DIFFERENTIAL EXPRESSION ANALYSIS BETWEEN MICROARRAY AND RNA-SEQ OVER ANALYTICAL METHODS ACROSS STATISTICAL MODELS

by

YUHAO WU

Submitted in partial fulfillment of the requirements

For the degree of Master of Science

Department of Electrical Engineering and Computer Science

CASE WESTERN RESERVE UNIVERSITY

May, 2020

Differential Expression Analysis between Microarray and RNA-seq over Analytical Methods across Statistical Models

Case Western Reserve University
Case School of Graduate Studies

We hereby approve the thesis[1] of

YUHAO WU

for the degree of

Master of Science

Dr. Fredrick Schumacher

Committee Chair, Advisor — Apr. 7th 2020
Department of Population and Quantitative Health Sciences

Dr. Jing Li

Committee Member, Academic Advisor — Apr. 7th 2020
Department of Electrical Engineering and Computer Science

Dr. Satya Sahoo

Committee Member — Apr. 7th 2020
Department of Population and Quantitative Health Sciences

Dr. Mehmet Koyuturk

Committee Member — Apr. 7th 2020
Department of Electrical Engineering and Computer Science

[1]We certify that written approval has been obtained for any proprietary material contained therein.

Dedicated to my advisor Fredrick Schumacher, who have always supervised me during my research time. Also to Jing Li, who had given me so many great suggestions.

Table of Contents

List of Tables

List of Figures

Acknowledgements

0.1 Acknowledgements

Foremost, I would like to thank my advisor Dr. Fredrick Schumacher in the Department of Population and Quantitative Health Sciences. He lead me into the research field and helped me a great deal in finishing this work. I was always enlightened by his advice.

In addition, I would like to thank Dr. Jing Li who was my academic advisor. I really appreciate the valuable suggestions he has given me.

Furthermore, I would like to express my thanks to the rest of my committee: Dr. Satya Sahoo and Dr. Mehmet Koyuturk. Thank you for your interests in my thesis, your support, comments and questions are all essential to this work.

Finally, thanks to my parents and those who have helped me throughout the years of study.

Abstract

Differential Expression Analysis between Microarray and RNA-seq over Analytical Methods across Statistical Models

Abstract

by

YUHAO WU

0.2 Abstract

Microarray and RNA-seq are two transcriptomics data types. Comparison between microarray and RNA-seq was popular in recent research fields. But few works analyzed both data types on the same sample groups and fewer have discussed whether different analytical methods or statistical models will impact on the results. To get an insight of what role analytical method and statistical model play, and also to eliminate variances that might be caused by data types, we applied microarray and RNA-seq data into 5 analytical methods across 4 statistical models to contrast the similarities and differences.

In this thesis, we processed and transformed raw microarray and RNA-seq data to fit in 5 different analytical methods across 4 statistical models. We did differential expression analysis on both data for all methods and evaluated the results. Both data types showed high consistency in those two methods applying the same model. All statistical models gave similar detected differentially expressed genes, the extent of similarity varied basing on specific pairs chosen to be compared. Among all analytical methods,

NBPSeq and NOIseq are the most consistent for microarray while DESeq and baySeq are the most consistent for RNA-seq. Between data types, RNA-seq gives more detected differentially expressed genes. Overall, statistical models and data types both impact greatly on differential analysis results while analytical method seems trivial.

1 Introduction

1.1 Transcriptomics Data: Microarray and RNA-seq

Among all analyses of genes, gene expression analysis has the potential to be the very impactful one of them. Gene expression is the process where genetic information is transcribed from DNA to mRNA then translated to proteins. Transcriptomics, the ability to measure the sum of the molecules from DNA to mRNA, is growing in the recent field of bioinformatics. RNA expression data, the most typical data for transcriptomics, has been widely applied across several research areas. For years, microarray data was the dominant form of transcriptomics data available for researchers. This changed in 2009 when technological advances made RNA-seq an alternative for microarray data.[1] As highlighted in the Literature Review section, several papers have discussed similar and different features of these two kinds of data types.[2] Both data types have pros and cons.

A microarray is a laboratory tool used to detect the expression of thousands of genes at the same time. DNA microarrays are microscope slides printed with thousands of tiny spots in defined positions, where each spot contains a known DNA sequence or gene.[3] Therefore microarrays can only detect pre-assigned printed sequences available from

commercial companies. Additional issues with microarray also include cross-hybridization artifacts and poor quantification of lowly and highly expressed genes.[4]

RNA-sequencing, short for RNA-seq, is a technique that can examine the quantity and sequences of RNA in a sample using next generation sequencing.[5] Typically this technique detects short segments of RNA, which may not be informative or cause errors, so it requires quality control, which can be complicated. Since most of the sequences have fairly low counts, dealing with them is a difficult task. Furthermore, different approaches applied to read mapping or alignment may show different detected gene types and expression levels.[6] Nookaew et al[7] compared *de novo* assembly and reference mapping in many aspects. Their conclusion was that in general they are in good agreement. The advantage is that RNA-seq is not limited to pre-assigned genes, however the expense of RNA-seq limits its full utility.

Cost is the primary reason researchers are limited from obtaining both microarray data and RNA-seq data, not to mention the redundancy and inefficiency, so one is always enough.[8] However, some results can be driven by the technology. For example, unlike microarray, RNA-Seq technique does not require species- or transcript-specific probes, so it can detect novel transcripts. This can lead to differences but how significant these differences can be are still unknown to us. Few papers have clearly discussed the differences in performance when both data types are used in one experiment.

1.2 Statistical Methods for Transcriptomics

Apart from different kinds of data types multiple statistical approaches, including the underlying model and the method, have emerged during the last decade.[9] Many researchers have proposed analytical approaches with varying underlying models, and

methods, to analyze expression data for purposes like finding differentially expressed genes. However, those approaches sometimes give different outcomes. Determining which method gives better results would be very helpful for future studies of gene expression. The statistical methods involved in this thesis will be further explained in the Methods section.

Yet detailed difference analysis on both microarray and RNA-seq data using different analytical methods is absent. Filling this scientific gap can give direction choosing on a method and data type for further research projects of gene expression analysis. In the Literature Review section, we will simply introduce what other researchers have done in the related scientific fields and why these works are valuable to our research. According to the literature review, microarray and RNA-seq had been compared in many aspects: mechanics, advantages and disadvantages and accuracy in assessing expression levels. However, their consistency in differential expression analysis has never been considered. Therefore, our aim is to reveal the consistency between microarray and RNA-seq data by conducting experiments using different methods.

In the Method chapter, we will briefly introduce the mechanics of each method we use in the experiments. Since our focus is not how each method works, but how microarray and RNA-seq data perform on different methods, we won't explain the methods across all details.

1.3 Following Experiments

In the Experimental chapter and Results chapter we will compare results across three aspects: data types, underlying statistical model and between methods using the same

model. Our aim is to prove that the comparison between data types will show that microarray and RNA-seq are consistent in highly differentially expressed genes, the comparison among models will give different levels of consistency and those two methods in the same model will have a high consistency.

For the experimental designs, we utilize flow chart in Figure 1.1. The raw data was generated by Trapnell et al[10]. The original experiment used lung fibroblasts and the difference between two groups is the transcription factor HOXA1 knock-down. It was designed for differential analysis of HOXA1 in adult cells at isoforms resolution by RNA-seq. We use their microarray and RNA-seq data to do differential analysis on five different methods, belonging to four statistical models. The three aspects of comparison we focus on are shown in the flow chart.

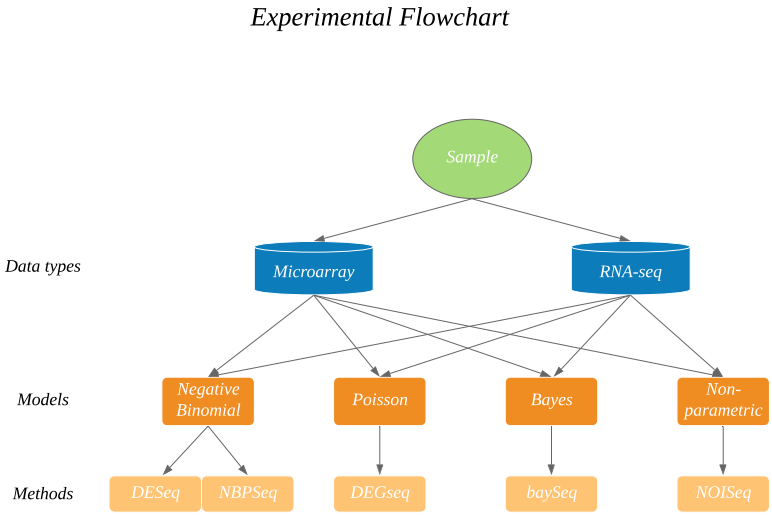


Figure 1.1. Experiment Flow Chart
Mainly focused differential analysis result comparisons in three aspects: between data types microarray and RNA-seq; between 4 different statistical models; between analytical methods using the same or different statistical models

In general, we will first assess sample gene expression levels and do differential analysis on both microarray and RNA-seq and compare the results. Then we will apply 4 statistical models on both data types to see the similarities and differences. At last, consistency between analytical methods using the same or different statistical models will be measured. The detailed experimental design will be discussed in Experimental section.

2 Literature Review

Publications in related fields will always reveal research trends. Studying what other scholars have achieved can help us understand what happened and determine the next steps. In this chapter, we will discuss about what we can get from differential expression analysis, what other researchers have achieved in comparing microarray and RNA-seq and what statistical methods and normalization we could use to complete the comparison.

2.1 Gene Differential Expression Analysis

Gene differential analysis has been widely used in transcriptomics studies. When two groups of samples show different features, their genes will most likely be significantly differentially expressed. In contrast to other analyses such as expression analysis or cluster analysis, differential analysis studies those differentially expressed genes between two conditions. The differentially expressed genes may solve the secret why the two groups show different features. Review papers[11] in the last decade talked about differential expression analysis in both microarray and RNA-seq, while up-to-date review papers[12] focus more on RNA-seq.

2.2 Microarray and RNA-seq

Mantione et al[13] compared microarray and RNA-seq in the perspective of how both techniques actually work. Their conclusion was that RNA-seq will eventually be used more routinely than microarray, but right now the techniques can be complementary to each other. Their work is an introduction and review of microarray and RNA-seq workflow and sample preparations. It focused on comparing the price and reliability but lacks experiments to support their point of view. We can see that comparison between microarray and RNA-seq has raised interest from researchers but solid experiments are needed to obtain a better insight of the differences.

Some researchers noticed and have talked about differences between RNA-seq and microarray when doing transcriptome profiling. For example, Zhao et al[14] performed both RNA-seq and microarray analyses on a sample taken in six different time points. It claimed that microarray and RNA-seq showed a high correlation between gene expression profiles generated by the two platforms. RNA-seq was superior in detecting low abundance transcripts, differentiating biologically critical isoforms, and allowing the identification of genetic variants. What they have tested are only gene expression level for six time-tags, it lacks comparison between sample groups, which might yield a different result.

The work by Fu et al[8] estimated accuracy of RNA-seq and microarrays with proteomics. They first measured mRNA expression levels from both microarray and RNA-seq data and observed agreement between them. Then they measured mRNA expression levels from protein data using 2D LC-MS/MS analysis. The results from protein data were their “golden standard” to estimate microarray and RNA-seq absolute mRNA

expression levels measurement accuracy. They brought a brilliant idea by using proteomics data as the “golden standard” when comparing microarray and RNA-seq in transcriptomics. It is novel to estimate accuracy using data from different sources. But still, their choice of a “golden standard” makes the reliability controversial. Since proteomics and transcriptomics provide data from different stages of gene transcription and translation, the respective expression analysis result can be a kind of validation, but using results from proteomics data as a golden standard for transcriptomics data seems unconvincing.

As shown in Castillo et al[15], microarray data can be integrated with RNA-seq data. This work used heterogeneous data from microarrays and RNA-seq technologies and implemented a new classification and diagnosis tool. This work gives credit to combining microarray and RNA-seq data in differential expression analysis. However, this integration needs preparation. According to Babu et al[16] Chapter 11, An Introduction to Microarray Data Analysis, the data from a microarray experiment requires transformation and normalization before analyses. Usually after the experiments we can get signals and intensities for the predesigned probes. But the raw intensities for probes is not exactly the same as the raw counts for genes, what we can get from RNA-sequencing. Since the statistical methods we will use are all originally designed for RNA-seq, raw intensities from microarray cannot be directly used. In order to get both data to work for the same models, we will need quality control, specifically transformation and normalization. This process can be considered as a preparation for microarray data. Babu et al[16] has detailed steps of converting signal into intensity and we can reverse those steps to get "raw counts" for the probes. After all the transformation and normalization, microarray data can now fit into statistical models.

Though microarray and RNA-seq data can be integrated, their data distributions are different. Most of the time, intensity of microarray, especially after log fold change, follows a normal distribution. While raw count from RNA-seq usually follows discrete probability distributions like a binomial distribution, negative binomial distribution or Poisson distribution. In most cases, these distributions have different features, but under some certain circumstances they can be approximated. For example, normal approximation to a negative binomial distribution is valid when the number of required successes, s, is large, and the probability of success, p, is neither very small nor very large. The normal distribution can be used as an approximation to the binomial distribution and Poisson distribution as well. Most of the time, the processed data may not strictly follow the distributions the models assumed but this doesn't influence model robustness significantly. Work by Lu et al[17] shows that an assumption of a negative binomial distribution can be robust even if the data are not truly negative binomially distributed.

2.3 Statistical Models and Analytical Methods

The review paper by Huang et al[6] listed the most up-to-date statistical models and analytical methods and reviewed how each model fits RNA-seq data. This paper listed 5 different models, respectively Poisson, negative binomial, beta binomial, Bayesian and empirical Bayesian and Non-parametric. Huang et al[6] collected different tools that can be used in differential expression analysis for RNA-seq.

In this study we will apply models using Poisson distribution and negative binomial distribution. A Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the

last event[18]. A negative binomial distribution is a discrete probability distribution of the number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified number of failures occurs.

2.4 Normalization

Normalization sometimes plays a significant role in differential expression analyses. As mentioned above, normalization is required for microarray, it is also crucial for RNA-seq. Normalization methods can remove bias, as Kreil et al[19] demostrated by comparing differential protein expression analysis results using different normalization methods. Work by Bullard et al[20] evaluated various normalization techniques and found that more general quantile-based procedures yield much better concordance with data and are hopefully more robust than normalization by a single housekeeping gene. Furthermore, Maza et al[21] revealed that bias due to the relative size of transcriptomes leads to poor estimations of ratios of gene expressions, and consequently to biased differential expression analysis. From those works we can tell that it's necessary to normalize data before making comparisons but normalization methods won't make significant difference.

3 Methods

Work by Trapnell et al[10] provided the inspiration to investigate the differences and similarities in all three aspects. Their work was to present Cuffdiff 2, an algorithm that estimates expression at transcript-level resolution and controls for variability evident across replicate libraries. The experiment they generated was designed for differential analysis of HOXA1 in adult cells at isoform resolution. Their original purpose was to prove that Cuffdiff 2 performs robust differential analysis in RNA-seq experiments at transcript resolution, revealing a layer of regulation not readily observable with other high-throughput technologies. Although their purpose was different from ours, their raw data generated from the experiments can be directly used for differential analysis. In the following part of this chapter, we will introduce the methods involved in accomplishing differential analysis.

All the experiment designs follow the flowchart in Figure 1.1. Our experiment focuses on comparisons across three aspects: data types, underlying statistical model and between analytical methods using the same statistical model. The raw data was generated by Trapnell et al[10]. Their original experiment used lung fibroblasts and the difference between two groups is the transcription factor HOXA1 knock-down. It was designed for differential analysis of HOXA1 in adult cells at isoform resolution by RNA-Seq.

The whole experiment process is shown in Figure 3.1. The raw data from Illumina and Agilent platforms are stored in NCBI databases. We extracted raw data and processed into count matrices to fit into differential analyses methods. Green icons represent platforms, blue for databases, pink for aligning tools, cyan for intermediate data, yellow for differential analyses methods and red for input and output data.

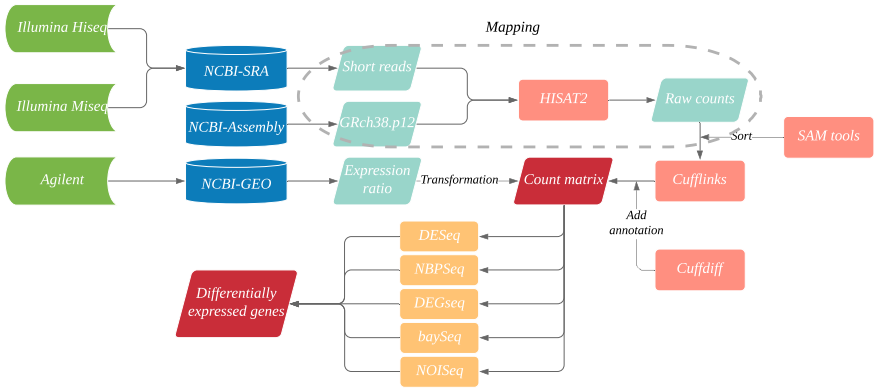


Figure 3.1. Experiment Design Flow Chart
The flow chart demonstrated how raw data are extracted, processed into count matrices and fit into analytical methods

3.1 Platforms

All biological experiments were conducted by Trapnell et al[10], and we evaluated data generated by those experiments. The platforms they use to generate data are briefly introduced in the following sections.

3.1.1 Microarray Platform

The platform for assaying microarray gene expression data was Agilent. Agilent Gene Expression Microarray Platform includes whole transcriptome gene expression for almost 30 different species, exon microarrays to analyze splicing variants and expression

microarrays with comprehensive content. This platform also offers all reagents needed to successfully process microarrays, delivering reliable and reproducible results. The probe used is Agilent-028004 SurePrint G3 Human GE 8x60K Microarray. The result matrix is a 164 by 384 array.

3.1.2 RNA-seq Platforms

There are several platforms for RNA-seq and the mostly used ones are Illumina. In this thesis we will use RNA-seq data from both Illumina MiSeq and Illumina HiSeq. MiSeq focused applications such as targeted resequencing, metagenomics, small genome sequencing, targeted gene expression profiling, and more. MiSeq reagents enable up to 15 Gb of output with 25 million sequencing reads and 2 × 300 bp read lengths. The HiSeq 2500 System is a powerful high-throughput sequencing system. High-quality data using proven Illumina SBS chemistry has made it the instrument of choice for major genome centers and research institutions throughout the world.

3.2 Quality Control

3.2.1 Transformation for Microarray

We can get access to raw intensities and expression levels from GEO. As mentioned above in literature review, expression level is log2 of normalized signal intensity ratio:

$$E_k = log_2(T'_k)$$

where E_k is expression level of gene k and T'_k is the normalized expression ratio T_k of gene k, which is:

$$T_k = \frac{R_k}{G_k}$$

where R_k represents the spot intensity metric for the test sample and $G_k$represent the spot intensity metric for the reference sample. In order to eliminate the influence of background intensity, the spot intensity is replaced by background subtracted median value then the median expression ratio for a given spot is:

$$T_{median} = \frac{R^{spot}_{median} - R^{background}_{median}}{G^{spot}_{median} - G^{background}_{median}}$$

where R^{spot}_{median} and $R^{background}_{median}$ are the median intensity values for the spot and background respectively, for the test sample. The normalization factor can be calculated as:

$$N_{total} = \frac{\sum_{k=1}^{N_{gene-set}} R_k}{\sum_{k=1}^{N_{gene-set}} G_k}$$

so that the normalized expression ratio becomes:

$$T'_k = \frac{R_k}{G_k \times N_{total}} = \frac{T_k}{N_{total}}$$

In order to use absolute value of expression to fit models designed for count matrix from RNA-seq data, we get:

$$R_{spot} = 2^{E_k} \times N_{total} \times (G_{spot} - G_{background}) + R_{background}$$

Since we know the starting amount of mRNA, we can make use of the absolute value.

3.2.2 Aligner for RNA-seq

HISAT2[22] is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes (as well as to a single reference genome). Based on an extension of BWT, Burrows–Wheeler Transform[23], for graphs, it is a graph FM index.[24] Although being mentioned a lot in many previous publications, TopHat[25], an efficient read-mapping algorithm designed to align

reads from an RNA-seq experiment to a reference genome, is almost replaced by HISAT2. The reference genome is Genome Reference Consortium Human Build 38 patch release 12 (GRCh38.p12). HISAT2 has mapping quality control functions built in so what we need to do is to set parameters and let the alignment tool to filter poorly mapped reads.

3.3 Differential Expression Analysis Methods

The methods designed to be used in the experiments are all for differential expression analysis for RNA-seq originally[6]. They can be used for microarray data as well if proper normalization techniques are applied. The normalization methods may vary according to models. In the following sections we will get a general idea about the five statistical methods and other methods used in the experiment.

3.3.1 Statistical Models and Analytical Methods

The models are: Poisson, negative binomial, beta binomial, Bayesian and empirical Bayesian and non-parametric. We will use $DEGseq$[26], $NBPSeq$[27], $BBSeq$[28], $baySeq$[29] and $NOIseq$[24] respectively. BBSeq lacks maintenance and requires R version older than 3.0, making it not suitable for comparison. Thus, we added $DESeq$[30], which uses negative binomial as a data distribution model, into comparison. Using two analytical methods in one statistical model can yield more comparison information within a statistical model.

DEGseq. DEGseq is a R package for identifying differentially expressed genes from RNA-seq data. It supports raw read counts or normalized gene expression values and assumes a Poisson distribution. It uses Fisher's exact test and likelihood ratio test to identify differentially expressed genes.

NBPSeq. NBPSeq uses RNA-seq data. It believes that commonly used probability distributions, such as binomial or Poisson, cannot appropriately model the count variability in RNA-seq data due to over-dispersion. It introduces an additional parameter to allow the dispersion to depend on the mean. It uses an adapted exact test proposed by Robinson and Smyth[31] to get differentially expressed genes and proved to be robust even when data departs from model assumptions.

DESeq. DESeq is a pretty mature algorithm for detection of differentially expressed genes using count data from RNA-seq. The algorithm detects differential expression by use of the negative binomial distribution. The count data needs to be normalized according to the effective library size. The tests it utilizes are exact test and likelihood ratio test.

baySeq. The algorithm baySeq assumes a negative binomial distribution for the data and derives an empirically determined prior distribution from the entire dataset. BaySeq uses an empirical Bayes approach to detect patterns of differential expression. The testing strategy is to first estimate an empirical distribution on the parameters of the negative binomial distribution, then evaluate posterior probability for inference.

NOIseq. NOISeq is a non-parametric approach for the differential expression analysis of RNA-seq data. NOISeq creates a null or noise distribution of count changes by comparing the number of reads of each gene in samples within the same condition. This reference distribution is then used to assess whether the change in count number between two conditions for a given gene is likely to be part of the noise or represents a true differential expression.

3.3.2 Gene Clustering

There are plenty of novel machine learning methods we can choose for clustering the detected differentially expressed genes.[32] Since we care more about correlations among

those genes, PCA, principal component analysis[33], will be used to do the clustering. PCA is a statistical procedure that uses an orthogonal transformation which converts a set of correlated variables to a set of uncorrelated variables. PCA is a widely used tool in exploratory data analysis and in machine learning for predictive models. Using PCA for gene clustering is reliable[34], and the results can be further interpreted to find biological meanings.

3.3.3 Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) (also functional enrichment analysis) is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with disease phenotypes. The method uses statistical approaches to identify significantly enriched or depleted groups of genes. Transcriptomics technologies and proteomics results often identify thousands of genes which are used for the analysis.[35] The tool we used in the thesis is PANTHER[36] on Gene Ontology.

3.4 Evaluation

When comparing performances between two methods, we will use both Pearson correlation coefficient and Spearman's rank correlation coefficient. The Pearson correlation coefficient is a measure of the linear correlation between two variables. Spearman's rank correlation coefficient can be used to assess monotonic relationships. When p-values from two methods are compared, Pearson correlation coefficient is the best

choice. When p-values are compared to likelihoods, Spearman's rank correlation coefficient would be a better choice. To evaluate the performances of clustering, mutual information will be calculated.[37]

4 Experimental

In this chapter we perform differential expression analysis for both microarray and RNA-seq data applying 5 different analytical methods: $DESeq$[30], $NBPSeq$[27], $DEGseq$[26], $baySeq$[29], and $NOISeq$[24]. In order to standardize our comparison of analytical methods we processed publicly available microarray and RNA-seq expression data.

4.1 Data description

The microarray and RNA-seq expression data is from GSE37704[10] and is publicly available. Briefly, the primary objective of the original study was to assess the impact of a transcription factor HOXA1 knock-down in lung fibroblasts. The lung fibroblasts were transfected with either a HOXA1 directed siRNA pool or a scramble non-targeting siRNA control. The samples were divided into two groups according to the treatment after HOXA1 knockdown. One is named HOX and the other SCR for the HOXA1 directed siRNA or the scramble non-targeting siRNA that transfected to the cells. RNA was collected 48 hours after transfection and changes in gene expression were assayed using Agilent microarrays and high throughput RNA sequencing.

The datasets are open source enabling access to raw data from GEO and SRA. The probe used is Agilent-028004 SurePrint G3 Human GE 8x60K microarray. High throughput RNA sequencing was performed on Illumina MiSeq and HiSeq. The following analysis details are described in the following paragraphs.

For each knockdown/control 200 ng of total RNA was amplified and labeled with CY3 using the Agilent Low Input Quick Amp Labeling One Color Kits and hybridized to Agilent SurePrint G3 Gene Expression Microarrays as per manufacturer's specifications. Probe intensities were extracted using the Feature Extraction Software (GE1 Sep09 protocol).

For each RNA sample, they prepared Illumina mRNA-seq libraries using the TruSeq RNA kit (version 1, rev A), using 1 µg of total RNA and prepared according to manufacturer's instruction. For HiSeq 2000 sequencing, eight libraries were pooled per sequencing lane (including libraries not described in this manuscript). One anti-HOXA1 siRNA library and one scrambled control library were pooled in each of three sequencing lanes, resulting in each of the six libraries discussed here being sequenced with 30 million reads. Human lung fibroblast reads are available at GEO accession GSE37704.

4.2 Data Pre-procession

In order to fairly compare microarray and RNA-seq, the same statistical model will be applied to each dataset thus eliminating variance caused by data distribution assumptions. As discussed above, models initially designed for RNA-seq will be applied to both datasets. Most of the software designed for RNA-seq require input format as gene-count

matrices. Since microarrays fail to provide raw matrix equivalents of gene-count matrices in RNA-seq, reasonable transformation is required for the raw data. As for RNA-seq data, we mapped the reads using alignment tools and underwent counting applying counting tools.

4.2.1 Microarray data transformation

The microarray raw data from GEO includes raw intensities and normalized expression levels. In general, gene differential expression analysis first transfers raw intensities to an expression ratio, then into a normalized expression ratio. Since models designed for RNA-seq will be applied to the microarray data, and these models typically use gene-count matrices, our aim is to transfer the raw intensities into absolute expression levels. This approach has been discussed in the Methods chapter and the whole process is implemented in R.

4.2.2 RNA-seq Alignment and Counting

The RNA-seq raw data is from SRA. We used $HISAT2$[22] as the alignment tool and $Cufflinks$[38] as the counting tool. $HISAT2$ is computationally fast and sensitive aligning program for RNA-seq and it has nearly replaced the mapping tool $TopHat$, which previously was the dominant aligner. $HISAT2$ supports SRA accession thus removing the need to directly download raw data. We used Genome Reference Consortium Human Build 38 patch release 12 (GRCh38.p12) as the reference genome. The output file was a sam file so we used Sam tools[39] to sort the mapped reads and then applied Cufflinks to count mapped reads. Annotations were added to the results using $Cuffdiff$[40], a method from Cufflinks.

4.3 Differential Expression Analysis

Since a major objective is to assess performance by different models applying five methods, a differential expression analysis for both data types will be performed. All five methods are implemented in the statistical program R. Common quality control and data normalization functions have been applied. The quality control and normalization skills include but not limited to: quantile normalization, low-count filtering, library size factor normalization, batch size normalization and gene length normalization. To minimize variability arising from differing normalization skills applied to the data, we opted to only normalize library sizes which can be implemented by all methods.

The two sample groups HOX and SCR were divided according to the treatment after HOXA1 knockdown. The detailed differences between them has been discussed. To select differentially expressed genes, we set p-value and q-value thresholds both smaller than 0.05. A smaller p-value threshold will limit false positive discoveries, thus detecting differentially expressed genes with higher proportion of true positives, but in the meanwhile giving more false negatives. While a smaller q-value threshold will provide a small number of false negatives and ensure a smaller false positive rate. DESeq also provides adjusted p-value, p-adj value, where the threshold was set to 0.05 as well. For Bayesian models using the likelihood to identify differentially expressed genes, we chose a cut-off probability value of 0.9.

4.4 Unify Annotations

Following the completion of the differential expression analysis, the microarray results were annotated using probe IDs while RNA-seq utilized Ensembl IDs. Unfortunately,

the annotated IDs across the two data types are neither in the same category system nor directly comparable. The microarray chips used in the experiment was Agilent-028004 SurePrint G3 Human GE 8x60K, where every probe is capable of detecting unique sequences consisting of 60 nucleotides. However, the size of a gene is commonly over 1,000 base pairs long. Thus, a certain sequence on a probe may be mapped to several isoforms, and a gene can occasionally contain sequences that can be found on multiple probes. Most of the time, one probe can be mapped to several isoforms but will finally be mapped to a certain gene. For example, as shown in Figure 4.1, we can see that A_33_P3212630 can be mapped to many Ensembl genes, but all of them are members from FAM90 family.

Probe name	Gene name	Ensembl ID
A_33_P3212630	FAM90A3P	ENSG00000233132
A_33_P3212630	FAM90A4P	ENSG00000249005
A_33_P3212630	FAM90A5P	ENSG00000215373
A_33_P3212630	FAM90A6P	ENSG00000248944
A_33_P3212630	FAM90A7P	ENSG00000285975
A_33_P3212630	FAM90A8P	ENSG00000285937
A_33_P3212630	FAM90A13P	ENSG00000223885
A_33_P3212630	FAM90A14P	ENSG00000285814
A_33_P3212630	FAM90A15P	ENSG00000230045
A_33_P3212630	FAM90A16P	ENSG00000285620
A_33_P3212630	FAM90A17P	ENSG00000285720
A_33_P3212630	FAM90A18P	ENSG00000285657
A_33_P3212630	FAM90A19P	ENSG00000285913
A_33_P3212630	FAM90A20P	ENSG00000233295
A_33_P3212630	FAM90A21P	ENSG00000234749
A_33_P3212630	FAM90A22P	ENSG00000285687
A_33_P3212630	FAM90A23P	ENSG00000285765

Figure 4.1. Probe ID mapping to Ensembl ID example screenshot

To interpret the detected differentially expressed genes from microarray and RNA-seq and make comparisons in the following studies, unifying annotations is essential. Mapping probe IDs to Gene IDs is one way of unifying annotations. The BiomaRt[41]

package provides function mapping probe IDs to Ensembl IDs. After the mapping process, probe IDs will be mapped and converted to Ensembl gene IDs, all using annotations from the same system. Until that can we analyze results from different data types.

5 Results

Differential expression analysis generates a list of differentially expressed genes and their statistical significance contrasting groups defined by treatment, exposure, or cell type, just to name a few. This chapter provides an overview of the analytical results and the consistency between each statistical method applying the same model across data types.

5.1 Differentially Expressed Genes

All the differentially expressed genes detected are listed according to the order of p-value or likelihood in a supplementary file. In statistical hypothesis testing, the probability value (p-value) is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming the null hypothesis is correct. In general, applying a p-value threshold of 0.05 indicates statistical significance.

The Human Genome Project estimated that humans have between 20,000 and 25,000 genes.[42] In this section, we quantify the number of differentially expressed genes identified by several methodological approaches. The counts of differentially expressed genes by statistical models and analytical methods are presented in Table 5.1 ($p<0.05$). It is obvious that by data type, more differently expressed genes were detected using RNA-seq

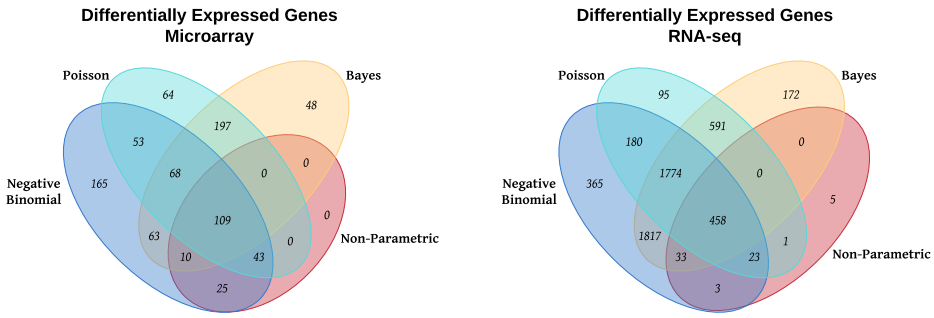
than microarray assays. Specifically, for RNA-seq, 9.1%, 5,517 from a total of 60,658 different Ensembl gene/transcripts were detected differentially expressed by HOXA1 knockdown. On the other hand, microarray data detected about 1.6% from all probes. However different probes can map to the same gene. In the original experiment, the microarray chips used was Agilent-028004 SurePrint G3 Human GE 8x60K, where every probe is capable of detecting unique sequences with 60 nucleotides. The sequence on one probe can be mapped to multiple isoforms and one gene can occasionally contain multiple sequences on different probes. Although we detected more than 1000 probes, a portion of the probes map to the same gene yielding 845 distinct genes. From Table 5.1 we also can see that, across analytical methods, DEGseq gives the most detected genes (534) for microarray and baySeq (4845) for RNA-seq while NOIseq gives the least detected genes for both data types (187 and 523).

Table 5.1. Differentially expressed gene (p-value < 0.05) counts and percentages for all statistical models and analytical methods. The union and overlapping portion between data types are also included.

Statistical Model	Analytical Method	Differentially Expressed Gene Count		Sum (Union)	Overlap/Percentage (Intersection)
		Microarray	RNA-seq		
Negative Binomial	DESeq	210	4631	4662	179 / 3.8%
	NBPSeq	528	1414	1654	288 / 17.4%
Poisson	DEGseq	534	3122	3357	299 / 8.9%
Bayesian and Empirical Bayesian	baySeq	495	4705	4845	355 / 7.3%
Non-parametric	NOIseq	187	523	581	129 / 22.2%
Sum (Union)		845	5517	5845	517 / 8.8%

The overlapping genes between data types are included in supplementary files. Overlapping genes across statistical models are presented in Venn diagrams in Figure 5.1. The figure highlights that across statistical models, most detected differentially expressed genes can be found in at least two models. The negative binomial and Bayesian models

detected the most genes and shared most of their findings. Non-parametric detected the least but most of its findings can be found in other models. All models show consistency to others to some extent, exact consistency levels are discussed in details in the following sections.



(a) (b)

Figure 5.1. Intersection Venn diagrams: (a) DEG counts for Microarray in all statistical models (b) DEG counts for RNA-seq in all statistical models

5.2 Consistency of Analytical Methods across a Statistical Model

In order to understand how two analytical methods applying the same statistical model either differ or remain consistent, we compared the results from DESeq and NBPSeq. Both analytical methods assume a negative binomial model. As shown in Table 5.2, 37.0% and 27.6% of the genes were found by both approaches in the microarray and RNA-seq data. For those differentially expressed genes detected by one but not both analytical methods, NBPSeq detected a greater portion of them (61.5%) for microarray compared to DESeq (47%). The reverse was observed for RNA-seq data where NBPSeq only detected 0.5% significant genes compared to 72.0% by DESeq. From Table 5.2, we observed that for both data types, the detected differentially expressed genes from one

analytical method is almost a subset of the results from the other analytical method. Specifically, for microarray, DESeq shares 96.2% of its findings with NBPSeq; for RNA-seq, NBPSeq shares 98.3% of its findings with DESeq. Therefore, we can draw a conclusion that two analytical methods in one statistical model basically yields consistent results, yet with different resolutions.

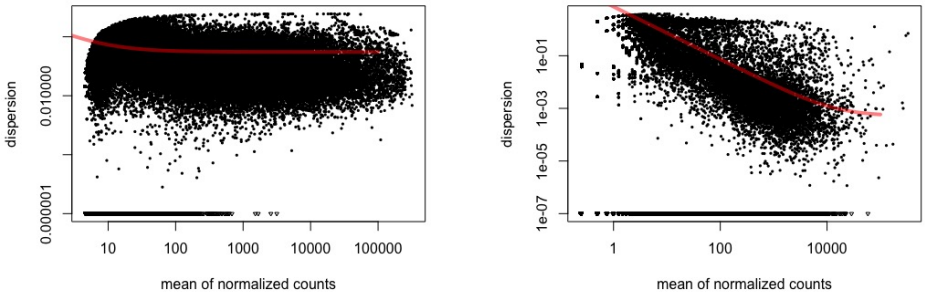
Table 5.2. Differentially expressed gene counts and percentage for DESeq and NBPSeq, both analytical methods using negative binomial distribution as statistical model

Model	Method	DEG count/percentage	
		Microarray	RNA-seq
Negative Binomial	DESeq only	8	3348
	both	202	1283
	NBPSeq only	336	22
Sum		546	4653

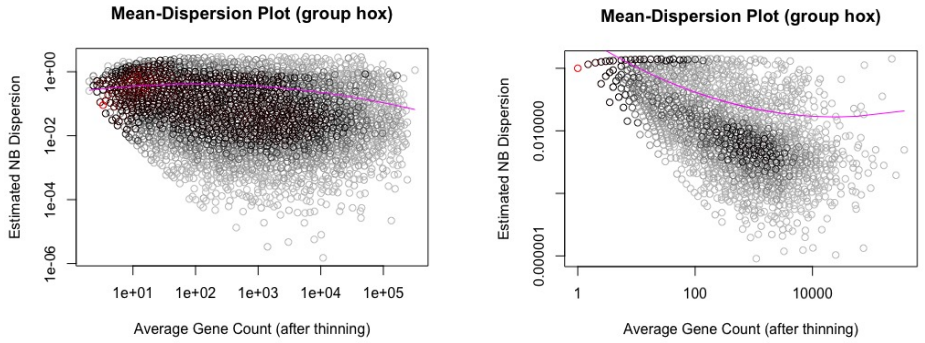
The dispersion of the average gene count (after thinning) and the estimated dispersion is plotted in Figure 5.2 and Figure 5.3. The figure panels include data types microarray and RNA-seq each analyzed with the analytical methods DESeq and NBPSeq. For NBPSeq, the results are divided into groups HOX and SCR. A fitted line (red) of the dispersion values is presented.

According to Figure 5.2, we can see that both analytical methods show similar dispersion, comparing figure panels (a) vs (c) and (b) vs (d), while the differences between data types, microarray and RNA-seq, are very apparent comparing figure panels (a) vs (b) and (c) vs (d). Specifically, fitted average dispersion for microarray are all in the range of 0.1 to 1, regardless of analytical methods. For RNA-seq, the fitted dispersion (red line) ranges from 0.001 to 1, having a greater inclination than microarray.

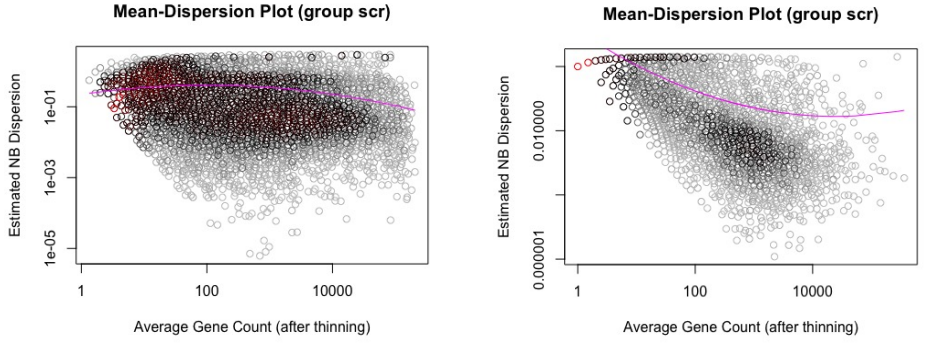
Similarly, according to Figure 5.3, variances are similar between the HOX and SCR groups, comparing figure panels (a) vs (b) and (c) vs (d), but different between data



(a) Microarray DESeq (b) RNA-seq DESeq



(c) Microarray NBPSeq HOX (d) RNA-seq NBPSeq HOX



(e) Microarray NBPSeq SCR (f) RNA-seq NBPSeq SCR

Figure 5.2. Empirical(dots) and fitted(line) dispersion values plotted against mean of normalized counts for combination of analytical methods, data types and groups: (a)DESeq Microarray (b)DESeq RNA-seq (c) NBPSeq Microarray group HOX (d) NBPSeq RNA-seq group HOX (e) NBPSeq Microarray group SCR (f) NBPSeq RNA-seq group SCR

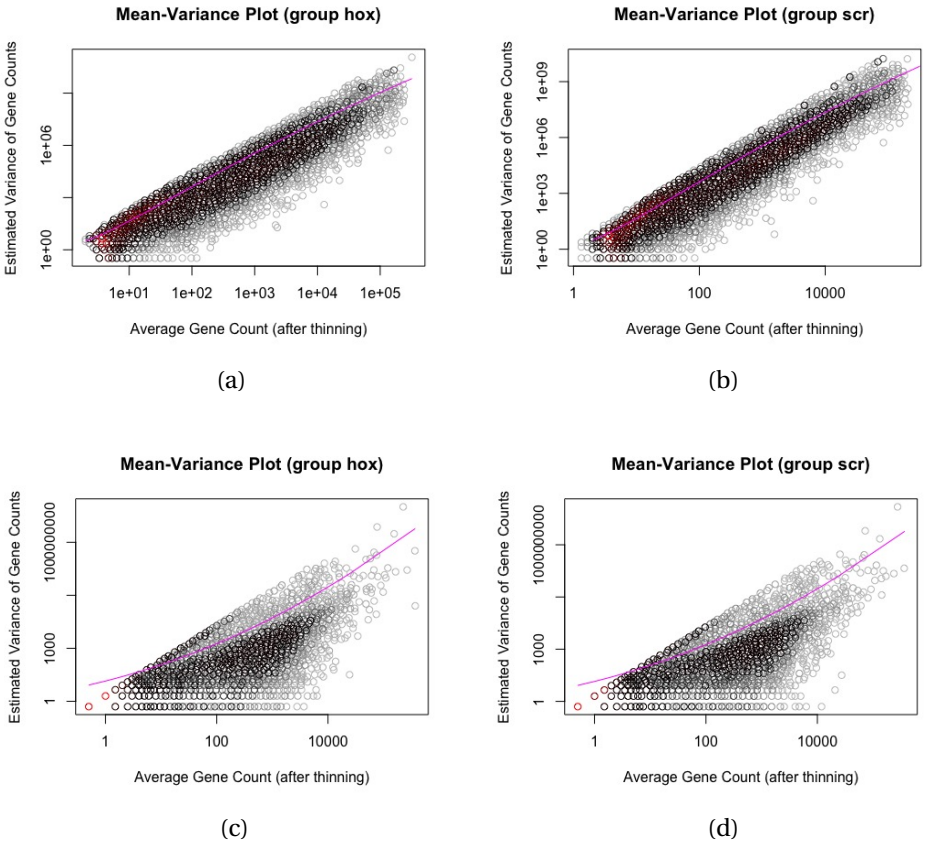
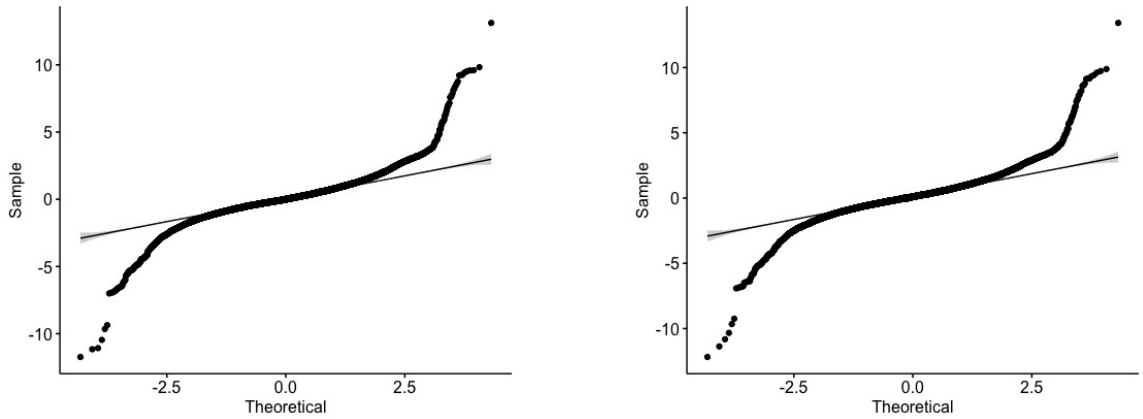


Figure 5.3. NBPSeq mean variance plotted against average gene count: (a) Microarray group HOX (b) Microarray group SCR (c) RNA-seq group HOX (d) RNA-seq group SCR

types, comparing figure panels (a) vs (c) and (b) vs (d). Overall, gene counts from RNA-seq are less dispersed than microarray.

Figure 5.4 highlights that microarray expression levels are normally distributed in both DESeq and NBPSeq. The x coordinate of the qq-plot is the theoretical distribution of the normal distribution and the y coordinate represents the observed distribution.

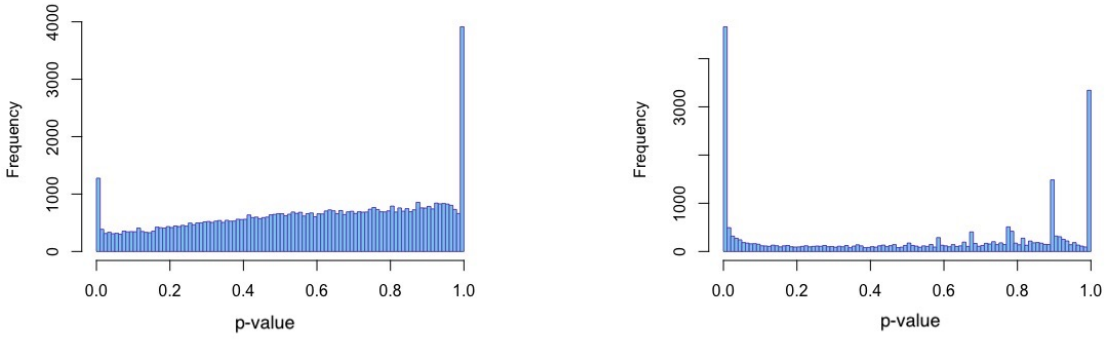
In addition, we plotted the p-value histogram in Figure 5.5 to show the p-value frequency by data type for DESeq. The x-axis is the p-values from microarray (panel a) and RNA-seq (panel b) and the y-axis is respective frequency. Both plots show high



(a) DESeq (b) NBPSeq

Figure 5.4. Normality check for microarray expression levels. X coordinate is for theoretical distribution and y coordinate is for observed distribution.

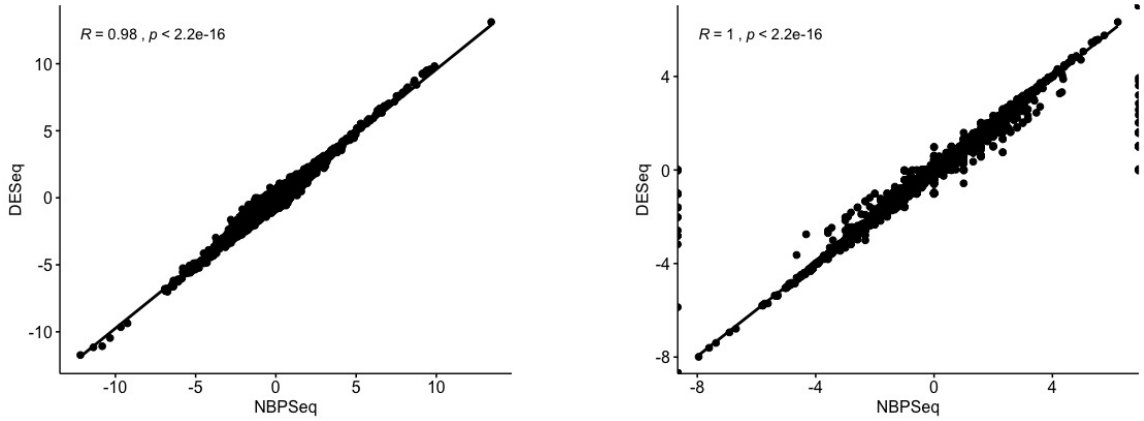
frequency where p-value equals 1 and close to 0, representing totally insignificant and highly significant genes. The p-value distribution also has similar middle parts between data types. This pattern indicates that differential analyses usually determine the majority of the genes either totally insignificant or highly significant.



(a) Microarray (b) RNA-seq

Figure 5.5. Histogram of p-value for DESeq. The frequency are obviously not normally distributed.

In order to contrast the consistency between DESeq and NBPSeq when expression levels are normally distributed, the Pearson correlation was applied. The Pearson correlation coefficient of expression level between method NBPSeq and DESeq for both microarray and RNA-seq are plotted. In Figure 5.6, DEseq's expression level was plotted against NBPSeq's expression level. We should notice that the scales between both plots are different. The Pearson correlations for p-values are 0.82 for microarray and 0.54 for RNA-seq (See Table 5.3 and 5.5).

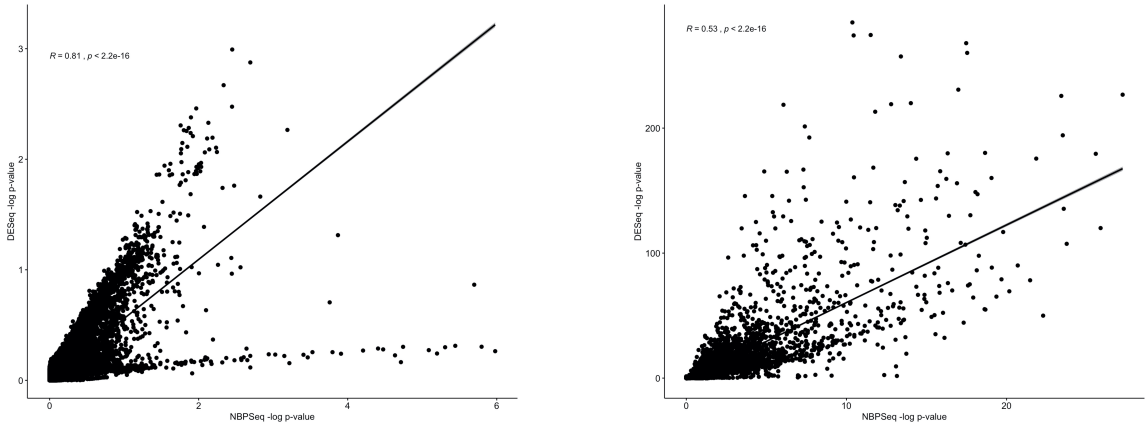


(a) Microarray (b) RNA-seq

Figure 5.6. NBPSeq vs DESeq expression level Pearson correlation. The coefficient between NBPSeq and DESeq for microarray is 0.98, for RNA-seq is 1.0. Should we notice that the scales between plots are different.

The Pearson correlation coefficient for p-value after log transformation has similar results, being 0.67 and 0.79 respectively. It should be noted that Pearson correlation coefficient measures correlation between two sets of values, so it will change after log transformation. All normality checking plots and Pearson correlation plots are included in the supplementary files. According to the results, DESeq and NBPSeq gave highly correlated gene expression levels and p-values, indicating that they are showing high consistency in measuring expression levels.

Similarly, we plotted minus log of p-values from NBPSeq against DESeq in Figure 5.7. For a better view, we removed top 20 statistically significant findings, the original plots can be seen in the supplementary files. The Spearman's rank correlation between them is 0.81 for microarray, 0.53 for RNA-seq (See Table 5.3 and 5.5). Using minus log transformation gives the same correlation from using p-value, because this transformation will not change ranks. The figures show that for both microarray and RNA-seq, NBPSeq and DESeq are highly consistent in assessing p-values.

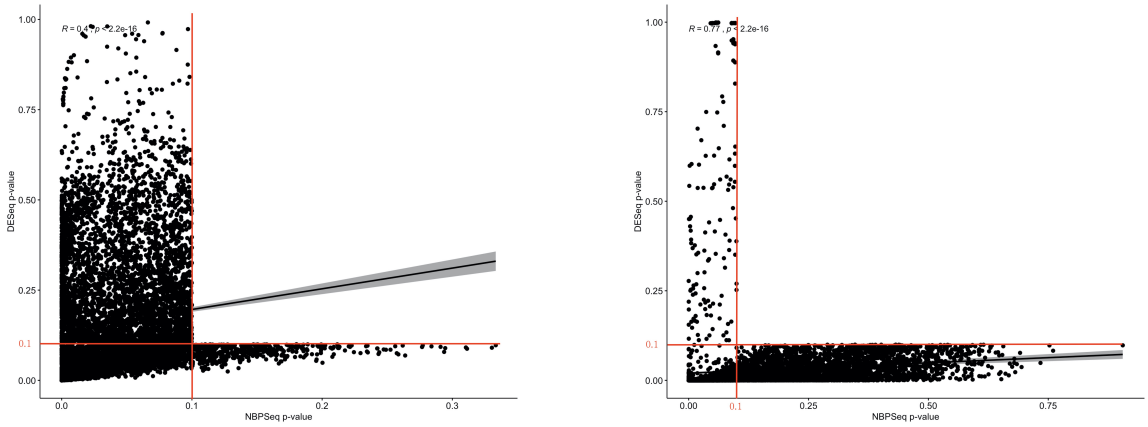


(a) Microarray (b) RNA-seq

Figure 5.7. NBPSeq vs DESeq minus log of p-value Spearman's correlation (removing top 20 findings). The coefficient between NBPSeq and DESeq for microarray is 0.81, for RNA-seq is 0.53.

We care more about those statistically significant differentially expressed genes, so we set p-value cut-off as 0.1, the p-values from DESeq and NBPSeq are plotted in Figure 5.8. The gene counts for microarray is 3984 in the upper left, 3853 in the bottom left and 425 in the bottom right. This indicates most detected genes from DESeq can be found in NBPSeq while only half of the genes detected by NBPSeq can be found in DESeq. Conversely, the results for RNA-seq indicate the opposite. There are 165 genes

in the upper left area, 3525 in the bottom left and 3280 in the bottom right. We are unable to determine which analytical method is better without a "golden standard", but it is clear that DESeq and NBPSeq are showing high consistency in both data types.



(a) Microarray (b) RNA-seq

Figure 5.8. NBPSeq vs DESeq p-value < 0.1 Spearman's correlation. Gene counts: (a) 3984 upper left, 3853 bottom left, 425bottom right (b) 165 upper left, 3525 bottom left, 3280 bottom right

However, for RNA-seq we removed points where their p-values are 1 in one method but smaller than 0.1 in another. This quality control removes 580 points in the upper left area and 11 points in the bottom right, thus changing the correlation from 0.65 to 0.77 (See supplementary files). This partially explains the observed Spearman rank correlation difference between microarray and RNA-seq (0.81 and 0.53), that DESeq and NBPSeq consistently detect significant genes but are not consistent in assessing those insignificant genes, while microarray is consistent all the time.

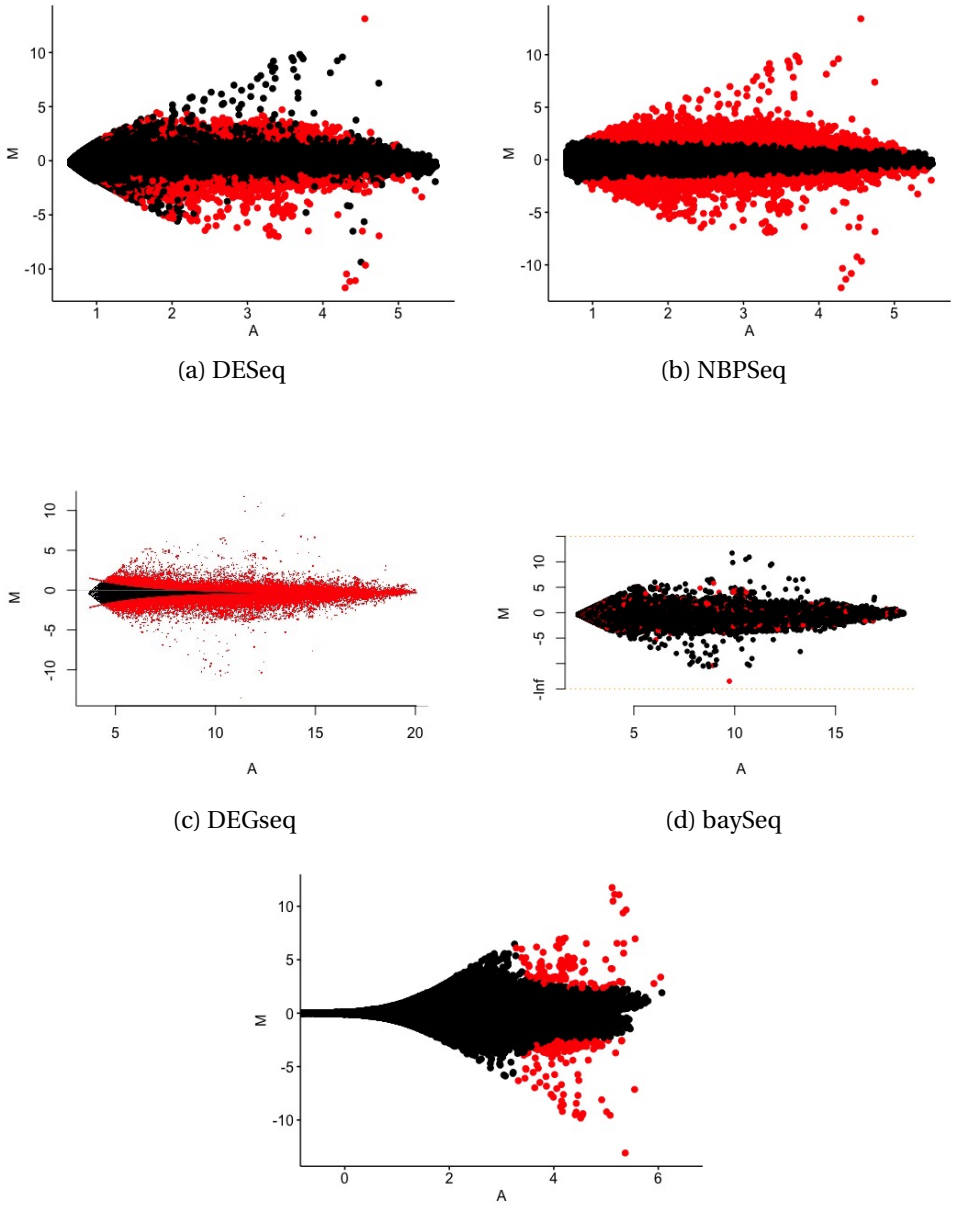
5.3 Consistency among Statistical Models

5.3.1 Microarray

We presented MA plots from all analytical methods for microarray in Figure 5.9. MA plots visualize the differential expression of genes by comparing the mean of normalized counts (x-axis) and the log fold change(y-axis). By comparing log ratio (M) and mean average (A), we have a general idea of the impact and the statistical significance of the data. The red dots are those differentially expressed and statistically significant genes, threshold being smaller than 0.05 for p-values and greater than 0.9 for likelihood. As indicated in the figures, different methods yield different results, in both expression level assessment and differentially expressed genes detected.

But revealing how different their results are relying on further analysis. Here, besides Pearson correlation coefficient, we applied Spearman's rank correlation coefficient which assess monotonic relationships.

See Table 5.3. Pearson correlation coefficients and Spearman's rank correlations were calculated between methods. DESeq, NBPSeq and DEGseq use p-value, baySeq and NOIseq use likelihood. Since some methods use p-value as criteria for selecting differentially expressed genes while some use likelihood, Pearson correlation coefficient may not work well to show consistency between models using different criteria. In case Pearson correlation coefficient performs poorly, we applied Spearman's rank correlation to assess correlations between analytical methods. From the table we can see similar results from both correlation coefficients. For comparison between models using different selection criteria, mutual information in Table 5.4 would be more precise.



(a) DESeq

(b) NBPSeq

(c) DEGseq

(d) baySeq

(e) NOIseq

Figure 5.9. M(log ratio)A(mean average)plots for microarray from analytical methods: (a) DESeq (b)NBPSeq (c) DEGseq (d) baySeq (e) NOIseq. Red dots are those detected differentially expressed genes. Threshold for p-value <0.05, likelihood > 0.9

Table 5.3. Microarray Pearson correlation coefficient / Spearman's rank correlation between analytical methods.

	NBPSeq	DEGseq	baySeq	NOIseq
DESeq	0.82/0.81	0.76/0.75	0.55/0.59	0.84/0.86
NBPSeq		0.65/0.64	0.46/0.46	0.85/0.86
DEGseq			0.42/0.58	0.73/0.69
baySeq				0.33/0.51

Table 5.4. Microarray mutual information between statistical models

	Poisson	Bayes	Non-parametric
Negative Binomial	0.34	0.32	0.35
Poisson		0.57	0.27
Bayes			0.21

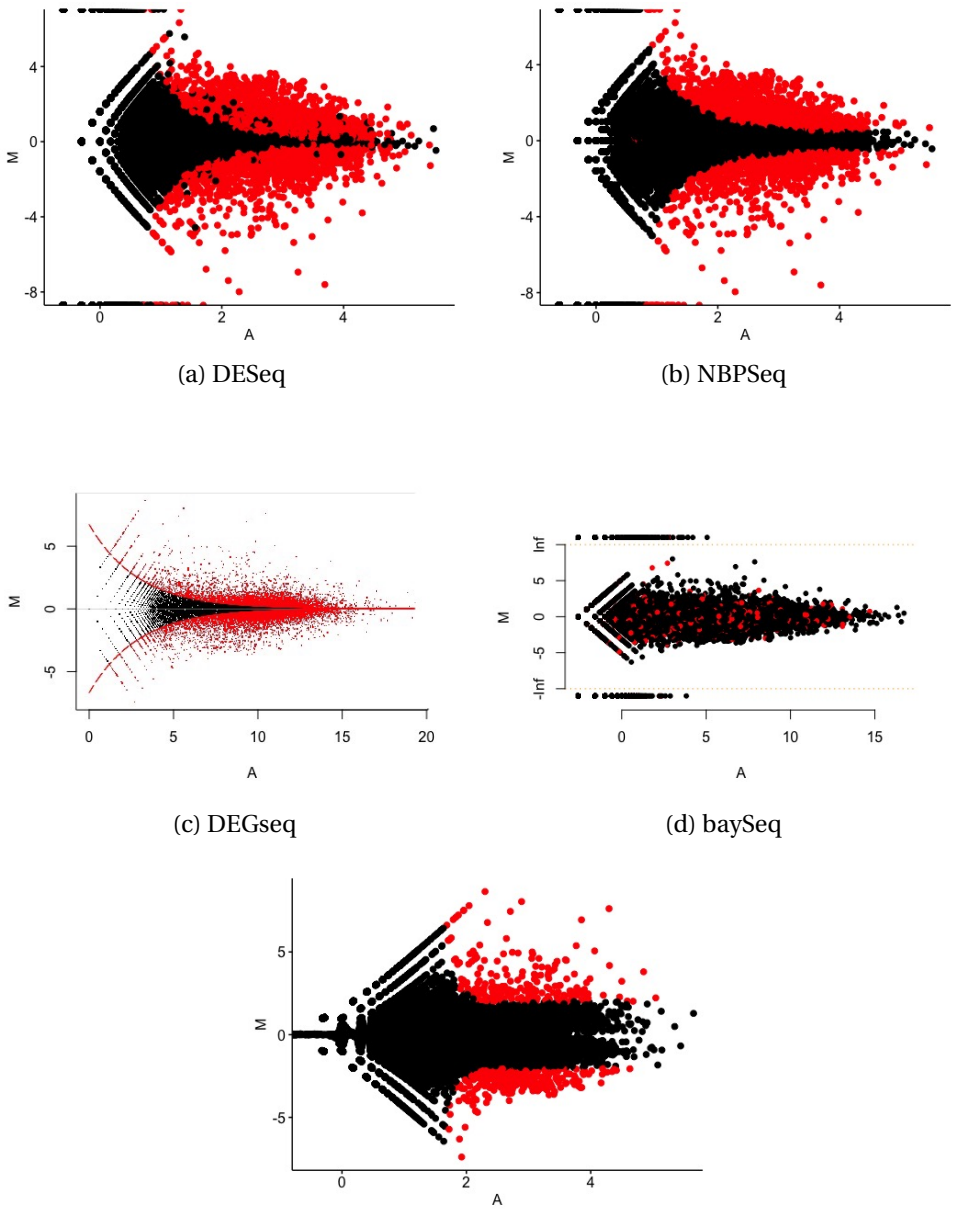
5.3.2 RNA-seq

We demonstrated MA plots from all analytical methods for RNA-seq in Figure 5.10. The results show similar patterns to microarray, that DESeq is highly similar to NBPSeq while baySeq and NOIseq show a somehow different distribution of differentially expressed genes.

In Table 5.5 we show the Pearson correlation coefficients and Spearman's rank correlations between analytical methods. In Table 5.6 we list the mutual information between statistical models. From the tables we can see similar correlation patterns from microarray.

Table 5.5. RNA-seq Pearson correlation coefficient / Spearman's rank correlation between analytical methods.

	NBPSeq	DEGseq	baySeq	NOIseq
DESeq	0.54/0.53	0.86/0.82	0.88/0.85	0.91/0.93
NBPSeq		0.79/0.68	0.57/0.48	0.56/0.55
DEGseq			0.79/0.63	0.72/0.72
baySeq				0.79/0.74



(a) DESeq

(b) NBPSeq

(c) DEGseq

(d) baySeq

(e) NOIseq

Figure 5.10. M(log ratio)A(mean average)plots for RNA-seq from analytical methods: (a) DESeq (b)NBPSeq (c) DEGseq (d) baySeq (e) NOIseq. Red dots are detected differentially expressed genes. Threshold for p-value <0.05, likelihood > 0.9

Table 5.6. RNA-seq mutual information between statistical models

	Poisson	Bayes	Non-parametric
Negative Binomial	0.46	0.75	0.11
Poisson		0.55	0.15
Bayes			0.10

5.4 Consistency between Data Types

Pearson correlation coefficient and Spearman's rank correlation should be the most ideal tools to measure correlation and check consistency between data types. But as mentioned before, probes in microarray can sometimes be mapped to multiple genes. When performing correlation analysis, there will be duplicates if those pairs are not eliminated; otherwise there will be lost information if those pairs are removed. So we only show mutual information between microarray and RNA-seq for all analytical methods in Table 5.7. From the table we can see that, similar from Table 5.1, microarray and RNA-seq show highest consistency in NOIseq and lowest in DESeq. The reason why DESeq is the worst is that DESeq detects much more genes for RNA-seq than for microarray, making the intersection proportion too small.

Table 5.7. Mutual information between data types

	DESeq	NBPSeq	DEGseq	baySeq	NOIseq	Sum
Mutual information	0.04	0.19	0.09	0.07	0.22	0.09

We also did gene set enrichment analysis on the detected differentially expressed genes using PANTHER on GO. We hope to see biological meaning revealed from the results. Here we present protein classes hit by microarray and RNA-seq results from analytical method DESeq in Figure 5.11 and results from all analytical methods in Figure 5.12. The figure panels (a) are protein class hit by microarray, panels (c) are protein class hit by RNA-seq and panels (b) and (d) are their category annotations. For both

data types, both protein classes "Nucleic acid binding" and "Hydrolase" have high hit numbers, protein classes "Cytoskeletal protein", "Enzyme modulator", "Transcription factor" and "Transferase" have hit numbers outstand others. One huge difference is the hit number for "Signaling molecule", the hit number detected is large for microarray but only above average for RNA-seq. It's obvious that despite some differences, microarray and RNA-seq are consistent in this aspect.

In Figure 5.13, we also compared the percentage of genes hitting each protein classes from microarray and RNA-seq for all analytical methods. Other than plotting hit numbers directly, we plotted the percentages. This figure demonstrated additional information of proportions rather than absolute values. Clearly, in both aspects, microarray and RNA-seq showed fair consistency. If we look into the results, we can see highly consistent results from Figure 5.12, that protein classes "Nucleic acid binding" and "Hydrolase" have rather high hit percentage and protein classes "Cytoskeletal protein", "Enzyme modulator", "Transcription factor" and "Transferase" have hit percentage higher than average. Notably, the huge difference mentioned before in protein class "Signaling molecule" remains in every analytical method except NOISeq. Possible reason is that NOISeq detects few differentially expressed genes, making the difference in percentage not that obvious between data types.

The total enrichment score was calculated and the most significant groups are plotted in Figure 5.14. The enrichment analysis has similar findings to genes hit per protein class, that "Nucleic acid binding" and "Signaling molecule" are still significant groups. Therefore, enrichment analysis is showing consistent results. Other results from other analytical methods are listed in the supplementary files.

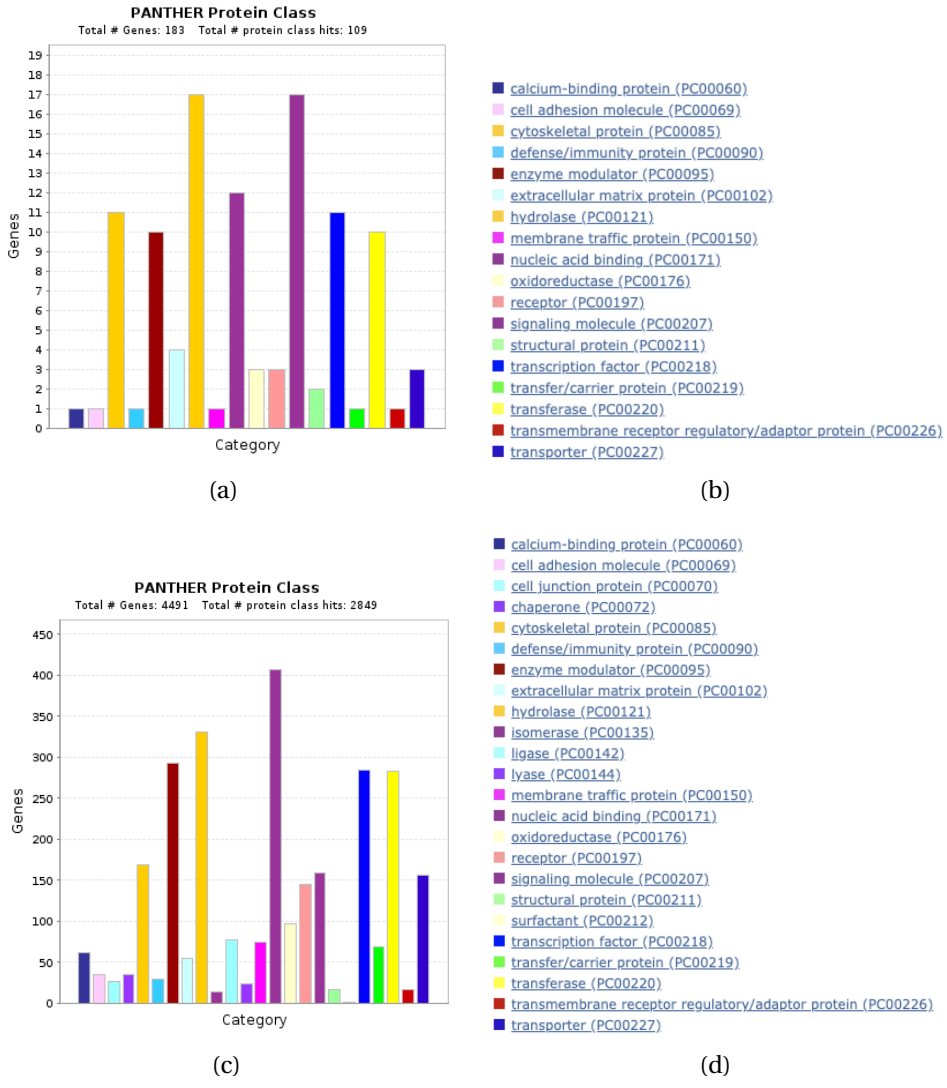


Figure 5.11. Protein class hit by microarray and RNA-seq results from DESeq package (a) Microarray results (b) categories (c) RNA-seq results (d) categories

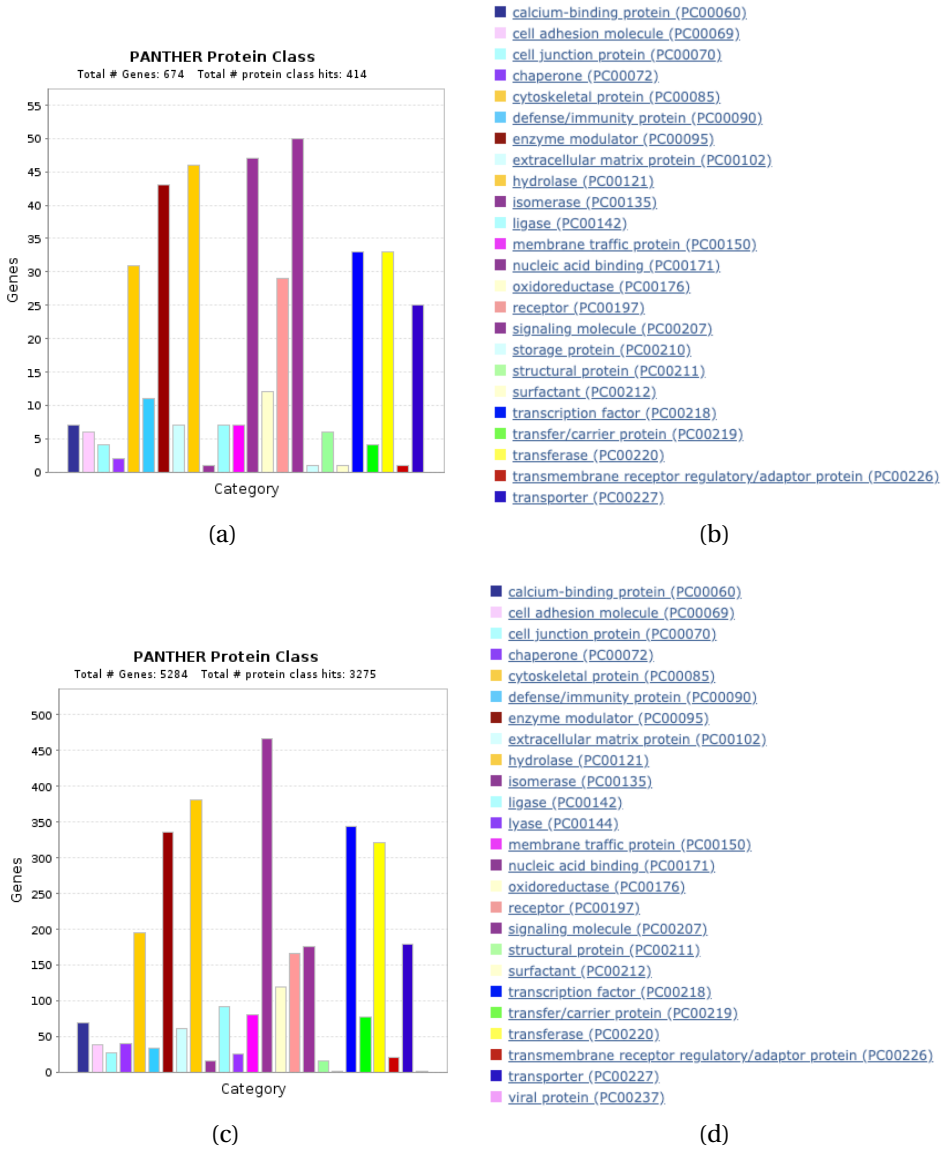


Figure 5.12. Protein class hit by microarray and RNA-seq results from all package (a) Microarray results (b) categories (c) RNA-seq results (d) categories

(a) (b)

(c) (d)

(e) (f)

Figure 5.13. Protein class hit percent comparison of microarray and RNA-seq (a) DESeq (b) DEGseq (c) NBPSeq (d) NOISeq (e) baySeq (f) all packages

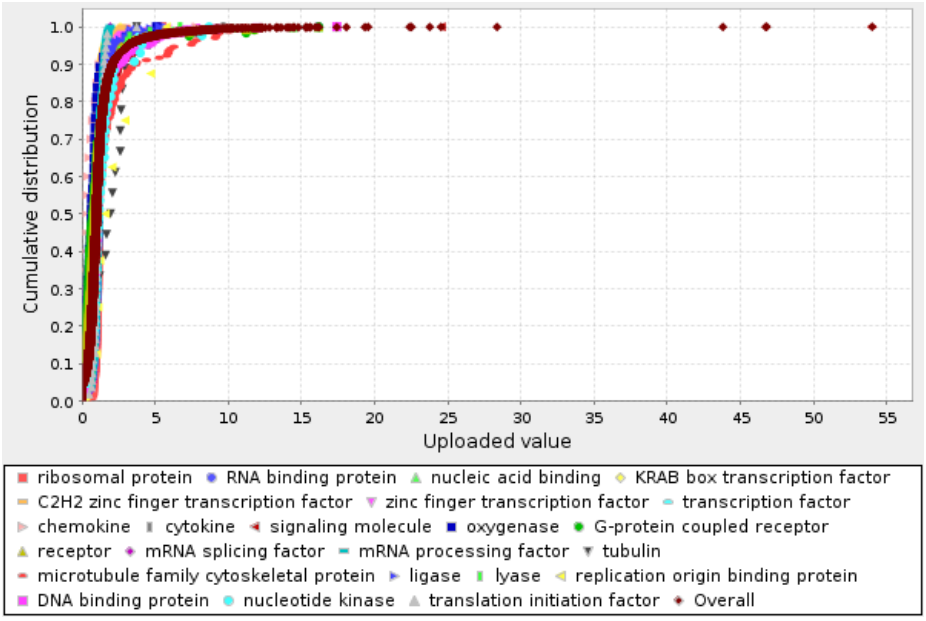


Figure 5.14. Enrichment score for the most significant groups

6 Discussion

In this chapter we will thoroughly discuss the results outlined in the previous chapter from the comparison of several statistical approaches, data methods and a combination of these factors. The discussion will include potential reasons for the observed results, the strengths and weaknesses of our approach and the future directions.

6.1 Discussion over Results

6.1.1 Consistency of Analytical Methods across a Statistical Model

Briefly, from the results we can see that, analytical methods applying the same statistical model show high consistency for both data types. Both DESeq and NBPSeq methods use negative binomial distribution model. From what we can see in chapter Results, DESeq and NBPseq show similar patterns in data distribution, data dispersion and variation (See Figure 5.2). Their measured gene expression levels and p-values are highly correlated (See Figure 5.6). To further interpret the similarity, we calculated and plotted Pearson and Spearman's correlation between DESeq and NBPSeq. The correlation coefficients are pretty high which means the results from the two analytical methods are highly correlated, or say consistent (See Figure 5.7). Microarray and RNA-seq share all these findings above, so overall, high performance consistency can be found in DESeq

and NBPSeq for both microarray and RNA-seq. It's probably because that both DESeq and NBPSeq use a negative binomial distribution model, results from these two analytical methods are very similar in every aspect.

6.1.2 Consistency among Statistical Models

Consistency across statistical models or analytical methods both depends on which two models or methods are compared. Some pairs show high consistency while others don't. Specifically, see Table 5.1, DEGseq gives the most detected genes (534) for microarray and baySeq (4845) for RNA-seq while NOIseq gives the least detected genes for both data types (187 and 523). In general, RNA-seq detects more differentially expressed genes than microarray. Work by Bullard et al[20] has evaluated various statistics for differential expression and find that the main difference between test statistics is their ability to handle low counts. So, we assume RNA-seq has stronger ability than microarray in this aspect.

Negative binomial model has very high consistency with Bayesian model, they detect the most differentially expressed genes and share most of their findings. The potential reason is that baySeq, the method applying Bayesian model, runs a binomial test, which may lead to similar results to methods applying negative binomial distributions. NOIseq, method using non-parametric model, detected the least differentially expressed genes but most of its findings can be found in other models. This indicates that all methods are consistent to others to some extent regardless of the models they are applied to.

6.1.3 Consistency between Data Types

Between data types, the results are generally consistent despite RNA-seq detects more differentially expressed genes than microarray assays. All statistical models for both data types give a similar set of detected differentially expressed genes. Analytical methods NBPSeq and NOIseq are the most consistent for microarray while DESeq and baySeq are the most consistent for RNA-seq. Microarray and RNA-seq have the best consistency in NOIseq and the worst in DESeq. Potential reason one is that although non-parametric model yields the least detected differentially expressed genes, the detected ones are so significantly differentially expressed that they can hardly be missed by any data type. Potential reason two is that non-parametric model depends the least on data distribution, since using specific distribution in the statistical model will reduce degrees of freedom. Overall, RNA-seq gives more detected differentially expressed genes. As mentioned above, it's probably because RNA-seq does a better job when dealing with low counts genes.

Analyzing the differential expression results, the detected differentially expressed genes, shows biological meaning. For both microarray and RNA-seq, genes from protein classes Nucleic acid binding and Hydrolase are significantly differentially expressed, genes from protein classes Cytoskeletal protein, Enzyme modulator, Transcription factor and Transferase are highly differentially expressed. One huge difference is that genes from Signaling molecule protein class is detected much more differentially expressed by microarray than by RNA-seq (See Figure 5.12). This could due to microarray probe design, RNA-seq read mapping process or other molecular level differences caused by different treatments.

6.2 Study Design

Our experiment design aims at studies at three levels: analytical methods, statistical models and data types, all results are based on differential expression analysis. Evaluating consistency across three aspects is the biggest strength of our study design. This study reveals what role analytical method, statistical model and data type plays in differential analyses.

The weakness of our study design is that we lack a "golden standard". Without a golden standard, we cannot calculate accuracy or false discovery rate for any analytical method, which could have provided more information on differential analysis performance. We are limited to consistency between analytical methods or statistical models since we cannot identify how many findings are accurate, but only how many are in common.

6.3 Future Directions

There are many potential ways to improve performances in many aspects. For example, since we see that RNA-seq may have stronger ability in handling low count genes, to improve microarray's performance, studies could focus on low count genes.

Despite the consistency observed in previous experiments, different statistical models give different analytical results. A better fitting model for both data types may improve differential expression analysis performance. With strong statistical skills, proposing a new statistical model that better fits data distribution may better serve this purpose.

Biological experiments can be conducted to verify whether the detected differentially expressed genes are truly meaningful. Furthermore, those truly differentially expressed genes can serve as a golden standard to assess accuracy for each analytical method and statistical model.

7 Conclusions

From the experiments we can draw the following conclusions. First, analytical methods applying the same statistical model are highly consistent in every inspected aspect. Second, all statistical models are consistent to others to some extent, varying based on specific pairs chosen to be compared. Lastly, differential expression analysis results are generally consistent despite RNA-seq detects more genes and there are some differences in gene set enrichment analysis. Overall, the statistical model and data type both impact greatly on differential analysis results while analytical method seems trivial.

Complete References

[1] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. Nature reviews genetics, 10(1):57, 2009.

[2] Radmila Hrdlickova, Masoud Toloue, and Bin Tian. Rna-seq methods for transcriptome analysis. Wiley Interdisciplinary Reviews: RNA, 8(1):e1364, 2017.

[3] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. Science, 270(5235):467–470, 1995.

[4] Kimberly R Kukurba and Stephen B Montgomery. Rna sequencing and analysis. Cold Spring Harbor Protocols, 2015(11):pdb–top084970, 2015.

[5] Elmar Pruesse, Christian Quast, Katrin Knittel, Bernhard M Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. Nucleic acids research, 35(21):7188–7196, 2007.

[6] Huei-Chung Huang, Yi Niu, and Li-Xuan Qin. Differential expression analysis for rna-seq: an overview of statistical methods and computational software: supplementary issue: sequencing platform modeling and analysis. Cancer informatics, 14:CIN–S21631, 2015.

[7] Intawat Nookaew, Marta Papini, Natapol Pornputtapong, Gionata Scalcinati, Linn Fagerberg, Matthias Uhlén, and Jens Nielsen. A comprehensive comparison of rna-seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in saccharomyces cerevisiae. Nucleic acids research, 40(20):10084–10097, 2012.

[8] Xing Fu, Ning Fu, Song Guo, Zheng Yan, Ying Xu, Hao Hu, Corinna Menzel, Wei Chen, Yixue Li, Rong Zeng, et al. Estimating accuracy of rna-seq and microarrays with proteomics. BMC genomics, 10(1):161, 2009.

[9] Manuel Garber, Manfred G Grabherr, Mitchell Guttman, and Cole Trapnell. Computational methods for transcriptome annotation and quantification using rna-seq. Nature methods, 8(6):469, 2011.

[10] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. Nature biotechnology, 31(1):46, 2013.

[11] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1):139–140, 2010.

[12] Juliana Costa-Silva, Douglas Domingues, and Fabricio Martins Lopes. Rna-seq differential expression analysis: An extended review and a software tool. PloS one, 12(12):e0190152, 2017.

[13] Kirk J Mantione, Richard M Kream, Hana Kuzelova, Radek Ptacek, Jiri Raboch, Joshua M Samuel, and George B Stefano. Comparing bioinformatic gene expression profiling methods: microarray and rna-seq. Medical science monitor basic research, 20:138, 2014.

[14] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. PloS one, 9(1):e78644, 2014.

[15] Daniel Castillo, Juan Manuel Gálvez, Luis Javier Herrera, Belén San Román, Fernando Rojas, and Ignacio Rojas. Integration of rna-seq data with heterogeneous microarray data for breast cancer profiling. BMC bioinformatics, 18(1):506, 2017.

[16] M Madan Babu. Introduction to microarray data analysis. Computational genomics: Theory and application, 225:249, 2004.

[17] Jun Lu, John K Tomfohr, and Thomas B Kepler. Identifying differential expression in multiple sage libraries: an overdispersed log-linear model approach. BMC bioinformatics, 6(1):165, 2005.

[18] Frank Avery Haight. Handbook of the poisson distribution. 1967.

[19] David P Kreil, Natasha A Karp, and Kathryn S Lilley. Dna microarray normalization methods can remove bias from differential protein expression analysis of 2d difference gel electrophoresis results. Bioinformatics, 20(13):2026–2034, 2004.

[20] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. BMC bioinformatics, 11(1):94, 2010.

[21] Elie Maza, Pierre Frasse, Pavel Senin, Mondher Bouzayen, and Mohamed Zouine. Comparison of normalization methods for differential gene expression analysis in rna-seq experiments: a matter of relative size of studied transcriptomes. Communicative & integrative biology, 6(6):e25849, 2013.

[22] Daehwan Kim, Ben Langmead, and Steven L Salzberg. Hisat: a fast spliced aligner with low memory requirements. Nature methods, 12(4):357, 2015.

[23] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. bioinformatics, 25(14):1754–1760, 2009.

[24] Jouni Sirén, Niko Välimäki, and Veli Mäkinen. Indexing graphs for path queries with applications in genome research. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 11(2):375–388, 2014.

[25] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. Bioinformatics, 25(9):1105–1111, 2009.

[26] Likun Wang, Zhixing Feng, Xi Wang, Xiaowo Wang, and Xuegong Zhang. Degseq: an r package for identifying differentially expressed genes from rna-seq data. Bioinformatics, 26(1):136–138, 2009.

[27] Yanming Di, Daniel W Schafer, and Maintainer Yanming Di. Package 'nbpseq'. Molecular Biology, 10:1, 2014.

[28] Yihui Zhou and Fred A Wright. Bbseq: A method to handle rna-seq count data. 2011.

[29] Thomas J Hardcastle and Krystyna A Kelly. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. BMC bioinformatics, 11(1):422, 2010.

[30] Simon Anders and Wolfgang Huber. Differential expression of rna-seq data at the gene level–the deseq package. Heidelberg, Germany: European Molecular Biology Laboratory (EMBL), 2012.

[31] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. Biostatistics, 9(2):321–332, 2007.

[32] Alexander Sturn, John Quackenbush, and Zlatko Trajanoski. Genesis: cluster analysis of microarray data. Bioinformatics, 18(1):207–208, 2002.

[33] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3):37–52, 1987.

[34] Leping Li, Clarice R Weinberg, Thomas A Darden, and Lee G Pedersen. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method. Bioinformatics, 17(12):1131–1142, 2001.

[35] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based

approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43):15545–15550, 2005.

[36] Huaiyu Mi, Anushya Muruganujan, Dustin Ebert, Xiaosong Huang, and Paul D Thomas. Panther version 14: more genomes, a new panther go-slim and improvements in enrichment analysis tools. Nucleic acids research, 47(D1):D419–D426, 2018.

[37] Amir Ben-Dor, Ron Shamir, and Zohar Yakhini. Clustering gene expression patterns. Journal of computational biology, 6(3-4):281–297, 1999.

[38] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. Nature protocols, 7(3):562, 2012.

[39] Stephane Joost, Michael Kalbermatten, and AurELie Bonin. Spatial analysis method (sam): a software tool combining molecular and environmental data to identify candidate loci for selection. Molecular Ecology Resources, 8(5):957–960, 2008.

[40] Cole Trapnell et al. Cuffdiff (v7).

[41] Damian Smedley, Syed Haider, Steffen Durinck, Luca Pandini, Paolo Provero, James Allen, Olivier Arnaiz, Mohammad Hamza Awedh, Richard Baldock, Giulia Barbiera, et al. The biomart community portal: an innovative alternative to large, centralized data repositories. Nucleic acids research, 43(W1):W589–W598, 2015.

[42] Bronwen L Aken, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos García Girón, Thibaut Hourlier, et al. The ensembl gene annotation system. Database, 2016, 2016.