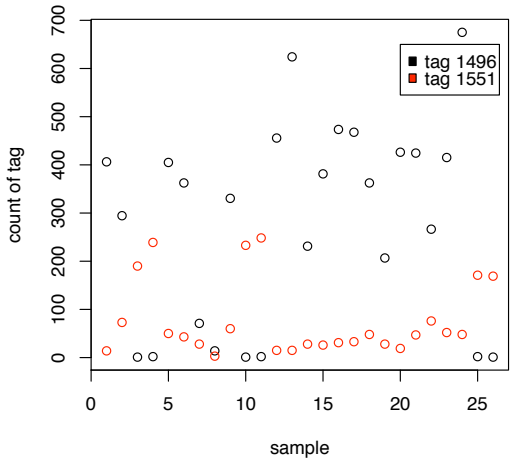


Figure 32: A plot of tag counts over all samples for the two least correlated samples in dataset 2.

3.2 Subjective impressions

Although in dataset 1 the clusters were known a-priori, initial inspection of the data suggests that cluster 3 is not drastically different from the other two. Different types of algorithms and different, more sensitive distance measures can be used to further separate the three clusters. It is likely that clustering of the data known to be in clusters 1 and 2 only will cluster distinctly into two clusters. However clustering of the entire dataset or cluster 3 with either cluster 1 or 2 is expected to give incorrect results due to the similarity of cluster 3 to the other two. It is probable that the expression profile of cluster 3 is too similar to that of clusters 1 and 2 to separate distinctly.

In dataset 2 there is no clear indication of distinct clusters. However it is expected that when using different clustering methods and distance measures, different clusters will be identified. Only 2 clusters are anticipated in this dataset due to the low number of samples. This was confirmed when applying the clustering algorithm discussed in Chapter

4 to the dataset, if more than 2 clusters were entered into the algorithm the samples repeatedly jumped from cluster to cluster and the algorithm did not converge. When the algorithm was applied to the dataset with two clusters entered the algorithm converged.

When looking at tags in both datasets using standard distance measures, there do not appear to be any distinct clusters. Using more sensitive distance measures and techniques developed specifically for the clustering of sequencing data is expected to distinguish distinct clusters of tags, which can then provide information on specific groups of tags that occur more frequently in cancerous and non cancerous tissues. Although no a-priori information was given about tags, various methods of identifying dissimilarities in expression profiles should detect different levels of tag expression between samples and between clusters of samples.

It is expected that samples from different clusters will have notably different expression profiles i.e. different groups of tags will be differentially expressed in samples from different clusters. Samples from the same cluster are expected to have more similar expression profiles. More traditional significance tests such as the 2-sample t-test are not sensitive enough to detect the levels of differential expression expected. Many techniques have been developed to assess differential expression both between individual samples and groups of samples. These will be evaluated in Chapter 4.

Chapter 4

Clustering

4.1 Overview

In the data analysed in this thesis several different tissue samples have been sequenced and the expression of the tags recorded. When data consists of several different samples, the first point of interest is whether any of these samples can be grouped together in homogeneous categories. These categories are a result of the differential expression of individual tags between samples. In order to identify these, clustering has to be performed.

Very few theories about clustering are concrete and definitive. The two most common ideas of what constitutes a cluster are internal structure and external separation. [32]

There exist many different clustering techniques that can be used to cluster both samples and tags to find patterns of interest in the data. In both cases clustering can be useful for a variety of reasons.

The central goal of clustering samples is to identify significant changes in tag expression between them. By dividing the samples into dissimilar groups of individuals (clusters), tag expression can be related to a specific response [32]. For example, in the given datasets ideally the tag expression in the cancerous tissue samples will be considerably different

from the tag expression in the non-cancerous tissue samples and therefore would cluster separately.

When looking at clustering of tags, the first point of interest is to reduce the quantity of information acquired due to the large number of individual tags sequenced. Clustering of the tags can make the vast quantity of information more controllable and also to distinguish if tags that are known to be similar have similar expression profiles. [32] Clustering tags with similar expression profiles can allow biologists to investigate the function and relevance of the tags with different expression profiles. [22]

There are many clustering techniques available; the main features required from any clustering technique are adaptability to different distance measures and the ability to deal with the high-dimensional and sparse nature of the data [32]. The two methods of clustering that will be explored in this thesis are ‘k-means clustering’ and ‘hierarchical clustering’.

‘K-means clustering’ aims to cluster a given number of observations (could be samples or tags) into the cluster with the closest mean. The method works by randomly assigning observations to one of k clusters and repeatedly moving the observations to the cluster with the closest mean until convergence. The main drawback to this method is that the number of clusters, k, must be specified beforehand and doing this incorrectly can produce the wrong results.

‘Hierarchical clustering’ works by linearly ordering observations that are being clustered. The most common type of hierarchical clustering is agglomerative. This works by first assigning each observation to a separate cluster and then, using a distance measure,

assigning the observations that are closest together to the same cluster. The main benefit of this method is that the number of clusters does not need to be pre-assigned and any kind of distance measure can be used. [32]

The reliability of any clustering technique is almost exclusively dependant on the distance (or similarity) measure chosen. The two most common distance measures used in any type of clustering are Euclidean and Manhattan distances. These measures have previously worked for the analysis of sequencing data following a normal distribution provided by sequencing methods such as microarray analysis. [32] However, the data provided by deep sequencing is count data due to the sampling nature of the sequencing process. Due to the discrete nature of this count data, distance measures previously used on microarray experiment analysis will not be suitable. Various distance measures have been introduced for use in the analysis of SAGE data, similar in nature to deep sequencing data. Measures developed for use in the cluster analysis of SAGE data are more statistically valid for deep sequencing data than those developed for microarray data, as they are more sensitive to the structure of the data. These measures were introduced in Chapter 2 and will be evaluated later in this chapter.

The two clustering methods investigated in this thesis are those developed by Cai et al [22] and Berninger et al [24]. These methods have been introduced and outlined in Chapter 2. Adaptations made to the algorithms and evaluations of the techniques are presented later in this chapter. Various different distance measures and models for the data have been investigated and the analysis is presented later in this chapter. These have all been introduced and outlined in Chapter 2.

4.2 Adaptations

4.2.1 PoissonC / PoissonL algorithm

In the original clustering algorithm presented by Cai et al [22], the data was modelled using a Poisson distribution and two distance measures, likelihood and chi square, were assessed. While the paper proves the reliability of this technique for clustering of tags in SAGE data, after translating the algorithm from the paper into R it was clear that in order to be used for the data produced by deep sequencing some alterations were needed. The aim was to create a function that could cluster into any given number of clusters using various distributions to model the data and different distance measures.

Similar to that presented in 2.2.1 the algorithm works based on a k-means principle and is outlined below. The algorithm is presented in terms of clustering of samples. However it can just as easily be used for the clustering of tags. The simplest way to achieve this is to transpose the input matrix so tags are columns and samples are rows.

1. The number of clusters K is selected a-priori.
2. A distribution is chosen to model the data. This distribution can be any of the Poisson, Negative Binomial or the Zero-Truncated Poisson.
3. A distance measure is selected. This can be any of likelihood, chi-square or trans chi-square.
4. $\hat{\theta}_t$ is calculated *(8)* for each individual sample and each sample is randomly assigned to a cluster.
5. A while loop is started and runs until convergence. Initialisation $r = 0$.

6. Cluster centres λ_k^r are calculated; this is a vector of length the number of tags; each element representing the value of λ_k^r for each tag over all samples in cluster k. If the chosen distribution is Poisson or Negative Binomial λ_k^r is calculated using *(8)*. If the Zero-Truncated Poisson is used λ_k^r is calculated using Newton's method to solve *(47)* for $\hat{\lambda}$.

$$\bar{y}(t) = \hat{\lambda}(t)\theta_t(1 - e^{\hat{\lambda}(t)}\theta_t) \qquad (47)$$

7. A for-loop is initialised to run through each sample individually.
8. For each sample the chosen distance measure is calculated for each cluster k. If the Poisson distribution has been chosen the chi-square and likelihood distance measures are calculated by *(48)* and *(49)* respectively, where $\hat{E}(y_i(t)) = \hat{\lambda}_k^r(i)\hat{\theta}_t$, denoting the expected value of a given tag i in a given sample t in cluster k:

$$S = \sum_i \sum_{t=1}^{T} \left(y_i(t) - \hat{E}(y_i(t))\right)^2 \Big/ \hat{E}(y_i(t)) \qquad (48)$$

$$L_{t,k} = -\log\left(\sum_i \frac{\exp\left(-\theta_t\lambda_k^r(i)\right)\left(\theta_t\lambda_k^r(i)\right)^{y_i(t)}}{y_i(t)!}\right) \qquad (49)$$

The method for calculating the trans chi-square distance measure has been outlined in Chapter 2 and is calculated using *(50)*:

$$S_{trans} = \sum_i \sum_{t_1 t_2} \left((y_i(t_1) - y_i(t_2)) - E(y_i(t_1) - y_i(t_2))\right)^2 \Big/ Var(y_i(t_1) - y_i(t_2)) \qquad (50)$$

where:

$$E(y_i(t_1) - y_i(t_2)) = \left(\lambda_i(t_1) - \lambda_i(t_2)\right)\theta_t \qquad (51)$$

$$Var(y_i(t_1) - y_i(t_2)) = \left(\lambda_i(t_1) + \lambda_i(t_2)\right)\theta_t \qquad (52)$$

If the distribution chosen is Negative Binomial, the chi-square and trans chi square distance measures are equivalent to that calculated for the Poisson. This is only true for these two distributions, as the expected value of a random variable following a Poisson distribution $y_i(t) \sim Po(\theta_t \lambda_i(t))$ is equivalent to that of a random variable following a negative binomial $y_i(t) \sim NegBin(\theta_t \lambda_i(t), \phi)$ distribution i.e. $E(y_i(t)) = \lambda_k^r(i)\theta_t$. *(53)* is used to calculate the likelihood distance measure for the Negative Binomial.

$$L_{t,k} = -\log\left(\sum_i \frac{\Gamma\left(y_i(t) + \phi^{-1}\right)}{\Gamma\left(\phi^{-1}\right)\Gamma\left(y_i(t)+1\right)}\left(\frac{1}{\left(1+\theta_t\lambda_k^r(i)\right)}\right)^{\phi^{-1}}\left(\frac{\theta_t\lambda_k^r(i)}{\phi^{-1}+\theta_t\lambda_k^r(i)}\right)^{y_i(t)}\right) \qquad (53)$$

where ϕ is the dispersion. There are various ways to estimate this: the ones assessed here are outlined in 2.1.3.

If the Zero-Truncated Poisson is selected the Chi-Square and Trans Chi-Square distances differ only due to the expected value being different, for the Zero-Truncated Poisson $E(y_i(t)) = \lambda_k^r(i)\theta_t \Big/ \left(1 - e^{\lambda_k^r(i)\theta_t}\right)$. The Likelihood distance measure is calculated using *(54)*.

$$L_{t,k} = -\log\left(\sum_i \frac{\exp\left(-\theta_t\lambda_k^r(i)\right)\left(\theta_t\lambda_k^r(i)\right)^{y_i(t)}}{\left(1-\exp\left(-\theta_t\lambda_k^r(i)\right)\right)y_i(t)!}\right) \qquad (54)$$

9. The sample is then assigned to the cluster to which the chosen distance measure is minimised.
10. New cluster centres λ_k^r are then calculated each time a sample is reassigned.
11. Steps 8-10 are then repeated for all samples individually until end of for loop.
12. Steps 6-11 are repeated until the algorithm converges, i.e. the clusters calculated in this iteration are equal to those calculated in the previous, and returns the clusters. However there is a special case where there are one or two samples that constantly jump between clusters preventing convergence. In this case, once 1000 iterations have passed, and if less than 5% of samples are constantly jumping between clusters the algorithm removes these samples and identifies them as outliers.

4.2.2 Bayesian algorithm

The method presented by Berninger et al [24] was developed for the clustering of small RNA expression profiles and as such would appear to be perfect for the clustering of the datasets provided.

Although no alterations were made to the algorithm [24], many problems were encountered when translating the algorithm from the paper into R. The main issue encountered was the calculation of the two likelihoods using equations *(18)* and *(19)*. Due to the high-count nature of the data the gamma functions in these equations could not be calculated directly so the log of each of the equations was calculated to make the computation possible. This proved mathematically awkward due to the abundance of zero tags in the dataset, as once this had been done problems were encountered when inserting the two logged likelihoods into the distance formula *(20)*.

Another problem encountered was when assigning the Dirichlet prior *(17)*. It was not made clear in the paper if the value of the pseudo count of the prior α was tag specific or was a constant throughout. After consultation with one of the authors the value of α was set at 0.05.

4.3 Results

4.3.1 PoissonC / PoissonL algorithm

4.3.1.1 Dataset 1

First the PoissonC / PoissonL algorithm was tested on dataset1 using the various distributions and distance measures to assess the algorithm's reliability. The algorithm was constructed in the R statistical computing language as a function in which the user inputs the dataset, the required number of clusters K, the number of loops the algorithm should run for (default=100), the desired distance measure and the distribution.

When the Negative Binomial distribution was used the dispersion parameter ϕ was calculated using the pseudo likelihood and quasi-likelihood methods outlined in 2.1.3. When testing the two methods it was found that, for these particular datasets, only the pseudo-likelihood method worked in the algorithm. When solving equation *(5)* to calculate the dispersion for each tag it was found that some of these values again did not work in the clustering algorithm so a common dispersion for all tags was found by calculating the dispersion for each tag and finding the mean of these values. This is all calculated in the algorithm itself for ease of use.

The algorithm was first tested on the entire dataset in which the information given states that there are three clusters. The results were recorded and collated below in Table 2. If the information given a-priori about the clusters is correct samples 1-22, samples 23-33 and samples 34-55 should appear in distinct clusters separately with no overlap. The algorithm was run three times for each condition and the same results were generated.

Table 2: Results from PoissonC / PoissonL clustering of the entire dataset 1. The dataset was clustered separately using each distribution with each distance measure. Cluster 1 consists of 22 samples; cluster 2, 11 samples and cluster 3, 22 samples.

	Cluster1		Cluster2		Cluster3	
Distribution & Distance measure	#Correct samples in cluster	#Wrong samples clustered alongside	#Correct samples in cluster	#Wrong samples clustered alongside	#Correct samples in cluster	#Wrong samples clustered alongside
Poisson						
Likelihood	22	4	3	0	18	8
Chi-Square	22	5	3	0	17	8
Trans-Chi	21	3	5	0	19	7
Negative Binomial						
Likelihood	17	2	3	0	20	13
Chi-Square	22	6	2	0	16	9
Trans-Chi	20	4	4	0	18	9
Zero-Truncated Poisson						
Likelihood	22	5	3	0	17	8
Chi-Square	15	5	9	5	13	8
Trans-Chi	20	3	4	1	19	8

As expected, the results indicate that there is definite overlap between the three clusters. The results presented in Table 2 are illustrated in Figure 33, Figure 34 and Figure 35 below, where each method has been shown with the samples from each different cluster highlighted in a different colour. Looking at these figures the overlap between the three clusters is evident which suggests that the clusters are very similar in nature.

Applying the algorithm with all distributions and distance measures, the results show that the samples did not cluster according to the pre-designated clusters using any of the options available. From Table 2, it would appear that the trans chi-square distance measure is the most effective in this case, particularly when used in conjunction with the Poisson distribution.

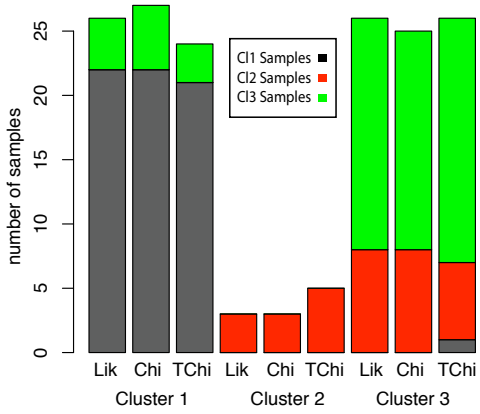


Figure 33: Bar chart showing the distribution of the samples using each distance measure for Poisson in the clustering algorithm on all 3 clusters.

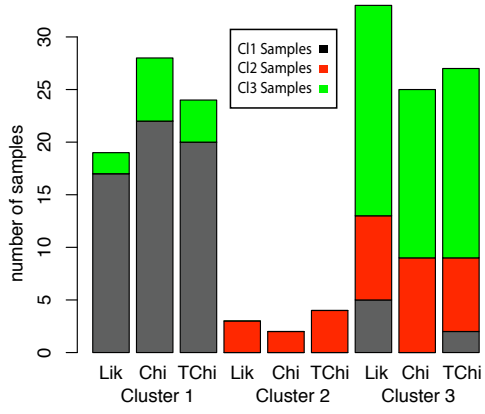


Figure 34: Bar chart showing the distribution of the samples using each distance measure for Negative Binomial in the clustering algorithm on all 3 clusters.

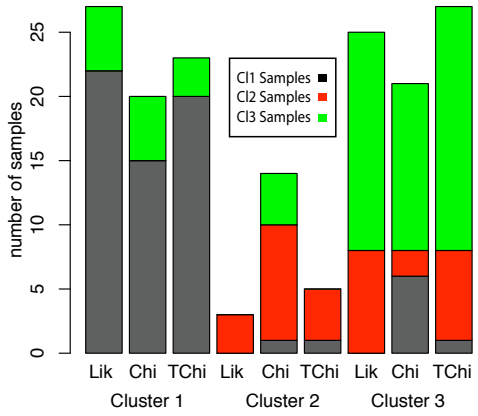


Figure 35: Bar chart showing the distribution of the samples using each distance measure for Zero Truncated Poisson in the clustering algorithm on all 3 clusters.

Sammon plots are given in Figure 36, 37 and 38 which show an approximation of the similarity between each samples in the dataset for each distribution using the Trans Chi Square similarity measure. It is clear from the overlap of the samples in the three clusters in these plots that there is a definite similarity between the samples in the three clusters. It is a distinct possibility that, using any method of clustering, the samples in the three clusters are too similar in nature to cluster distinctly. Another possibility is that the information given about the samples in each cluster is wrong. This however is speculation as no other information was given about the data.

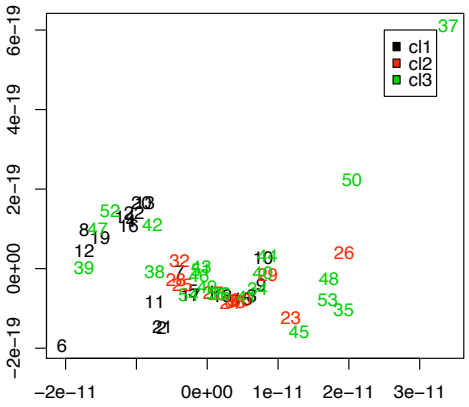


Figure 36: Sammon plot of all clusters. Distribution used is Poisson and distance measure used is Trans Chi Square.

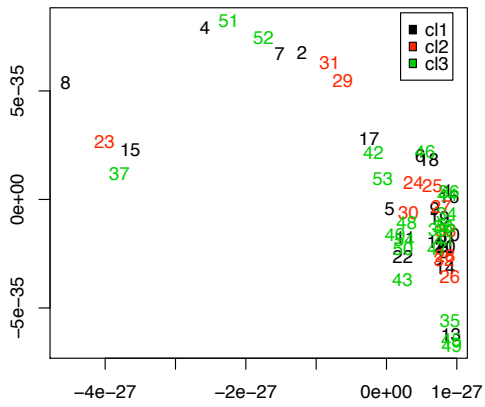


Figure 37: Sammon plot of all clusters. Distribution used is Negative Binomial and distance measure used is Trans Chi-Square

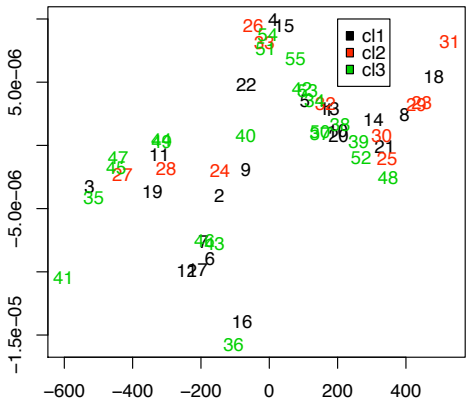


Figure 38: Sammon plot of all clusters. Distribution used is Zero-Truncated Poisson and distance measure used is Trans Chi-Square.

In order to obtain more information, each pair of clusters were investigated separately. Looking at the results in Table 3 it is clear that there is a well-defined dissimilarity between the samples contained in cluster 1 and those contained in cluster 2. Using all distributions and all distance measures the clusters were identified correctly as was expected from initial analysis of the data. These results suggest that the samples contained in clusters 1 and 2 definitely come from two distinctly separate groups of

individuals. Although no information has been given about the samples other than their groups it is possible that the samples contained in clusters 1 and 2 are the results from cancerous and non-cancerous tissue samples that have been sequenced and the algorithm has clustered these correctly.

Table 3: Results from PoissonC / PoissonL clustering of samples contained in clusters 1 and 2. The samples were clustered using each distribution with each distance measure. Cluster 1 contains 22 samples and cluster 2 contains 11 samples.

	Cluster1		Cluster2	
Distribution & Distance measure	#Correct samples in cluster	#Wrong samples clustered alongside	#Correct samples in cluster	#Wrong samples clustered alongside
Poisson				
Likelihood	22	0	11	0
Chi-Square	22	0	11	0
Trans-Chi	22	0	11	0
Negative Binomial				
Likelihood	22	0	11	0
Chi-Square	22	0	11	0
Trans-Chi	22	0	11	0
Zero-Truncated Poisson				
Likelihood	22	0	11	0
Chi-Square	22	0	11	0
Trans-Chi	22	0	11	0

A bar chart illustrating the results in Table 3 is displayed below along with a Sammon plot investigating the clustering of the samples using total likelihood, Chi Square and Trans Chi Square similarity measures and modelling the data with a Poisson distribution.

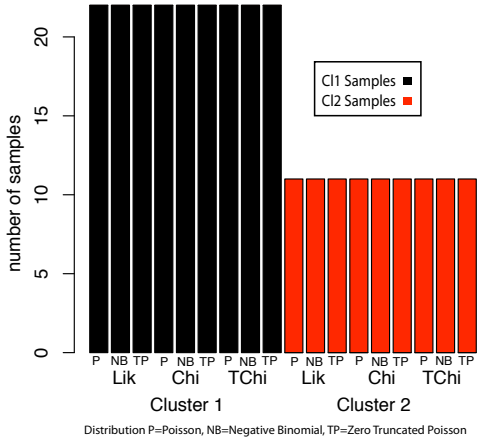


Figure 39: Bar plot of clustering results for clusters 1 and 2 using each distribution and each distance measure.

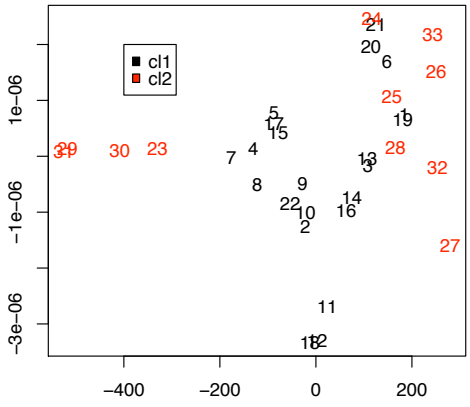


Figure 40: Sammon plot of clusters 1 and 2. Distribution used is Poisson and distance measure used is Likelihood.

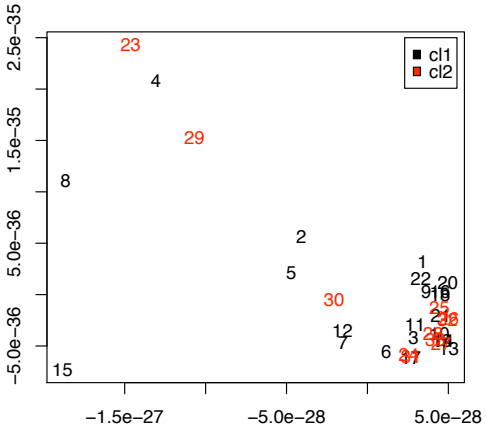


Figure 41: Sammon plot of clusters 1 and 2. Distribution used is Poisson and distance measure used is Chi-Square.

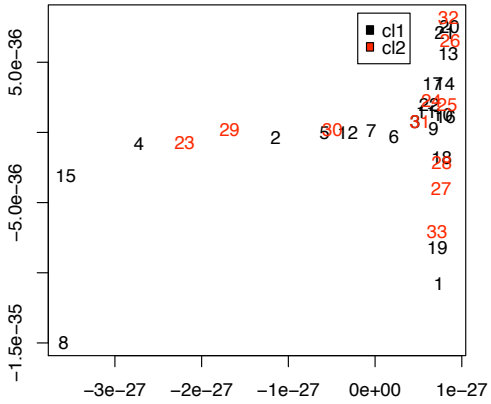


Figure 42: Sammon plot of clusters 1 and 2. Distribution used is Poisson and distance measure used is Trans Chi-Square.

Figure 39 illustrates the effectiveness of the algorithm on the clustering of the samples in these two clusters. It shows for each distribution using each distance measure the number of samples assigned to the correct cluster, which in this case is all of the samples.

Figure 39 illustrates the difference between each of the samples contained in clusters 1 and 2. Due to the fact that clusters 1 and 2 cluster perfectly for each available distribution and distance measure, only Sammon plots of the Poisson distribution using each of the similarity measures has been shown.

Looking at Figure 40 there is some evidence of clustering. However, the samples in cluster 2 seem to be widely spread. However, Figure 41 and Figure 42 show large overlap between the two clusters. These plots have been made using total likelihood/ chi square/ trans chi square of each sample, which in itself does not seem sensitive enough for the clustering. Clearly the method adopted in the algorithm of finding a cluster centre and calculating the required similarity for each sample is sensitive enough to distinguish between these clusters.

Clustering of samples contained in clusters 1 and 3 gave the results presented in Table 4. These results indicate that whilst there is some indication of a similarity between the samples contained in clusters 1 and 3, the algorithm is sensitive enough to detect these and clusters the majority of the samples correctly. The most successful implementation of the algorithm was modelling the data with the Zero-Truncated Poisson distribution using the similarity measure Trans Chi-Square, illustrated in Figure 45. The results from clustering of samples 1 and 3 indicate that the issue encountered when clustering the entire dataset arises because clusters 2 and 3 are very similar. This is investigated further below. These results are also shown in Figure 43, illustrating the slight overlap between the two clusters more clearly.

Table 4: Results from PoissonC / PoissonL clustering of samples contained in clusters 1 and 3. The samples were clustered using each distribution with each distance measure. Cluster 1 consists of 22 samples and cluster 3 contains 22 samples.

	Cluster1		Cluster3	
Distribution & Distance measure	#Correct samples in cluster	#Wrong samples clustered alongside	#Correct samples in cluster	#Wrong samples clustered alongside
Poisson				
Likelihood	19	2	20	3
Chi-Square	20	3	19	2
Trans-Chi	20	3	19	2
Negative Binomial				
Likelihood	10	0	22	12
Chi-Square	20	2	20	2
Trans-Chi	13	5	17	9
Zero-Truncated Poisson				
Likelihood	20	1	21	2
Chi-Square	12	3	19	10
Trans-Chi	21	0	22	1

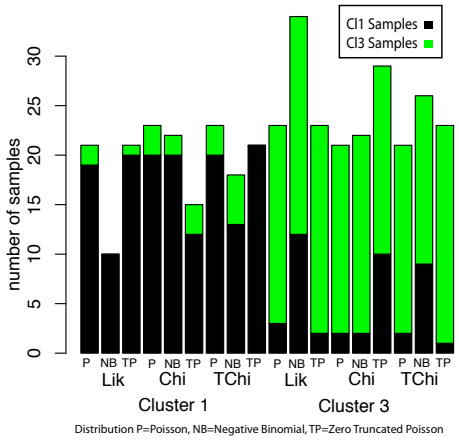


Figure 43: Bar plot of clustering results for clusters 1 and 3 using each distribution and each distance measure.

Table 5 shows the results obtained when using the algorithm to cluster the samples contained in clusters 2 and 3. The similarity between the samples in these two clusters is evident when looking at these results. Clustering of these samples modelling the data using the Negative Binomial distribution and using the likelihood as a similarity measure seems to work notably better on this section of the data than any other method.

Table 5: Results from PoissonC / PoissonL clustering of samples contained in clusters 2 and 3. The samples were clustered using each distribution with each distance measure. Cluster 2 consists of 11 samples and cluster 3 contains 22 samples.

	Cluster2		Cluster3	
Distribution & Distance measure	#Correct samples in cluster	#Wrong samples clustered alongside	#Correct samples in cluster	#Wrong samples clustered alongside
Poisson				
Likelihood	3	0	22	8
Chi-Square	4	0	22	7
Trans-Chi	4	0	22	7
Negative Binomial				
Likelihood	10	1	21	1
Chi-Square	4	0	22	7
Trans-Chi	4	0	22	7
Zero-Truncated Poisson				
Likelihood	3	0	22	8
Chi-Square	5	6	16	6
Trans-Chi	4	0	22	7

These results are illustrated in Figure 44, showing the results using each distribution and each distance measure of the samples assigned to each cluster. There is a very distinct overlap between the two clusters, which would indicate that they are very similar in nature. From the plot it can be seen that often more of the samples in cluster 2 are assigned with the samples in cluster 3 than in a separate cluster.

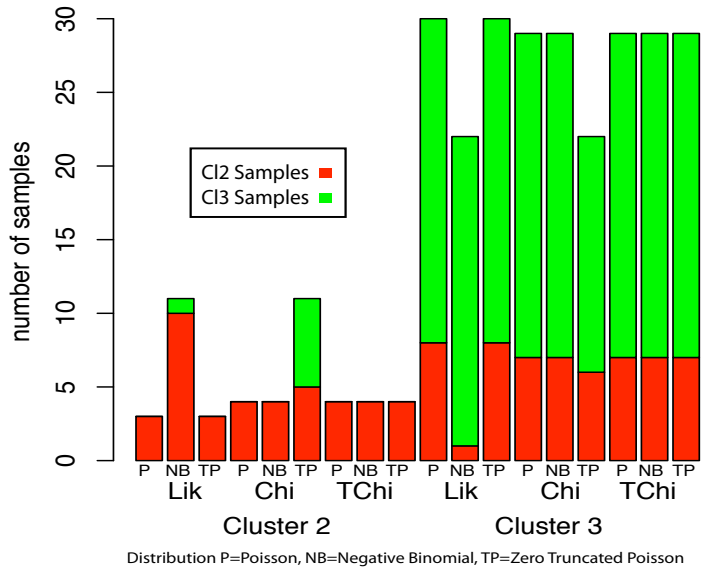


Figure 44: Bar plot of clustering results for clusters 2 and 3 using each distribution and each distance measure.

Figure 45 is a Sammon plot showing the approximate similarity of each of the samples in clusters 1 and 3, modelling the data using the Zero Truncated Poisson distribution and using the Trans Chi-Square as a measure of similarity. From this plot there is evidence of clustering but the two clusters overlap each other, which indicates that the samples from each cluster would not cluster distinctly. As above, using most of the distance measures for each distribution the algorithm seems to be more sensitive and identifies the majority of the samples in the correct cluster. Figure 46 is a Sammon plot showing the approximate similarity of each of the samples in clusters 2 and 3, modelling the data using the Negative Binomial distribution and using the Likelihood as a measure of similarity. There is a clear spread of data here, which would indicate that the samples contained in clusters 2 and 3 will not cluster. The majority of the results in Table 5 support this conclusion and therefore confirms the suggestion that clusters 2 and 3 are too similar in nature, which has a detrimental effect when clustering the entire dataset.

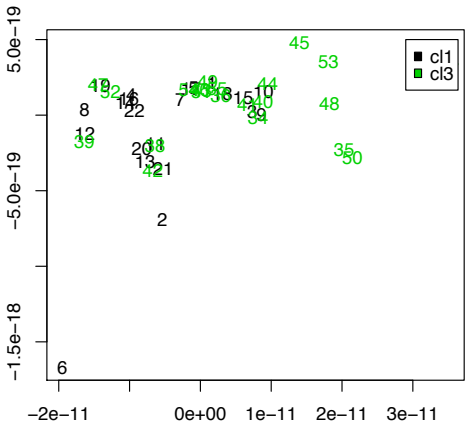


Figure 45: Sammon plot of clusters 1 and 3. Distribution used is Zero Truncated Poisson and distance measure used is Trans Chi-Square.

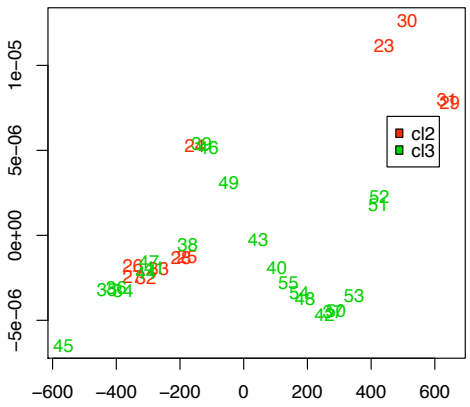


Figure 46: Sammon plot of clusters 2 and 3. Distribution used is Negative Binomial and distance measure used is Likelihood

When the interest is in clustering of tags, the issue then arises of how to display this information as there can be hundreds or possibly thousands of individual tags sequenced. What is of interest is if there are any specific group of tags that appear together when using each of the clustering methods.

Due to no a-priori information being given about the clustering of tags, the algorithm was simulated with various numbers of clusters as the input and from visual analysis of the results it was decided that three clusters were appropriate. After the results were recorded for each method in each distribution, a similarity matrix was constructed by counting the number of tags in common with all of the clusters output from each method. To show this in a more understandable manner a graphical display of this similarity matrix was plotted using the image() command in R, shown in Figure 47. This plot illustrates the elements of the matrix; where white illustrates elements that are exactly the same, elements with a high similarity are shown by a light colour such as yellow, areas with low similarity with darker colours like orange and red if there is no similarity.

Figure 47 shows the similarity for each cluster produced by each method of clustering. Looking at the plot is seems that there are several clusters that have tags in common but does not provide much useful information.

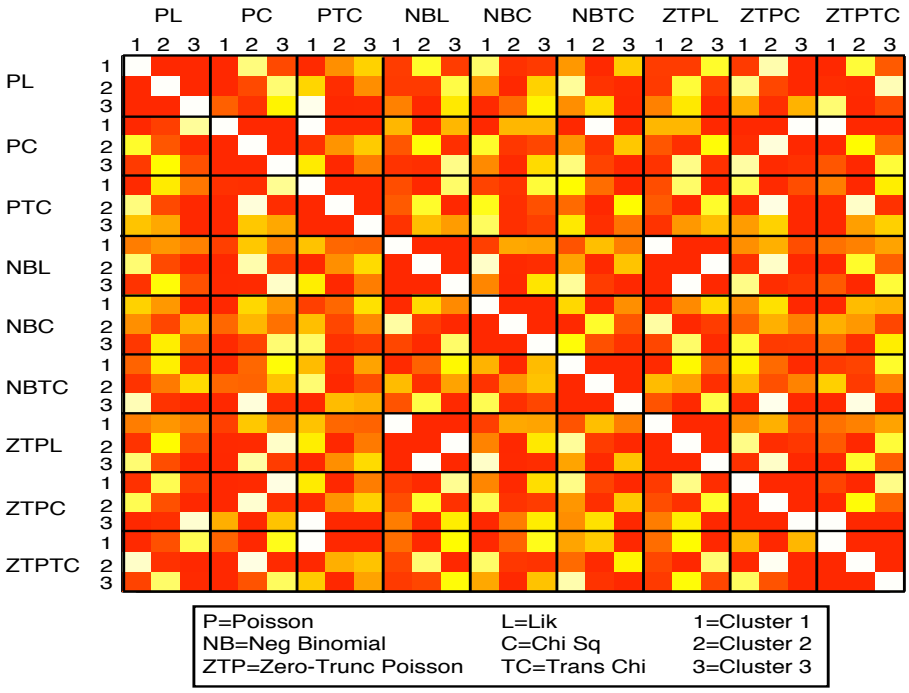


Figure 47: Graphical image of tag cluster similarities matrix.

In order to determine which tags cluster commonly using every method of clustering the similarity matrix constructed above was converted into a distance matrix and hierarchical clustering was performed in R to find out which of the clustering methods gave the most similar results. The dendrogram is given below in Figure 48. As expected, the three output clusters from each pair of inputs cluster together and it seems that the three Poisson methods of clustering produce very similar results as they cluster quite distinctly together, as do both the Negative Binomial and Zero-Truncated Poisson distributions using each of the similarity measures.

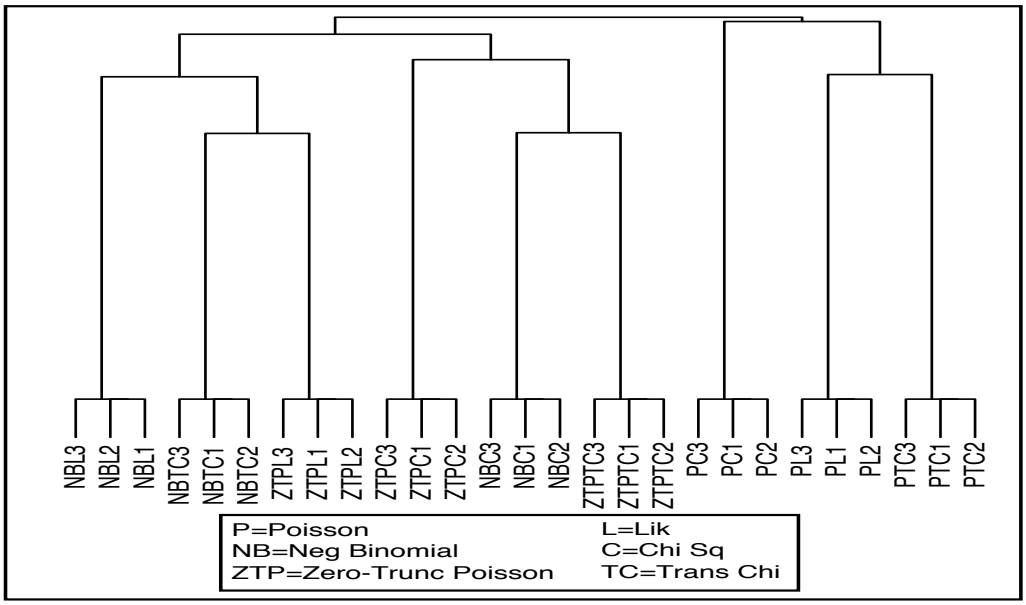


Figure 48: Dendrogram displaying the similarities of the results obtained from clustering of tags using each of the methods available in the algorithm.

From these results the output was then analysed to find out if any tags cluster together throughout using each of the available clustering options. It was found that only 40 tags commonly cluster. It is likely that these 40 tags will be tags that will not be differentially expressed between clusters of samples, due to the lack of information known about the grouping of the tags, no biological inferences can be made or assumptions confirmed. Due to the lack of analytical information obtained when clustering of the tags in dataset 1, the same analysis was not attempted for dataset 2, as no a-priori information was known about that dataset.

4.3.1.2 Dataset 2

Due to the lack of information given about the grouping of samples in dataset 2 the algorithm was run using each of the available distributions and similarity measures. These results were then evaluated to assess which samples most frequently appear in each cluster. As a result of the lower number of samples in this dataset (26 samples), it was

assumed that only two clusters exist. This was confirmed by repeatedly running the algorithm for both two and three clusters and when run for two the results were the same each time.

Table 6: Results from PoissonC / PoissonL clustering of dataset 2. Each method was used and from this the most optimal clusters were selected.

Distribution & Distance measure used	Results given for samples contained in Cluster 1	Results given for samples contained in Cluster 2
Poisson		
Likelihood	1 2 3 4 6 7 9 10 11 21 22 23 25 26	5 8 12 13 14 15 16 17 18 19 20 24
Chi-Square	1 2 3 4 6 7 10 11 22 25 26	5 8 9 12 13 14 15 16 17 18 19 20 21 23 24
Trans-Chi	1 2 3 4 6 7 8 9 10 11 21 22 23 25 26	5 12 13 14 15 16 17 18 19 20 24
Negative Binomial Distribution		
Likelihood	1 2 3 4 5 6 9 10 12 13 14 16 22 23 26	7 8 11 15 17 18 19 20 21 24 25
Chi-Square	1 3 4 8 9 10 11 21 22 23 25 26	2 5 6 7 12 13 14 15 16 17 18 19 20 24
Trans-Chi	1 2 3 4 6 7 10 11 22 25 26	5 8 9 12 13 14 15 16 17 18 19 20 21 23 24
Zero-Truncated Poisson Distribution		
Likelihood	1 2 3 5 6 7 9 21 22 23 24	4 8 10 11 12 13 14 15 16 17 18 19 20 25 26
Chi-Square	1 3 4 6 8 9 10 11 21 22 25 26	2 5 7 12 13 14 15 16 17 18 19 20 23 24
Trans-Chi	1 2 3 4 6 7 8 10 11 22 25 26	5 12 13 14 15 16 17 18 19 20 23 24

Table 7: The percentage of most common cluster in which each sample is contained, evaluated from Table 6

Sample	Cluster	% Occurrence	Sample	Cluster	% Occurrence
1	1	100%	14	2	89%
2	1	78%	15	2	100%
3	1	89%	16	2	89%
4	1	89%	17	2	100%
5	2	78%	18	2	100%
6	1	88%	19	2	100%
7	1	67%	20	2	100%
8	2	56%	21	1	67%
9	1	78%	22	1	100%
10	1	89%	23	1	56%
11	1	78%	24	2	89%
12	2	89%	25	1	78%
13	2	89%	26	1	89%

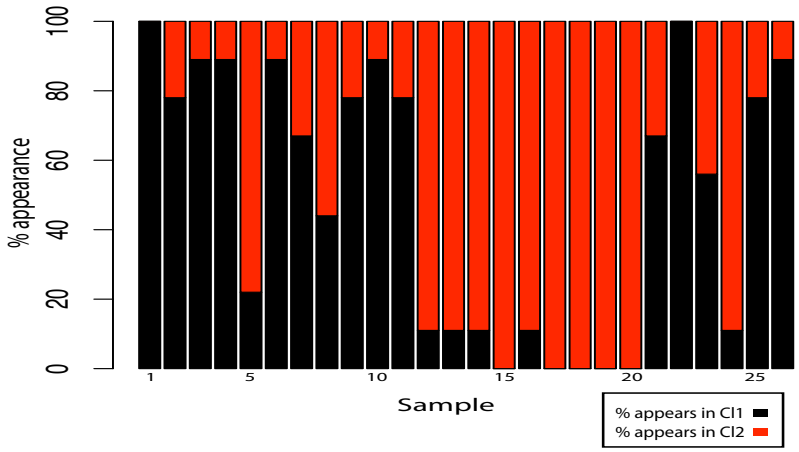


Figure 49: Bar Chart displaying the percent occurrence of each sample in each cluster.

Assessing the results presented in Table 7 and Figure 49 gives the most optimal clusters as:

1. 1,2,3,4,6,7,9,10,11,21,22,23,25,26
2. 5,8,12,13,14,15,16,17,18,19,20,24

where samples 7, 8 21 and 23 could be outliers. Looking at Table 6 the only method of clustering that has given these results exactly is using the Poisson distribution to model the data and using the likelihood as a similarity measure. Below is a Sammon plot of this, in which there is some evidence of the clustering that the algorithm suggests but the two clusters appear to be very similar.

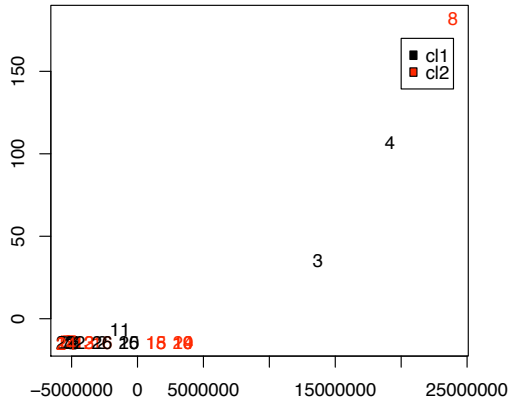


Figure 50: Sammon plot of optimal clusters in dataset 2.

Poisson and distance measure used is Likelihood.

4.3.2 Bayesian Algorithm

The Bayesian algorithm was used to construct a distance matrix for the samples in both dataset 1 and dataset 2. This distance matrix was then put into the hierarchical clustering function, *hclust*, in R. Dendrograms were then plotted to observe the clustering hierarchy. Figure 51 shows the results from the clustering of dataset1 and Figure 52 shows those from dataset 2. As is clear from the two figures, no hierarchy has been established, suggesting that the algorithm is not sensitive enough leading to the conclusion that the clustering algorithm will not be successful on any data of this format. This could be due to a variety of reasons such as the mathematics being interpreted wrongly when translating from paper to code or the data is not suitable for the algorithm. These will be discussed further in section 0.

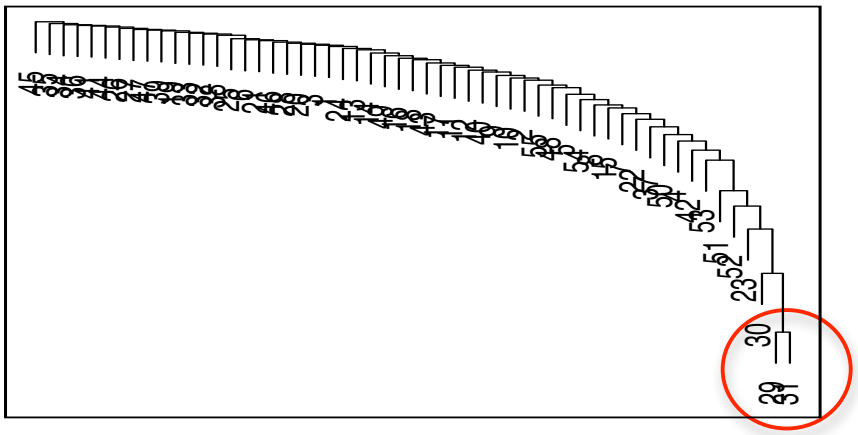


Figure 51: Dendrogram produced upon applying Bayesian algorithm to Dataset 1.

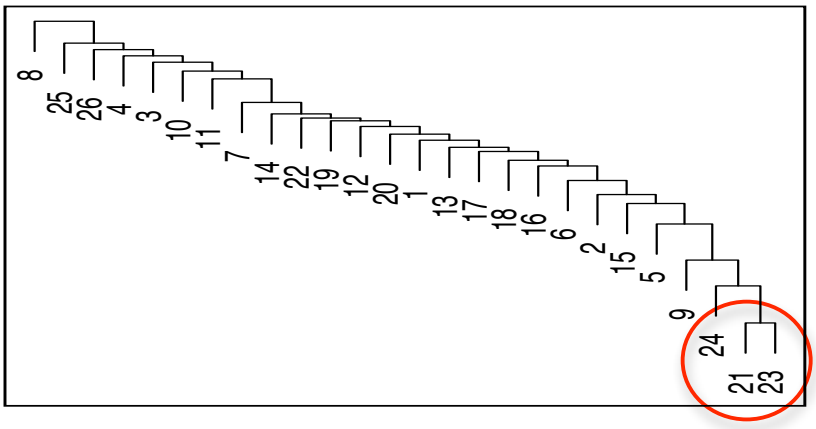


Figure 52: Dendrogram produced upon applying Bayesian algorithm to Dataset 2.

Looking at both of the dendrograms, what is interesting is that the outliers found in Chapter 3 cluster together first and then each of the other samples follow in no particular order. The Bayesian clustering analysis for dataset 1 shown in Figure 51 highlights the outlying samples 29, 30 and 31 in a red circle. For dataset 2 the outliers are samples 21, 23 and 24 again highlighted in Figure 52 in a red circle.

Summary

The results presented above imply that the Poisson C / Poisson L algorithm is sensitive enough to detect the dissimilarities and cluster samples distinctly in some cases.

However, when the clustering of all three clusters was attempted on dataset 1 it appeared that the overlap between the three clusters was too great and the algorithm failed to separate them distinctly. This could be due to a variety of factors: the three clusters may overlap and the differences between them may be too small for the algorithm to detect, the information given about the grouping of the samples may have been wrong or the algorithm may not be adequate for clustering of more than two groups. When the three clusters were analysed pair-wise it became clear that there was a large overlap, particularly between clusters 2 and 3, suggesting that the fault lies in the information given about the grouping of samples.

Due to the lack of information given about the grouping in dataset 2, the results presented cannot be confirmed or rejected. The algorithm was run in triplicate and obtained the same, recorded results each time.

The Bayesian algorithm yielded surprising results. It is assumed in the paper it was proposed in [24] that this method of constructing a distance (or similarity) matrix is adequate for all typed of small RNA cloning data. The problems encountered with this algorithm could lie in the translation of this method from the paper into R. The mathematics had to be translated into code and altered due to the large scale of the datasets used in this analysis. It is also possible that this algorithm is just not suitable for the analysis of next generation sequencing data. Compared to other sequencing datasets such as SAGE, the data provided by MSKCC and analysed in this thesis has an extensively larger scale and proportion of zero counts.

Chapter 5

Differential Expression

5.1 Overview

One of the most important questions in the analysis of any type of sequencing data is whether a given tag is differentially expressed. The goal of differential expression is to 'find statistically significant associations of biological conditions or phenotypes with gene expression.'[34] Differentially expressed miRNAs (or equally genes, proteins, exons etc.) are detected from variations in the expression profiles of the tag associated with that miRNA.

The interest in differential expression is in how the expression of different tags changes between individual samples or groups of samples. Ideally the goal of this, particularly in the analysis of cancer data, is to find groups of tags that are highly expressed in only the cancerous tissue samples and other groups of tags that are highly expressed only in the non-cancerous tissue samples. This can then lead to a long-term goal of discovering certain miRNAs or groups of miRNAs (or genes, exons, proteins) that occur more frequently in cancerous tissue, which in turn could lead to further development of treatments. Another use for this is to detect if there are certain genetic traits that can lead to early diagnosis of cancer (or any disease) in members of the same family.

It has previously been shown that simple significance tests such as the 2-sample t-test and Chi square test are often not sensitive enough to detect differential expression of tags between samples. This could be due to the influence of sample size or the random fluctuations that occur in the data [25][26]. Countless methods have been developed for the detection of differentially expressed tags. Some of these are used to detect differential expression of tags between two individual samples and some are used to detect differential expression of tags between groups of samples. The problem with using the methods developed to detect differential expression between individual samples is that, while adequate at detecting between and within library variation for the two individual samples, if the interest is in differential expression between groups of samples (or clusters) the samples are just pooled and the analysis run on the two pooled groups treating them as two individual samples. This pooling of the samples often results in the information about the within library variation and between individual library variation being lost.

In the analysis presented in this chapter various different methods developed for the detection of differential expression will be evaluated for use on next generation sequencing data. Firstly, simple significance tests such as the 2-sample t test and the Wilcoxon signed rank test will be used to illustrate the need for different techniques that can adapt more to the type of data being analysed. Next, various methods developed for the analysis of SAGE data were translated from research papers into R code and tested. These methods are a significance test developed by Audic and Claverie [25], a weighted t-test [26], a model for the data using over-dispersed logistic regression[17], an adaptation to [17] modelling the data using an over-dispersed log-linear approach[28], a log ratio

method[27] and modelling the data using a Poisson mixture model [29]. All of these methods are outlined in Chapter 2.

It is important to note that the grouping (or clusters) must be known beforehand, these can found using the methods presented in Chapter 4 and the results from these methods are used as input for the analysis presented in this chapter. However, more often than not this information should be known a-priori as ideally samples of certain types of tissue should cluster together (i.e. cancerous and non-cancerous). All of the methods assessed in this chapter can only be used for the analysis of two clusters of samples.

5.2 Adaptations

Most of the algorithms used to assess differential expression were translated from each of the research papers into code in R and used as presented in the given paper. Code was provided for the over-dispersed logistic regression[17], the over-dispersed log linear [28] and the Poisson mixture model [29] methods. Very few adaptations were made to these algorithms mainly due to the lack of time. However, some changes were made to the over-dispersed logistic regression and log ratio methods in an effort to make the methods more suitable for the data.

5.2.1 Over-dispersed logistic regression method

When the original method was tested on the data the code failed every time on both datasets. After contacting the author Keith A Baggerly it was suggested that this could be due to 'fake counts' where the count of an individual tag in one of the clusters is zero for all samples in the cluster. This causes the logistic regression to fail due to the proportions being so small. Two scenarios were recommended by the author, the first was to adjust

the data slightly by adding one count to each of the tags with an original count of zero and adding a count of two to the library sizes to account for this. The second was to add different weights into the logistic regression that take into account both library size and the level of over-dispersion. The results are presented and discussed further later in this chapter.

5.2.2 Log Ratio Method

The log ratio method presented in [27] does not take into account the specific groups or clusters in the data so it is possible to use this when no information is known or assumed about the grouping of the samples. However, in order to make a more sensitive measure this was adapted slightly to take the separate clusters into account. The outline of the algorithm used was the same as that presented in 2.3.3 but the alternative hypothesis was changed to include two alternative hypotheses of the form:

H_1 (alternative): The tag is differentially expressed, so the frequency of the gene is different in at least some of the samples in each cluster *(55)*:

$$L_i^{alt}(k) = \prod_{t=1}^{m} \frac{e^{-\lambda_i(t)\theta_t}\left(\lambda_i(t)\theta_t\right)^{y_i(t)}}{y_i(t)!} \qquad (55)$$

where k=1,2 represents the cluster and m is the number of samples in the given cluster. The two likelihoods are calculated for each tag, $L_i^{alt}(1)$ and $L_i^{alt}(2)$ for clusters 1 and 2 respectively and the log ratio statistic is calculated using *(56)*.

$$R = \log\left(\left[L_i^{alt}(1) - L_i^{alt}(2)\right] / L_i^{null}\right) \qquad (56)$$

From *(56)*, the decision can be made whether or not to reject the null hypothesis and hence, determine if a tag is or is not differentially expressed.

5.3 Results

Many previous methods developed for the identification of differential expression in the analysis of sequencing data consider the comparison of only two samples. Methods such as a normal approximation based on the z-test statistic (equivalent to the chi-squared test [34]) and significance of gene expression profiles [25] have been reviewed previously by Ruitjer et al [35]. It has been shown in this review that these methods work well in the case of studying two individual samples. This is illustrated below for the most and least correlated samples in dataset 1. It is evident from Table 8 that the two methods are mostly in agreement when detecting the differentially expressed tags.

Table 8: Results from differential expression analysis comparing only two samples at a time. This was done for both the most and the least correlated samples of dataset 1.

	# Diff expressed tags detected using the Z-statistic	# Diff expressed tags detected using the Significance of gene expression profiles	# Diff expressed tags appear in both
Most correlated	219	249	200
Least correlated	340	404	306

However, these methods do not adapt well to the analysis of two groups (or clusters) of samples. As discussed by Baggerly et al [26][17] and Lu et al [28], previous methods for the analysis of groups of samples have often relied on pooling the data in a specific group into one individual sample. It is suggested [17][26][28] that this is due to between sample

variability being lost within the group and, particularly in the case where the samples within a group are not replicate sequences from the same source, a certain proportion of within sample variability may also be lost. This is due to pooling of the data overemphasising the significance of the results as the normal variation between the results of different samples within a group is ignored. If they were adequate at detecting differential expression between groups, significance tests such as the 2-sample t-test and the normal approximation based z-test statistic mentioned above (which is equivalent to the chi-squared test) ought to be in agreement when detecting for differential expression.

Shown below is a graph displaying how the 2-sample t and the Chi-Square test statistic give contrasting results as to which tags are differentially expressed. These statistics were calculated for the tags that have a high count (>40) across all samples contained in clusters 1 and 2. It was found that 197 tags have a high count over all samples.

The t-statistic was calculated using $\frac{P1 - P2}{\sqrt{V1 + V2}}$, taken from the Baggerly paper [26], where P1 and P2 are the proportions of that particular tag in cluster 1 and 2 respectively and V1 and V2 are the sample variances of cluster 1 and 2 respectively. The chi-squared statistic was calculated for each of these tags using (57). Each element of *(57)* is explained in Table 9. This equation and table were given in the analysis presented by Man et al [35]. The two statistics were calculated for these high-count tags and plotted in Figure 53.

$$\chi^2 = \frac{N(y_{1,1}y_{2,2} - y_{1,2}y_{2,1})^2}{N_{1.}N_{2.}N_{.1}N_{.2}} \qquad (57)$$

Table 9: Table explaining each element of the equation to calculate the chi-square test statistic, where N is the total number of tags in the entire dataset.

	Cluster1	Cluster2	Total tags
Count of Tag i in dataset	$y_{1,1}$	$y_{1,2}$	$N_{1.}$
All other tags in dataset	$y_{2,1}$	$y_{2,2}$	$N_{2.}$
Total count of tag I in given cluster	$N_{.1}$	$N_{.2}$	N

If the two methods concur with each other a U-shape would be observed in Figure 53 where certain tags were found equally extreme by both statistics. However this is not the case. Looking at the graph, most of the tags being highlighted as significant by the chi-square statistic are not significant according to the t-statistic.

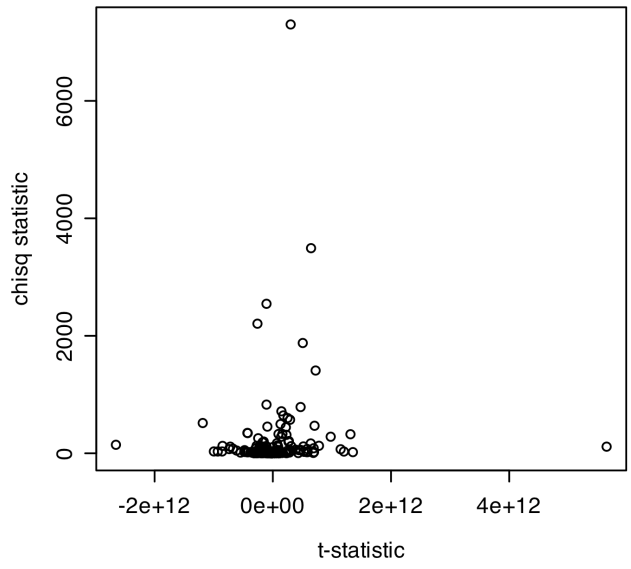


Figure 53: Plot of the chi-square statistic versus the t-statistic

While the 2-sample t-test does capture some of the between library variance it has an inherent problem when analysing this type of data as it assumes a normal distribution to the data and also applies equal weights to each of the samples. This would be somewhat

acceptable in the case where the samples are replicate libraries from the same source. However, the problem of largely differing sample sizes still exists and this is undesirable in the case of the data analysed in this thesis as the proportions of each tag vary greatly over the samples.

To try to account for the departures from the distributional assumptions of the 2-sample t test, the Wilcoxon signed rank test was also applied to the datasets. However this test also applies equal weights to the samples and does not take into account the within sample variability. Various methods have been developed to account for these issues, as explained in Chapter 2. Six of these methods were translated from the papers into code for the R statistical language [37] and used on the given datasets. The main issue that emerges when applying these to the given data is that there is no way to confirm the results as no information is known about the tags themselves or the nature of their grouping. This will be investigated in the simulation study presented in Chapter 6.

The clusters of samples were given a-priori in dataset 1, so the differential expression analysis could potentially be done without using the results from the clustering analysis in Chapter 4. Although there are three clusters present in this dataset all of the analysis techniques only work for two groups of clusters, so the differential expression analysis of clusters 1,2 and 3 was implemented on each pair of clusters separately. The number of differentially expressed tags was then recorded and the overlap of the differentially expressed tags detected using each testing method was found and recorded. Table 10 Table 11 and Table 12 contain the results of each pair of the differential expression analysis of the three pairs of clusters in dataset 1.

Table 10: Table of results from differential expression analysis of clusters 1 and 2 from dataset 1 using 9 different methods. The diagonal is the count of differentially expressed tags found using each method. Every other element represents the overlap of tags when using the two methods

D.E method	Simple t	Wilcox	Weighted t	Overdisp Log.reg	Overdisp Log.lin	Ratio paper	Ratio adapt	Pois mix
2-Sample t	108	97	27	0	102	44	30	71
Wilcox	97	179	59	0	133	58	92	96
Weighted t	27	59	110	0	84	42	83	52
Overdisp Log.reg	0	0	0	0	0	0	0	0
Overdisp Log.lin	102	133	84	0	226	85	88	122
Ratio paper	44	58	42	0	85	116	51	56
Ratio adapt	30	92	83	0	88	51	178	63
Pois mix	71	96	52	0	122	56	63	132

Table 11: Table of results from differential expression analysis of clusters 1 and 3 in dataset 1 using 9 different methods. The diagonal is the count of differentially expressed tags found using each method. Every other element represents the overlap of tags when using the two methods

D.E method	Simple t	Wilcox	Weighted t	Overdisp Log.reg	Overdisp Log.lin	Ratio paper	Ratio adapt	Pois mix
2-Sample t	106	14	54	0	84	26	61	49
Wilcox	14	95	15	0	35	25	23	29
Weighted t	54	15	128	0	99	36	100	83
Overdisp Log.reg	0	0	0	0	0	0	0	0
Overdisp Log.lin	84	35	99	0	185	45	88	134
Ratio paper	26	25	36	0	45	76	35	32
Ratio adapt	61	23	100	0	88	35	239	51
Pois mix	49	29	83	0	134	32	51	150

Table 12: Table of results from differential expression analysis of clusters 2 and 3 in dataset 1 using 9 different methods. The diagonal is the count of differentially expressed tags found using each method. Every other element represents the overlap of tags when using the two methods

D.E method	Simple t	Wilcox	Weighted t	Overdisp Log.reg	Overdisp Log.lin	Ratio paper	Ratio adapt	Pois mix
2-Sample t	43	16	32	0	39	8	37	24
Wilcox	16	64	8	0	44	15	19	25
Weighted t	32	8	152	0	116	25	151	98
Overdisp Log.reg	0	0	0	0	0	0	0	0
Overdisp Log.lin	39	44	116	0	228	42	162	118
Ratio paper	8	15	25	0	42	79	50	51
Ratio adapt	37	19	151	0	162	50	412	120
Pois mix	24	25	98	0	146	51	120	135

It is clear from all of the tables above that the over-dispersed logistic regression method even after applying the adaptations was completely unsuccessful. This could be due to a number of factors – particularly the large number of zero counts in the data and the high-count range of the data. Looking at Table 10, Table 11 and Table 12 although no information is known about the differential expression of the tags, it seems that the over-dispersed log-linear method is the most promising as it has the highest overlap with all of the other methods. It also appears that the adaptation of the log ratio method was successful, as it seems to have given much better results than the method presented in the paper. This is most likely due to the fact that it takes the grouping of the samples into account.

The differential expression analysis was then performed on dataset 2. The optimal clustering results obtained in Chapter 4 for this dataset were used, as no information was

previously known about the grouping of the samples or the tags. The differential expression analysis for dataset 2 is presented in Table 13. Again the over-dispersed log-linear method seems to be promising, but the analysis using the adapted log ratio method raises questions as it has detected a considerably larger number of differentially expressed tags compared to all of the other methods. This dataset contains over 3 times as many individual tags as dataset 1 so while the result that nearly half of the tags are differentially expressed is not impossible, it seems unlikely as no other methods have detected that large a number of differentially expressed tags.

Table 13: Table of results from differential expression analysis of the 2 clusters in dataset 2 using 9 different methods. The diagonal is the count of differentially expressed tags found using each method. Every other element represents the overlap of tags when using the two methods

D.E method	Simple t	Wilcox	Weighted t	Overdisp Log.reg	Overdisp Log.lin	Ratio paper	Ratio adapt	Pois mix
Simple t	84	75	43	0	69	6	38	53
Wilcox	75	120	68	0	87	9	62	63
Weighted t	43	68	111	0	61	5	85	52
Overdisp Log.reg	0	0	0	0	0	0	0	0
Overdisp Log.lin	69	87	61	0	126	10	57	92
Ratio paper	6	9	5	0	10	16	2	8
Ratio adapt	38	62	85	0	57	2	847	35
Pois mix	53	63	52	0	92	8	35	103

5.4 Summary

Due to the fact that no information is known about the grouping of the tags in either of the datasets, no formal assumptions or biological inferences can be made about the

differential expression analysis presented in this chapter. Many of the methods have previously been assessed on other types of sequencing data as mentioned above. However, in those cases information about the tags was known a-priori.

Looking at the results presented above it can be deduced that the over-dispersed log linear method for the analysis of differential expression, particularly when compared to simple tests such as the 2-sample t-test and the Wilcoxon signed rank test is the most reliable. This deduction is made based upon the results of the overlapping with other methods and the more reasonable number of differentially expressed tags detected, in contrast to those detected using the adapted log ratio method. However, none of this can be confirmed, as no information was known about the tags in either dataset.

Chapter 6

Simulating the data

6.1 Overview

Due to the fact that no information is known about the tags in dataset 1 and dataset 2 the validity of the differential analysis techniques cannot be assessed properly. Also the 'true counts' of the tags are not known in these datasets as the counts given are produced during the sequencing process. This makes it difficult to calculate the rate of false counts (or false positives) that are likely to appear in the data.

In order to account for this, data can be simulated from selected true counts with the desired conditions of differential expression set beforehand. The performance of the clustering algorithm for samples and the differential expression analysis for tags can then be analysed in detail and the rate of false positives (wrongly flagged differentially expressed tags) and false negatives (differentially expressed tags that have not been flagged) can be calculated.

In this chapter an algorithm is introduced to first simulate two vectors of true counts for the tags (miRNAs) for two conditions - differentially expressed and non-differentially expressed. The differential expression is set in designated tag numbers to make the change in expression significant. This is implemented so as when the differential

expression analysis is introduced the correct number of differentially expressed tags identified can be recorded. From these true counts the libraries (or samples) are then sampled from three different distributions: the Poisson, the Negative Binomial and the Zero-Truncated Poisson using pre-designated library sizes. Five of these libraries are simulated from the proportions of the non-differentially expressed true counts and five from the differentially expressed true counts.

Once this data has been simulated, the tests performed in Chapter 4 and Chapter 5 will be performed on the data and the results recorded in 6.3. The simulation algorithm is described in 6.2 below.

6.2 Algorithm

To simulate the data in the fashion that is suggested here a matrix of the form that is to be simulated is required. This is in order to calculate the power-law exponent γ introduced in Chapter 2. The algorithm to simulate the data works as follows:

1. Read in the dataset of the data to be simulated.
2. Remove all tags that have a count of zero across all samples.
3. Dispersion ϕ is calculated using the pseudo-likelihood method outlined in 2.1.3 for the matrix. To be used when the data is simulated from the Negative Binomial distribution.
4. The power-law exponent γ is calculated using the powerlaw() function provided by Khanin and Wit [30].

5. A designated number of tags are sampled; this is a list of tags which are to be differentially expressed. In the case of this study, 200 tags were chosen. Create a vector of these tag numbers (call this de.tags).
6. Now fold changes between 2 and 5 are sampled for each of these differentially expressed tags. So what is obtained here is a vector (call this fc.values) the same length as the vector created above.

   The fold change of a gene or miRNA is the ratio of the gene expression in one sample (or groups of samples) over another. Positive numbers indicate increases in expression, whereas negative numbers indicate decreases in expression.
7. Now the 'true' counts are simulated using the power-law exponent γ into the rpowerlaw() function provided by Khanin and Wit. This returns a vector (call this cell 1) of length the required number of tags, with a count for each tag representing the 'true' count.
8. Cell 1 is duplicated and this vector is named cell 2, now all of the tags in cell 2 that are contained in the vector de.tags are altered to have a fold change of the values in fc.values with the same increment. These tags that have had the fold change altered are now differentially expressed.
9. Change cell 1 and cell 2 into proportions.
10. The library sizes are simulated from a uniform distribution, using the maximum and minimum library sizes of the original dataset. In this case there will be 10 libraries.
11. The data is simulated from one of three distributions: the Poisson, Negative Binomial and Zero Truncated Poisson. This is done using the proportions cell 1 and cell 2 and the sampled library sizes. In this case five libraries were sampled using

the proportions in cell 1 and 5 using the proportions in cell 2. The five libraries that have been sampled from the proportions in cell 1 are in a separate cluster from the 5 libraries that have been sampled from the proportions in cell 2.

12. The tags that have a count of zero across all samples must be removed. The list of differentially expressed tags (de.tags) also has to be altered to account for this.
13. Clustering algorithms and differential expression analysis can now be tested on this data to look for false positives and assess the viability of these algorithms.

6.3 Results

6.3.1 Clustering

Data was simulated using the algorithm described above from each of the three distributions: Poisson, Negative Binomial and Zero-Truncated Poisson. Due to the way the data was simulated, it proved difficult to test if the data followed a specific distribution. This was due to the number of different distributions used in the simulation process. The proportions were simulated from the Power-law distribution, the library sizes from the Uniform distribution and the count of each tag in each sample from one of the designated simulation distributions.

A repeated Wilcoxon signed-rank test was set up to assess the probability that each of the tags in the particular dataset arose from the simulated distribution. This worked by performing multiple Wilcoxon tests on the vector of counts of the particular tag across all samples with 10000 numbers from the given distribution, with a mean equal to the mean count of the tag. The mean, maximum, minimum and mean standard deviation of the p-values were recorded and this was repeated for each individual tag in the simulated dataset. Once this had been done for each tag the mean of each of these values across all

tags was recorded and if the mean, maximum and minimum p-values are greater than the significance level of 0.05 and the standard deviation was not large then the tags could be said to come from the desired distribution. The results are displayed below in Table 14 and suggest that the tags in each of the simulated datasets follow the desired distributions. Ideally the range of the p-values should be relatively close together, as it is expected that each of the individual tags follow the mean of all counts of that specific tag.

Table 14: Repeated Wilcox test for assessing Simulated data, shows the mean, max, min and average standard deviation of the p-values for each simulated dataset.

	Mean	Max	Min	Std Dev
Poisson	0.63	0.64	0.60	0.01
Neg Bin	0.5889	0.6192	0.559	0.019
Z Trunc Pois	0.39	0.40	0.385	0.00479

Due to the performance of the Bayesian clustering algorithm on datasets 1 and 2, only the Poisson C / Poisson L algorithm was applied to each of the simulated datasets. The results are presented below. Each of the simulated datasets was modelled using each available distribution and distance measure in the algorithm and recorded below.

Assessing each of the algorithms on this dataset, samples 1 to 5 are expected to cluster together as are samples 6 to 10, as these groups of samples have been simulated from two separate vectors of true counts. Looking at Table 15 and Table 17 it is clear that the Poisson and Zero-Truncated Poisson data has clustered perfectly as would be expected. However Table 16 suggests that the Negative Binomial distribution is less trustworthy as the clustering results vary notably from the expected results. The row of zeros observed in Table 16 is due to the algorithm failing.

Table 15: Clustering results for Poisson simulated data. It is expected that five samples, samples 1:5, will be contained in cluster 1 and five samples, samples 6:10, in cluster 2.

	Cluster1		Cluster2	
Distribution & Distance measure	#Correct samples in cluster	#Wrong samples clustered alongside	#Correct samples in cluster	#Wrong samples clustered alongside
Poisson				
Likelihood	5	0	5	0
Chi-Square	5	0	5	0
Trans-Chi	5	0	5	0
Negative Binomial				
Likelihood	5	0	5	0
Chi-Square	5	0	5	0
Trans-Chi	5	0	5	0
Zero-Truncated Poisson				
Likelihood	5	0	5	0
Chi-Square	5	0	5	0
Trans-Chi	5	0	5	0

Table 16: Clustering results for Negative Binomial simulated data. It is expected that five samples, samples 1:5, will be contained in cluster 1 and five samples, samples 6:10, in cluster 2.

	Cluster1		Cluster2	
Distribution & Distance measure	#Correct samples in cluster	#Wrong samples clustered alongside	#Correct samples in cluster	#Wrong samples clustered alongside
Poisson				
Likelihood	3	2	3	2
Chi-Square	3	2	3	2
Trans-Chi	3	2	3	2
Negative Binomial				
Likelihood	4	1	4	1
Chi-Square	4	1	4	1
Trans-Chi	3	3	2	2
Zero-Truncated Poisson				
Likelihood	0	0	0	0
Chi-Square	3	2	3	2
Trans-Chi	3	2	3	2

Table 17: Clustering results for Zero-Truncated Poisson simulated data. It is expected that five samples, samples 1:5, will be contained in cluster 1 and five samples, samples 6:10, in cluster 2.

	Cluster1		Cluster2	
Distribution & Distance measure	#Correct samples in cluster	#Wrong samples clustered alongside	#Correct samples in cluster	#Wrong samples clustered alongside
Poisson				
Likelihood	5	0	5	0
Chi-Square	5	0	5	0
Trans-Chi	5	0	5	0
Negative Binomial				
Likelihood	5	0	5	0
Chi-Square	5	0	5	0
Trans-Chi	5	0	5	0
Zero-Truncated Poisson				
Likelihood	5	0	5	0
Chi-Square	5	0	5	0
Trans-Chi	5	0	5	0

These results are illustrated graphically below in Figure 54 and Figure 55. Due to the fact that both the Poisson and Zero-Truncated Poisson simulated data have the same results these are both represented by Figure 54 and the Negative Binomial data simulation's deviation from the true clustering results is illustrated in Figure 55. This would suggest that looking into the Negative Binomial distribution simulation for means of clustering would be unwise. Figure 56 and Figure 57 show Sammon plots for both the Poisson and Zero-Truncated Poisson simulated datasets using likelihood as a distance measure and Poisson and Zero-Truncated Poisson distributions to model each of the datasets respectively. Looking at the Sammon maps, Figure 56 shows some evidence of clustering, however Figure 57 suggests that there is no clustering whatsoever, similar to the Sammon mapping obtained in Chapter 4. It is possible the algorithm is more sensitive than the approximation that the Sammon mapping harnesses.

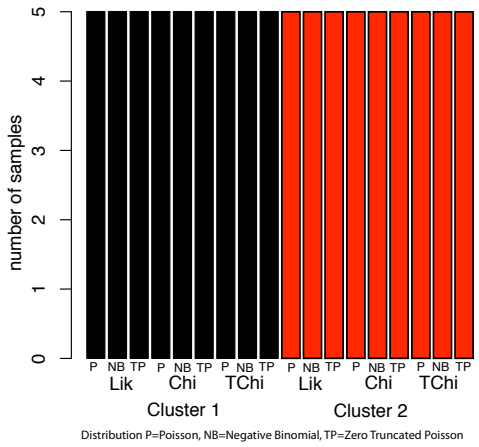


Figure 54: Bar-plot of clustering results for the two clusters for both Poisson and Zero-Truncated Poisson simulated data. Clustering analysis was performed using each distribution and each distance measure.

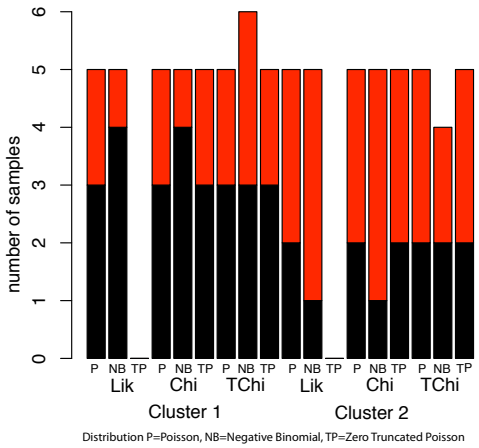


Figure 55: Bar-plot of clustering results for the two clusters for Negative Binomial simulated data. Clustering analysis was performed using each distribution and each distance measure.

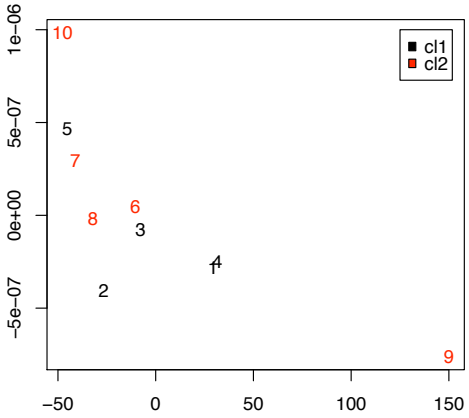


Figure 56: Sammon plot of Poisson simulated data. Distribution used is Poisson and distance measure used is Likelihood.

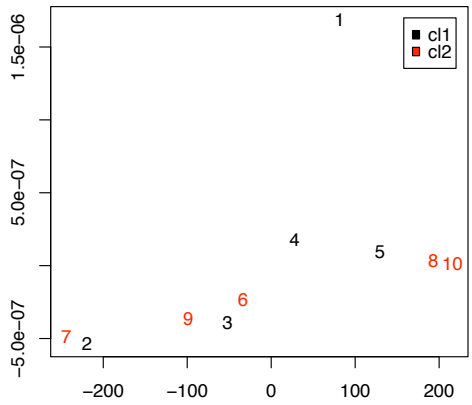


Figure 57: Sammon plot of Zero-Truncated Poisson simulated data. Distribution used is Zero-Truncated Poisson and distance measure used is Likelihood.

6.3.2 Differential expression

The next step of the simulation study was to assess the various methods of differential expression and to calculate the number of false positives and false negatives produced using each method. A false positive occurs when a tag has been flagged as differentially expressed but is not one of the tags in which differential expression should occur, a false

negative is a differentially expressed tag that has not been flagged. In the true counts, 200 tags were modified to exhibit differential expression. However, when the samples were simulated from the true counts, some of these designated tags had a zero count across all samples. In the Poisson simulated data 44/200 differentially expressed tags had a count of zero across all samples so the number of differentially expressed tags in the dataset was reduced to 156. In the Negative Binomial simulated dataset, 29/200 of these tags were zero across all samples so the number of differentially expressed tags in the dataset was reduced to 171. In the Zero-Truncated Poisson simulated dataset this problem did not occur so all of the original 200 tags were differentially expressed.

Displayed below in Table 18, Table 19 and Table 20 are the results of each of the methods of differential expression on each of the simulated datasets. The only method that was not assessed was the over-dispersed logistic regression method suggested by Baggerly et al [17], as it did not work for the given data or simulated data. In each table below, for the given simulated dataset, the number of correctly flagged differentially expressed tags using each differential expression analysis was recorded along with the false positives, false negatives and the overlap with each of the other methods. The overlap recorded is not that of the correct differentially expressed tags, but the overlap of all the tags flagged as differentially expressed by each method. This is due to the fact that if there is a large overlap between methods in tags that are not truly differentially expressed but are flagged as such a large proportion of the time. The method of simulation itself may be the problem. In the cell of the table where the overlap of a method is given with itself, what is recorded here is the number of tags flagged as differentially expressed using this method.

Table 18: Results of differential expression analysis testing 7 different methods on the Poisson simulated data.

Poisson Simulation 156 d.e tags	Simple t	Wilcox	Weighted t	Overdisp log lin	Ratio Paper	Ratio Adapt	Poiss mix
Flagged correctly	2	11	0	87	27	90	64
False +	1	1	87	139	24	260	98
False -	154	145	156	69	129	66	92
Overlap simple t	3	2	0	2	0	1	0
Overlap Wilcox	2	12	0	10	4	8	6
Overlap Weighted t	0	0	87	75	12	84	48
Overlap Overdisp log lin	2	10	75	226	51	200	63
Overlap Ratio Paper	0	4	12	51	51	51	31
Overlap Ratio adapt	1	8	84	200	51	350	82
Overlap Poiss mixture	0	6	48	63	31	82	162

Table 19: Results of differential expression analysis testing 7 different methods on the Negative Binomial simulated data.

Negative Binomial Simulation 171 d.e tags	Simple t	Wilcox	Weighted t	Overdisp log lin	Ratio Paper	Ratio Adapt	Poiss mix
Flagged correctly	0	2	6	43	86	77	32
False +	0	10	9	58	153	200	45
False -	0	169	165	128	85	94	139
Overlap simple t	0	0	0	0	0	0	0
Overlap Wilcox	0	12	5	7	10	9	4
Overlap Weighted t	0	5	15	6	7	15	3
Overlap Overdisp log lin	0	7	6	101	89	32	29
Overlap Ratio Paper	0	10	7	89	239	123	61
Overlap Ratio adapt	0	9	15	32	123	277	24
Overlap Poiss mixture	0	4	3	29	61	24	77

Table 20: Results of differential expression analysis testing 7 different methods on the Zero-Truncated Poisson simulated data.

Zero-Trunc Pois Simulation 200 d.e tags	Simple t	Wilcox	Weighted t	Overdisp log lin	Ratio Paper	Ratio Adapt	Poiss mix
Flagged correctly	0	36	0	107	35	151	78
False +	0	132	107	132	24	320	102
False -	0	164	200	93	165	49	122
Overlap simple t	0	0	0	0	0	0	0
Overlap Wilcox	0	168	90	151	38	162	79
Overlap Weighted t	0	90	107	94	19	107	86
Overlap Overdisp log lin	0	151	94	239	19	107	86
Overlap Ratio Paper	0	38	19	59	59	59	45
Overlap Ratio adapt	0	162	107	220	59	471	137
Overlap Poiss mixture	0	79	86	103	45	137	180

Looking at the tables above it is clear that the differential expression techniques have failed to distinguish between tags that are differentially expressed and tags that are not. Looking at Table 19 it is clear that the simulation of the data using the Negative Binomial distribution detects fewer differentially expressed tags. Adding to this the results of the clustering, it is safe to assume that the Negative Binomial distribution is unreliable for this type of data simulation. Looking now at Table 18 and Table 20 it is clear, as expected, that the 2-sample t test and the Wilcoxon signed rank test are not sensitive enough to detect differential expression in this data-type. In differential expression analysis of both the

Poisson simulated dataset and the Negative Binomial simulated dataset the weighted t-test has failed to flag any differentially expressed tags correctly. As a result only the over-dispersed log linear [28], the log ratio method proposed by Stekel et al [27], the adapted log ratio method described in 5.2.2 and the Poisson mixture model method [29] will be looked at into any further detail on both the Poisson and Zero-Truncated Poisson simulated datasets. The results are represented graphically below applying each of the methods mentioned above to both the Poisson simulated dataset and the Zero-Truncated Poisson simulated dataset.

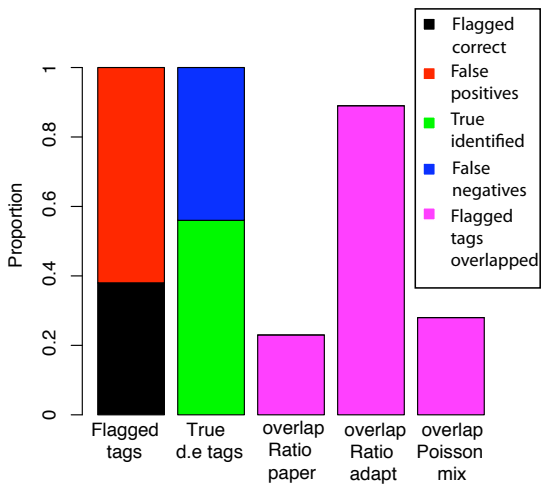


Figure 58: Bar plot outlining the results for over-dispersed log-linear differential expression analysis. What is shown is the proportion of false positives, false negatives and overlapping of the flagged tags in all the methods in relation to the true counts. This is for the Poisson simulated dataset

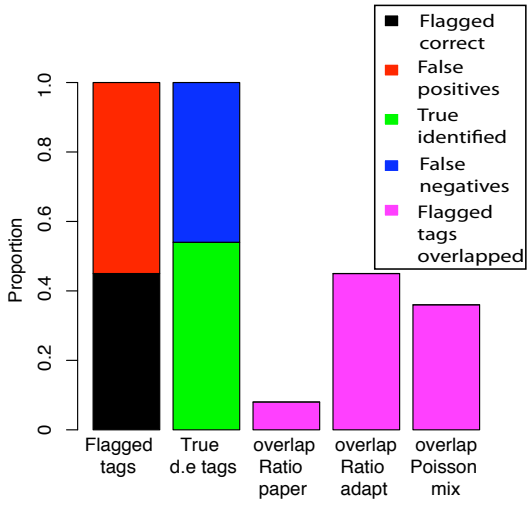


Figure 59: Bar plot outlining the results for over-dispersed log-linear differential expression analysis. What is shown is the proportion of false positives, false negatives and overlapping of the flagged tags in all the methods in relation to the true counts. This is for the Zero-Truncated Poisson simulated dataset.

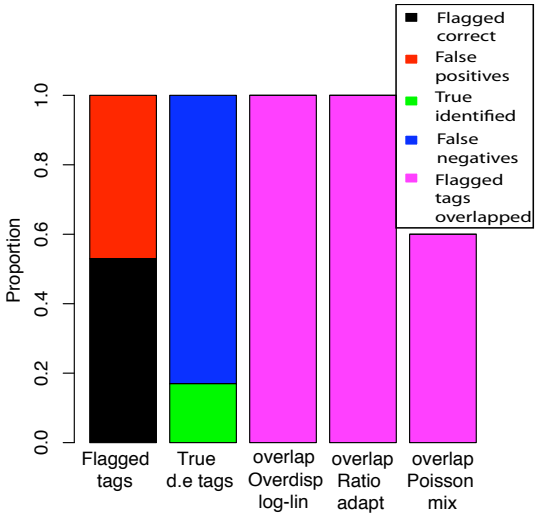


Figure 60: Bar plot outlining the results for log ratio differential expression analysis. What is shown is the proportion of false positives, false negatives and overlapping of the flagged tags in all the methods in relation to the true counts. This is for the Poisson simulated dataset

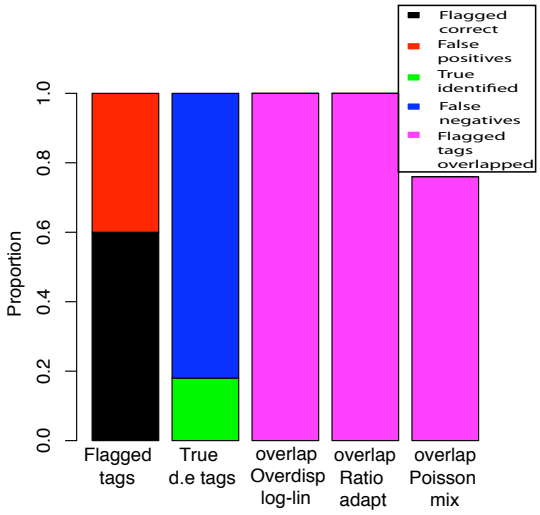


Figure 61: Bar plot outlining the results for log ratio differential expression analysis. What is shown is the proportion of false positives, false negatives and overlapping of the flagged tags in all the methods in relation to the true counts. This is for the Zero-Truncated Poisson simulated dataset.

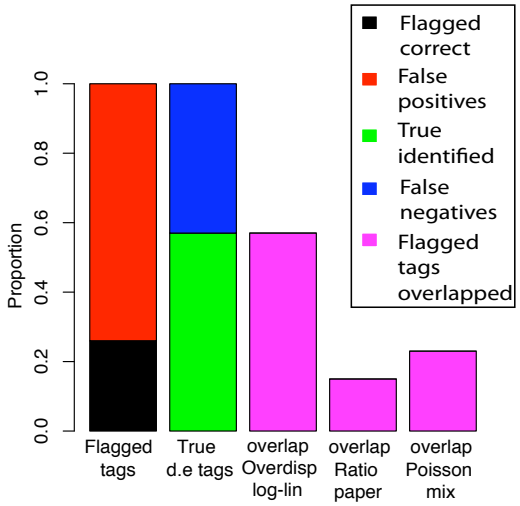


Figure 62: Bar plot outlining the results for adapted log ratio differential expression analysis. What is shown is the proportion of false positives, false negatives and overlapping of the flagged tags in all the methods in relation to the true counts. This is for the Poisson simulated dataset

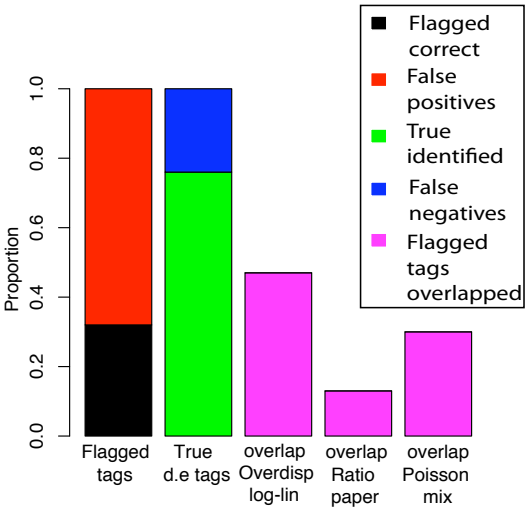


Figure 63: Bar plot outlining the results for adapted log ratio differential expression analysis. What is shown is the proportion of false positives, false negatives and overlapping of the flagged tags in all the methods in relation to the true counts. This is for the Zero-Truncated Poisson simulated dataset.

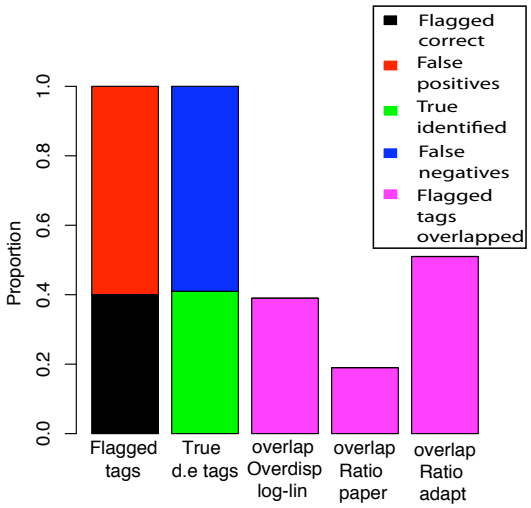


Figure 64: Bar plot outlining the results for Poisson mixture differential expression analysis. What is shown is the proportion of false positives, false negatives and overlapping of the flagged tags in all the methods in relation to the true counts. This is for the Poisson simulated dataset

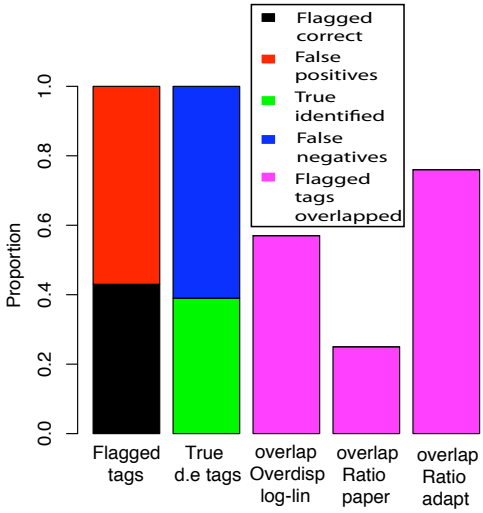


Figure 65: Bar plot outlining the results for Poisson mixture differential expression analysis. What is shown is the proportion of false positives, false negatives and overlapping of the flagged tags in all the methods in relation to the true counts. This is for the Zero-Truncated Poisson simulated dataset.

Looking at the plots above, it is clear that the error lies in the setup of the simulation study. When looking at the tags that have been flagged as differentially expressed, the proportion of false positives is greater than the proportion of correctly flagged for all of the methods on both datasets except when looking at the log ratio method (Figure 60 and Figure 61). However, looking at Table 18 and Table 20 the log ratio method flags a very low number of tags as differentially expressed in comparison to the other methods so this could be the reason for these results. The method itself does not take into account the grouping of the samples and these results would indicate that this method is an unreliable means for assessing differential expression as it flags a low number of tags, which suggests that the method is not adequately sensitive.

In all of the methods excluding the adapted log ratio method the proportion of false negatives is greater than the proportion of true tags identified. Looking at Table 18 and

Table 20 the adapted log ratio method has flagged a very large number of tags as differentially expressed, which explains the resulting difference. The magnitude of the differentially expressed tags flagged by the adapted log ratio method in both Table 18 and Table 20 coupled with similar results obtained in Chapter 5 leads to the inference that this method is an unreliable means for the analysis of differential expression.

After discounting the two log ratio methods, the interest lies in the over-dispersed log linear and Poisson mixture model methods. Looking at Figure 58, Figure 59, Figure 64 and Figure 65 the results show, for both datasets, that the proportion of false positives is greater than that of the correctly flagged tags. This could be due to either the method of analysis or the method of simulation. Looking at the results for the over-dispersed log linear method, for both datasets (Figure 58 and Figure 59), the proportion of correctly identified differentially expressed tags is greater than the proportion of false negatives. This would suggest that this method is the most reliable for differential expression analysis when compared to the other methods assessed.

6.4 Summary

Evaluation of the results above indicates that further work is needed on the method of simulating the data to increase its reliability, particularly for the assessment of differential expression methods. This will be discussed further in Chapter 7.

The assessment of the Poisson C / Poisson L algorithm on the simulated datasets yielded promising results, apart from when it was used on the Negative Binomial dataset. This suggests that the method of simulation is acceptable for the assessment of clustering algorithms developed for use on sequencing data.

From the results presented above coupled with those presented in Chapter 5, it is possible to conclude that the over-dispersed log linear method for the analysis of differential expression is the most reliable. However, due to the further investigation needed into the method of simulation this cannot be confirmed.

Chapter 7

Discussion and Conclusions

The use of clustering on sequencing datasets is one of the most common ways to identify similarities for the grouping of both samples and tags. While not always entirely accurate, it is a method of unsupervised learning so it is ideal when no information is known about the specific dataset. Two approaches to clustering - one based on k-means, incorporating different models to fit the data and various distance measures to assess similarity, and the other Bayesian hierarchical - have been presented and assessed in this thesis.

In dataset 1, due to the grouping of the samples being known a-priori, both of the clustering algorithms were applied to the samples to assess the reliability of the two algorithms. Looking at the results presented in Chapter 4 it is clear that the Poisson C / Poisson L algorithm was successful when clustering the first two groups of samples contained in this dataset. This would suggest that these particular groups are distinctly different and, as identical results were obtained when running the algorithm repeatedly for the same conditions with different random starting clusters, suggests consistency and reliability when using the algorithm.

When this algorithm was applied to all of the samples in the dataset, as well as samples in groups 1 and 3 and samples in groups 2 and 3 separately, the results diverged considerably from the expected results. The algorithm was run repeatedly under each condition and equivalent results were obtained. This, together with the results from the exploratory analysis using Sammon plots, would suggest that due to the successful

clustering of the samples in groups 1 and 2, group 3 appears to overlap the first two groups and the algorithm is not adequately sensitive to detect this. This could be due to a variety of reasons. For example, the samples in cluster 3 may belong to groups similar to those in clusters 1 and 2 but were assigned to a separate group. However, in order to investigate this further, more information would be needed about the dataset.

Applying the Bayesian algorithm to this dataset yielded unusual results as no hierarchy was observed, suggesting no clustering is present in the samples. These results whilst clearly wrong are interesting as the outliers observed in Chapter 3 cluster first and each of the other samples follow after in no particular order. The unusual results obtained could be due to an error in the translation of the mathematics from the paper into R. Another possibility is that the algorithm, whilst suitable for certain types of small RNA cloning data [24] is not suitably sensitive to detect differences between samples of the deep sequencing data provided.

It is natural that in most cases the grouping of the samples is known a-priori. For example, cancerous and non cancerous tissue samples, samples taken from patient A and samples taken from patient B. However there are cases when this information is not given, or where the interest lies in if specific samples do or do not cluster together.

This issue is raised when applying the clustering algorithms to the samples of dataset 2, as no information was given about the grouping of the samples or tags a-priori. Cluster analysis using the Poisson C / Poisson L algorithm was carried out using each available condition and the results were compiled to find the most likely clusters (Chapter 4). The algorithm was run repeatedly under different starting conditions for each set of conditions and the same results were obtained each time for the given conditions. Whilst

each condition yields different results, the percentage occurrence of an individual sample appearing in the same cluster using different starting conditions is relatively high (Figure 49) and the robustness of the algorithm is illustrated by the fact that the same results are obtained repeatedly. However, due to the lack of information known about the grouping of the samples the results obtained cannot be confirmed. The Bayesian algorithm was applied to this dataset and the same results were obtained as those for dataset 1 suggesting that the algorithm coded has been interpreted wrongly or is not suitable for the dataset.

Clustering of the tags was attempted for dataset 1 (Chapter 4) and results recorded. However, due to the lack of information given about the grouping of the tags in both datasets this could not be confirmed and analysis of the tags in dataset 2 was not deemed worthwhile. Due to the lack of information given about the tags it was not possible to see if certain tags that appear together in dataset 1 appear together in dataset 2.

Ideally the clustering algorithms would be tested for both tags and samples of a dataset with known (and distinct) groupings but due to the unpublished nature of this dataset, this was not possible. What could have been done for the grouping of samples instead of clustering was a classification analysis on dataset 1. This would have used the given grouping of the samples of dataset 1 and examined their interrelationship setting a class for each group (or cluster) that can then be used on datasets with no information of the grouping to find the similar classes of samples. A problem with this however is that it is specific to a certain type of data and the classes found for one will not be the same as that for a different type of data. For example, cancer tissue sequencing data and AIDS tissue sequencing data. Another option for further analysis would be to use the GAP [36]

statistic to estimate the number of clusters in the dataset. This could potentially be used on both samples and tags.

Although some promising results were obtained for the Poisson C / Poisson L algorithm further investigation is needed to confirm whether it is suitable for the clustering of deep sequencing data in general, especially where the situation of three or more groups of interest occurs. More distance measures and distributions could also be considered to suit different data types. Classification using cross validation on either test and training sets, published datasets or leave one out cross validation on the given dataset could have been carried out.

In the case of the Bayesian algorithm further investigation is needed to determine why it was not successful. The issue could be the large scale of the dataset and having to account for this in the gamma functions adopted in the analysis. Another possibility is that this algorithm is not sensitive enough for deep sequencing data. This could perhaps be accounted for by using a different Dirichlet prior in the analysis.

Due to the limitations of the datasets given, clustering of the tags provided too much information to handle easily without prior knowledge of the grouping of the tags. Further work could be done on different methods for the clustering of tags, however more information would be needed a-priori to assess the methods. Differential expression analysis is a more informative way of finding out key tags that are significantly up or down regulated across two groups (or clusters) of samples.

Once the grouping (or clusters) of the samples was calculated using the Poisson C / Poisson L algorithm, various methods of differential expression analysis were performed

on the data and the results recorded in Chapter 5. Due to the lack of information known about the grouping of the tags in both datasets, inferences made from the results obtained are subjective. It appears, upon examination of the results presented in Chapter 5, that routine tests of significance such as the two-sample t test and the Wilcoxon signed rank test are not adequately sensitive to detect differential expression in the dataset.

Significance tests previously developed for the analysis of differential expression in other types of sequencing data such as SAGE were researched and applied to the two datasets. The over-dispersed logistic regression method taken from Baggerly et al [17] failed on every analysis. After discussing this with one of the authors, Keith A Baggerly, adaptations were made and tested but to no avail. After evaluating the results given in Table 10, Table 11, Table 12 and Table 13 it appears that the over-dispersed log linear method for assessing differential expression is the most reliable. The adapted log ratio method while detecting a large number of differentially expressed tags would appear to be overly sensitive as in some cases it declares over 70% of the tags as differentially expressed which is rather implausible biologically. These results cannot be confirmed due to the lack of information known about the dataset.

In order to do any further work on differential expression methods more information needs to be known about the data being analysed. If more information was known the reliability of each of these methods could be assessed by calculating the correct number of differentially expressed tags flagged and the number of false positives. Once these methods were evaluated, the need for other methods or adaptations could be evaluated.

In an effort to do this a simulation study was proposed in Chapter 6, which aimed to provide a stable framework for evaluation of both clustering and differential expression

techniques with the tag and sample information pre-designated. While this simulated data was sensitive enough to assess the Poisson C / Poisson L algorithm, it is clear that further work is needed to make this simulation method more suitable for analysis of differential expression.

It appears that the method of incorporating the differential expression needs further work - perhaps the method of simulating the proportions using the Power-law distribution is not adequately sensitive. Another possibility is that the methods for assessing differential expression may not be sufficiently sensitive for such a large number of individual tags.

The work presented here could be further extended with additional investigation into methods of detecting differential expression, error rates and false positives in the datasets. With fewer limitations to the dataset more robust conclusions could be drawn about the algorithms that, in theory, could be adapted for use on any form discrete data. However from the work done here, there is there is evidence that the adapted Poisson C / Poisson L algorithm is a promising technique for the analysis of deep sequencing data.

While the emergence of deep sequencing techniques is relatively new and still somewhat unexplored in terms of statistical analysis, it has the potential to become the most prominent technique in the sequencing of DNA due to the large number of tags it can identify. There is a need to develop appropriate analysis techniques for the analysis of such large but sparse datasets. The work represented here provides a useful contribution in this direction.

Bibliography

[1] National Human Genome Institute. (2008, Nov.) DNA sequencing fact sheet. [Online]. http://www.genome.gov/10001177

[2] C. A. Hutchison, "DNA sequencing: bench to bedside and beyond." *Nucleic Acids Research*, vol. 1, no. 11, pp. 6227-6337, 2007.

[3] A. L. Hopkins and C. R. Groom, "The druggable genome." *National Review of Drug Discovery*, vol. 1, 2002.

[4] M. A. Metzker, "Emerging technologies in DNA sequencing." *Genome Research*, vol. 15, pp. 1767-1776, 2005.

[5] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain terminating inhibitors." *Proceedings of the National Academy of Science*, vol. 74, pp.5463-5467

[6] (2008) mondofacto medical dictionary. [Online]. http://www.mondofacto.com/facts/dictionary?peptide+chain+termination

[7] N. Hall, "Advanced sequencing technologies and their wider impact in microbiology." *Journal of Experimental Biology*, vol. 210, no. 9, pp.1518-1525, 2007.

[8] M. Ronaghi, "Pyrosequencing sheds light on DNA sequencing." *Genome Research*, Vol. 11, pp. 3-11, 2001.

[9] (2007, Dec.) 454-life sciences. [Online]. http://www.454.com/downloads/news-events/publications/HIV%20Resistance%20Workshop_454_final.jpg

[10] B. Wire. (2009, Jan.) Thomson Reuters. [Online]. http://www.reuters.com/article/pressRelease/idUS118722+27-Jan-2009+BW20090127

[11] R. D. Morin, et al., "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells." *Genome*

Research, vol. 18, pp. 610-621, 2008.

[12] S. Nygaard, et al., "Identification and analysis of miRNAs in human breast cancer and teratoma samples using deep sequencing." *BMC Medical Genomics*, vol. 2, Article no. 35, 2009.

[13] C. M. Croce, "MicroRNAs and lymphomas." *Annals of Oncology*, vol. 19, no. 4, pp. 28-29, 2008.

[14] T. Thum, P. Galuppo, and C. E. A. Wolf, "MicroRNAs in the human heart: A Clue to Fetal Gene Reprogramming in Heart Failure." *Circulation*, vol. 116, pp. 258-267, 2007.

[15] Thermo Fisher Scientific. (2008) MicroRNAs: Review of, discovery, biogenesis, and research areas [Technical Review]

[16] C. Wei, J. Li, and R. E. Bumgarner, "Sample size for detecting differentially expressed genes in microarray experiments." *BMC Genomics*, vol. 5, Article no. 87, 2004.

[17] K. A. Baggerly, L. Deng, J. S. Morris, and C. M. Aldaz, "Overdispersed logistic regression for SAGE: Modelling multuiple groups and covariates." *BMC Bioinformatics,* vol. 5, Article no. 144, 2004.

[18] M. Sun, et al., "SAGE is far more effective than EST for detecting low-abundance transcripts." *BMC Genomics*, vol. 5, Article no. 1, 2004.

[19] F. N. David and N. L. Johnson, "The Truncated Poisson." *IBS*, vol. 8, no. 4, pp. 275-285, 1952.

[20] M. D. Robinson and G. K. Smyth, "Small Sample Estimation of Negative Binomial Dispersion, with applications to SAGE data." *Biostatistics,* Vol.9*,* no. 2, pp. 321-332, 2007.

[21] J. A. Hartigan, *Clustering algorithms*. Yale: Wiley series in probability and mathematical statistics, 1975.

[22] E. A. Cai, "Clustering analysis of SAGE data using a Poisson approach." *Genome Bology*, vol 5, Issue 7, Article R51, 2004.

[23] K. Kim, et al., "Measuring similarities between gene expression profiles through new data transformations." *BMC Bioinformatics,* vol. 8, Article no.

29, 2007.

[24] P. Berninger, D. Gaidatzis, E. Van Nimwegen, and M. Zavolan, "Computational analysis of small RNA cloning Data." *Methods in Computational Biology*, vol. 44, no. 1, pp. 13-21, 2008.

[25] S. Audic and J.-M. Claverie, "The significance of digital gene expression profiles." *Genome Research,* vol. 7, pp.986-995, 1997.

[26] K. A. Baggerly, L. Deng, J. S. Morris, and M. C. Aldaz, "Differential esxpression in SAGE: accounting for normal between-library variation." *Bioinformatics*, vol. 19, no. 12, 2003.

[27] D. J. Stekel, Y. Git, and F. Francesco, "The comparison of Gene expression from multiple cDNA libraries." *Genome Research,* vol. 10, pp.2055-2061, 2000.

[28] J. Lu, J. K. Tomfohr, and T. B. Kepler, "Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach." *BMC Bioinformatics*, vol. 6, Article no. 165, 2005.

[29] S. D. Zuyderduyn, "Statistical analysis and significance testing of serial analysis of gene expression data using a Poisson mixture model." *BMC Bioinformatics*, vol. 8, Article no. 282, 2007.

[30] R. Khanin and E. Wit, "How Scale-Free Are Biological Networks." *Journal of Computational Biology*, vol. 13, no. 3, pp. 810-818, 2006.

[31] (2004,Jun.) About HiSee. [Online]. http://hisee.sourceforge.net/about.html

[32] E. Wit and J. McClure, *Statistics for Microarrays: Design, analysis and Inference*. WileyBlackwell, 2004.

[33] Portal Robert Gentleman. (2009, Sep.) Differential Expression. [Online]. http://gentleman.fhcrc.org/Ghent/Lectures/DiffExpr/

[34] M. Z. Man, X. Wang, and Y. Wang, "POWER_SAGE: comparing statistical tests for SAGE experiments." *Bioinformatics*, vol. 16, no. 11, pp. 953-959, 2000.

[35] J. M. Ruijter, A. H. C. Van Kampen, and F. BAAS, "Statistical evaluation of SAGE libraries: consequences for experimental design." *Physiological Geneomics*, vol. 11, pp. 37-44, 2002.

[36] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a dataset via the GAP statistic.” Journal of the Royal Statistical Society, vol. 63, no. 2, 411-423, 2001.

[37] R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna,

Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[38] Casella G, Berger RL; “Statistical Inferences” 2nd edition. Pacific Grove, CA; DuXBURY; 2002