ESTIMATING THE NUMBER OF GENES THAT ARE DIFFERENTIALLY EXPRESSED IN TWO DEPENDENT EXPERIMENTS OR ANALYSES

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Hammed Oluwashola Lawal

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

April 2022

Fargo, North Dakota

North Dakota State University
Graduate School

Title

ESTIMATING THE NUMBER OF GENES THAT ARE DIFFERENTIALLY EXPRESSED IN TWO DEPENDENT EXPERIMENTS OR ANALYSES

By

Hammed Oluwashola Lawal

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Megan Orr
Chair

Dr. Bong-Jin Choi

Dr. Changhui Yan

Approved:

April 13, 2022
Date

Dr. Magel Rhonda
Department Chair

ABSTRACT

Many researchers have used the intersection method to compare the results of differential expression analysis between two or more gene expression experiments. Some methods have been proposed to estimate the number of genes commonly differentially expressed in two independent gene expression experiments or analyses, but there has not been a method for estimating this number using dependent experiments or analyses other than the intersection method. In this thesis project, we propose a method for estimating the number of differentially expressed genes in two dependent experiments or analyses. Simulation studies are performed to compare the proposed to existing methods and an analysis of a real gene expression data set is performed to illustrate the use of the proposed method.

ACKNOWLEDGMENTS

I want to thank God almighty for the success of this program. I also want to sincerely thank my supervisor, Dr. Orr, for her efforts, guidance, support, and encouragement throughout this thesis project. In addition, I want to thank Dr. Choi and Dr. Yan for their willingness to support this project.

DEDICATION

I dedicate this thesis project to my late parents. Mum, Dad, I am a step closer! Continue to rest in power.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1: INTRODUCTION

1.1. Background Study

1.1.1. Gene Expression

Gene expression can be described as the process by which information from a gene is used in the synthesis of a functional gene product. Recent advances in technologies have contributed significantly to the understanding of the regulation of gene expression, and this regulation is done at the transcriptional level. A common purpose of these technologies is to quantify gene expression by measuring the abundance of mRNA in a sample organism. Some common technologies such as real-time PCR, microarray analysis, next-generation sequencing, and RNA-seq used in functional genomics depend on the scale and intent of the experiment.

1.1.2. Microarray Technology

Microarray technologies have become more relevant tools since their development in the 1990s for clinical research purposes. Microarray technologies are easier to use compared to several gene expressions profiling methods such as differential display and serial analysis of gene expression. They do not require large-scale DNA sequencing and allow the parallel quantification of thousands of genes from multiple samples (Russo, Zegar, and Giordano, 2003). There have been several research studies conducted using these technologies; for instance, toxicologists can now define specific patterns of gene expression under a given set of experimental conditions and provide a mechanistic rationale for such changes (Shankar and Mehendale, 2014); they have been used to analyze normal and cancerous tissues and cell lines (Bull et al., 2001). Microarrays technologies are also used for several other purposes; for example, they are efficient in the interrogation of chromosome structure and integrity; and are of two types, namely, comparative genomic hybridization (CGH) arrays and single nucleotide polymorphism (SNP) – based arrays (Martin, 2020).

1.1.3. RNA-seq Technology

RNA sequencing technologies are high-throughput transcriptome profiling technologies used to directly determine the cDNA sequence. There have been many developments in sequencing technologies over time, starting with the Double helix structure in 1953. Sanger sequencing was developed in 1977 and is referred to as the first-generation sequencing technology. The first high-throughput sequencing platform was developed in 2005, followed by many next-generation sequencing platforms (NGS). The comparison between NGS platforms' accuracy and reproducibility are measured using several factors such as the features and their corresponding analysis pipelines. NGS can detect unknown gene expression sequences as compared to microarrays but is time-consuming (Hong et al., 2020).

1.2. Differential Gene Expression Analysis

A common goal of gene expression experiments is to identify genes that are differentially expressed (i.e., genes whose mean expression values differ between two groups). In microarray experiments, expression levels are considered continuous measurements, and statistical methods assuming normality of gene expressions are regularly employed. In RNA-seq experiments, expression measurements are discrete counts, and often, the assumption is that the expression measurements follow negative binomial distributions. Because gene expression analysis involves performing inferences on thousands of genes simultaneously and could lead to multiple testing errors, there is a need to control the number of false-positive results, i.e., the number of equivalently expressed genes identified as differentially expressed. The False Discovery Rate

(FDR), proposed by Benjamini and Hochberg (1995), is a popular method for controlling multiple testing errors in gene expression experiments.

1.3. Intersection Method

It is common for researchers to compare the results of two differential expression analyses, and in many cases, a list of genes identified as commonly differentially expressed is determined for each analysis. The number of genes common to both lists is often reported using the Venn diagram graphical approach. We would refer to this method as the "intersection method". Using the intersection method, the number of genes identified as differentially expressed in both analyses or experiments relies heavily on the level at which the FDR is controlled. This is a potential drawback if researchers are interested in evaluating the degree of differential expression common to both analyses or experiments.

1.4. Estimating Differentially Expressed Genes in Two Experiments or Analyses

The intersection method allows us to determine the genes that are common to both lists of two differentially expressed genes in gene expression experiments when the FDR is controlled at a nominal level α. The results of the intersection method, or the number of genes declared to be differentially expressed in two experiments rely heavily on the value of α. The number of commonly differentially expressed genes in the two experiments increases as α increases.

(Orr et al., 2012) proposed a method to estimate the number of genes that are differentially expressed in two experiments. This method uses the p-values from the two experiments simultaneously, to produce an estimate that does not depend on the False Discovery Rate (FDR).

1.5. Research Objectives

Many researchers have used the intersection method to perform statistical analysis on gene expression experiments between two or more groups or two or more methods. Some methods have been proposed to estimate the number of genes commonly differentially expressed in two independent experiments/analyses. To our knowledge, there has not been a method for estimating this number using dependent experiments or analyses other than the intersection method. The goal of this research is to propose a method for estimating the number of differentially expressed genes in two dependent experiments or analyses. We would perform simulation studies and compare the proposed method to the existing ones and illustrate the use of the proposed method on real gene expression data sets.

1.6. Organization

The rest of this work is organized as follows; In Chapter 2 – we perform a literature review of topics related to this project. Chapter 3 describes the proposed methods, simulations studies, and real data analysis; In chapter 4, we present the results of the simulation study and real data analysis using the methods in chapter 3 and finally, we present our overall findings and conclusions of the analysis as well as recommendations for future work in chapter 5.

CHAPTER 2: LITERATURE REVIEW

2.1. Estimating the Number of Equivalently Expressed Genes in a Single Experiment Using p-values

Consider the problem of simultaneously testing null hypotheses $H_{i1}, \dots, H_{im}$ for experiment i based on the corresponding p-values $p_{i1}, \dots, p_{im}$. For $j = 1, \dots, m$, we assume that $p_{ij} \sim Uniform(0,1))$ when H_{ij} is true and that p_{ij} has a distribution that is stochastically smaller than the uniform distribution when H_{ij} is false. These are standard assumptions which imply that an unbiased size α test can be obtained for each gene j by rejecting H_{ij} if and only if $p_{ij} \leq \alpha$.

(Storey and Tibshirani, 2003) showed that, for any fixed $\lambda_i \in [0, 1)$ in experiment i,

$$m_0^{(i)}(\lambda_i) = \frac{\sum_{j=1}^{m} \mathbf{1}\{p_{ij} > \lambda_i\}}{1 - \lambda_i} \tag{2.1}$$

is an estimator of $m_0^{(i)}$, the number of true null hypotheses among $H_{i1}, \dots, H_{im}$. As illustrated by (Storey and Tibshirani, 2003), to demonstrate the method for estimating $m_0^{(i)}$, the histogram-based method proposed by (Liang and Nettleton, 2012) was used. This method selects a value of λ_i from a set of candidate values so that a histogram of p-values less than λ_i is approximately decreasing while a histogram of p-values greater than λ_i is approximately uniform.

The algorithm of Liang and Nettleton (2012) can be described as follows:

1. Partition the interval [0,1] into B bins of equal width. Let $c_b = \left(\frac{b-1}{B}, \frac{b}{B}\right]$ $for\ b = 1, 2, \dots, B$.
2. Denote the number of p-values in the interval $c_b\ as\ n_b\ for\ b = 1,2, \dots, B$.
3. For each $b = 1, 2, \dots, B$, calculate

$$\bar{n}_b = \frac{\sum_{k=b}^{B} n_b}{B - b + 1}.$$

4. Let $b^* = \min\{min\{b\colon nb \leq \bar{n}b\}, B - 1\}$Select $\lambda_i = \frac{b^*}{B}$.

We use $B = 20$ when applying this algorithm according to the recommendations by (Nettleton et al., 2006) and (Liang and Nettleton, 2012).

There have been many other methods proposed to estimate the number of equivalently expressed genes and the number of differentially expressed genes when performing a hypothesis test for each gene in one gene expression data set, (Storey, 2002), (Storey and Tibshirani, 2003), (Storey, Taylor, and Siegmund, 2004), (Langaas, Ferkingstad, and Lindqvist, 2005), (Nettleton et al., 2006), and (Liang and Nettleton, 2012) amongst others.

2.2. Estimating the Number of Equivalently Expressed Genes in Two Experiments Using p-values

Now considering the problem of testing m pairs of null hypotheses $(H_{11}, H_{21}), (H_{12}, H_{22}), \dots, (H_{1m}, H_{2m})$, where H_{ij} is the null hypothesis for experiment (or analysis) i $(i = 1,2)$ and gene j $(j = 1, 2, .., m)$. Each hypothesis H_{ij} is either true (gene j in experiment i is equivalently expressed), or false (gene j in experiment i is differently expressed).

Table 2.1: Contingency table of the estimate of equivalently expressed and differentially expressed genes for each of the m genes in each experiment described by (Orr et al., 2012).

		Experiment 2		
		Gene EE	Gene DE	Total
Experiment 1	Gene EE	m_{00}	m_{01}	$m_0^{(1)}$
	Gene DE	m_{10}	m_{11}	$m_1^{(1)}$
	Total	$m_0^{(2)}$	$m_1^{(2)}$	m

In table 2.1., m_{00} represents the number of equivalently expressed (EE) genes in both experiments; m_{11}, the number of differentially expressed (DE) genes in both experiments; m_{01}, the number of EE genes in experiment 1 but DE in experiment 2; m_{10} is the number of genes that are DE in experiment 1 but EE in experiment 2; $m_0^{(i)}, i = 1,2$ is the total number of EE genes in experiments 1 and 2; $m_1^{(i)}, i = 1,2$ is the total number of DE genes in experiments 1 and 2; and m is the overall total number of genes in both experiments.

Following the setup of (Orr et al., 2012), the histogram-based method was used to estimate the number of equivalently expressed genes in each experiment. For p-value pairs (p_{1j}, p_{2j}) corresponding to gene j, if both null hypotheses in the pair (H_{1j}, H_{2j}) are true, then it is assumed that p_{1j} and p_{2j} are independent and both follow a Uniform$(0,1)$ distribution.

From this assumption,

$$Pr\left((p_{1j}, p_{2j}) \in [\lambda_1, 1] \times [\lambda_2, 1]\right) = (1 - \lambda_1)(1 - \lambda_2) \tag{2.2}$$

Now, let n_{00} be the number of p-values pairs that fall into the upper right quadrant defined by λ_1 and λ_2, i.e.,

$$n_{00} = \sum_{i=1}^{m} \mathbf{1}\{(p_{1j}, p_{2j}) \in [\lambda_1, 1] \times [\lambda_2, 1]\} \tag{2.3}$$

Thus, a conservative estimate of m_{00}, the number of genes that are EE in both experiments, is given as

$$\widehat{m}_{00} = \frac{n_{00}}{(1 - \lambda_1)(1 - \lambda_2)} \tag{2.4}$$

2.3. Estimating the Number of Differentially Expressed Genes in Two Experiments Using p-values

To estimate m_{11}, we need to first estimate $m_0^{(i)}$, the number of EE genes for a single expression experiment i $(i = 1,2)$ using the method proposed by (Storey and Tibshirani, 2003; Liang, and Nettleton, 2012). The p-values from both experiments are paired for each gene j to estimate m_{00} using the methods described by (Orr et al., 2012; Lai, 2007) but these methods assume that the experiments are independent.

Following the work of (Orr et al., 2012) in obtaining p-value pairs, we used a truncated bivariate normal distribution method in estimating m_{00} for two dependent experiments, and then obtain m_{11} by combining $m_0^{(i)}$ and m_{00} which can be expressed as

$$m_{11} = m - m_0^{(1)} - m_0^{(2)} + m_{00} \tag{2.5}$$

We estimate $m_0^{(i)}$ for experiment 1 and 2 using the previously proposed methods, and then estimate m_{00}, which we propose later in this section, and finally, we estimate m_{11} which is our goal, and is given as:

$$\widehat{m}_{11} = m - \widehat{m}_0^{(1)} - \widehat{m}_0^{(2)} + \widehat{m}_{00} \tag{2.6}$$

2.4. Estimating the Number of Differentially Expressed Genes in Two Experiments Using the Intersection Method

Several research studies on gene expression experiments have been conducted to identify genes that are differentially expressed in multiple experiments or analyses using the intersection method. (Miyama and Hanagata, 2006) analyzed gene expression profiles in salt-stressed burma mangrove to identify genes that are likely to be of important to salt tolerance in burma mangrove. They used Venn diagrams to visualize the results of the analysis that showed tissue-specific and overlapping between genes that were up-regulated more than five-fold and genes were suppressed less than one-fifth.

(Liu et al., 2020) performed a transcriptome analysis of schistosoma japonicum derived from SCID mice and BALB/c mice to identify differentially expressed genes using an intersection method. Significant differential expression between the two independent experiments was computed using false discovery rate (FDR) estimation with q values. They presented the results of the intersection method using a Venn diagram to reveal four DEGs that are present in both comparisons.

Other authors that used the intersection method include (Sandford et al., 2012; HAI-XIA GAO et al., 2021; Glen et al., 2015) amongst others.

2.5. Calculating p-values in Gene Expressions Experiments

We divide each experiment into two groups of equal samples (replicates), and then use student t-test and moderated t-test to perform hypothesis testing and obtain the p-values corresponding to gene j for each experiment.

2.5.1. Student's t-test

We wish to test

$$H_{0j}: \mu_{1j} = \mu_{2j} \ \ vs. \ H_{1j}: \mu_{1j} \neq \mu_{2j} \tag{2.7}$$

for each $gene\ j = 1,2, \dots m$. The test statistic for each gene j is given as

$$t_j = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{\sqrt{s_j^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{2.8}$$

where $\bar{y}_{ij}$ is the sample mean of treatment $i\ (i = 1, 2)$ observations for gene j, s_j^2 is the estimate of σ_j^2, and the pooled sample variance is

$$s_j^2 = \frac{(n_1 - 1)s_{1j}^2 + (n_2 - 1)s_{2j}^2}{(n_1 - 1) + (n_2 - 1)} \tag{2.9}$$

s_{ij}^2 is the sample variance of treatment $i\ (i = 1, 2)$ observations for gene j.

If H_{0j} is true (i.e., whenever gene j is equivalently expressed for $j = 1, \dots, m$), t_j will have a *central* t-distribution with $d = n_1 + n_2 - 2$ degrees of freedom such that

$$t_j \sim t_{n_1+n_2-2}(0) \tag{2.10}$$

If H_{0j} is false (i.e., whenever gene j is differentially expressed), t_j will have a *non-central* t-distribution with $d = n_1 + n_2 - 2$ degrees of freedom and non-centrality parameter such that

$$t_j = \frac{\mu_{1j} - \mu_{2j}}{\sqrt{\sigma_j^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{2.11}$$

2.5.2. Moderated t-test

(Smith, 2004) proposed that the estimator of the variance for gene j is given as

$$\widetilde{s_j}^2 = \frac{d s_j^2 + d_0 s_0^2}{d + d_0} \tag{2.12}$$

such that for each gene j, $\widetilde{s_j}^2$ is a weighted average of the prior variance (s_0^2) and the sample variance of the jth gene $\left(s_j^2\right)$. The weights are the prior degrees of freedom (d_0) and the standard degrees of freedom for a pooled two-sample t-test (d). It is stated that this method shrinks the individual estimate s_j^2 towards s_0^2.

Thus, using $\widetilde{s_j}^2$, the test statistic for the hypothesis testing is given as

$$\widetilde{t_j} = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{\sqrt{\widetilde{s_j}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{2.13}$$

where $\widetilde{t_j}$ is referred to as the *moderated t-statistic* and can be shown that

$$\widetilde{t_j} = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{\sqrt{\widetilde{s_j}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{d+d_0}(0) \tag{2.14}$$

when H_{0j} is true (i.e., whenever gene j is equivalently expressed), and d_0 and s_0^2 are known.

2.6. Multiple Testing

2.6.1. Type I and II Errors

When performing a hypothesis test using gene expression data, a Type I error (false positive) is committed when a gene is declared to be differentially expressed when it is truly equivalently expressed, and a Type II error (false negative) is committed when the test fails to identify a gene that is truly differentially expressed. However, when we perform multiple hypothesis testing and the Type I error rate is controlled at α for each test, the probability of committing one or more Type I errors increases from α as the number of tests increases. The Family-Wise Error Rate (FWER) is a common method for controlling multiple testing error rate defined as the probability of committing at least one Type I error in a family of tests. The common procedures for controlling the FWER are the Bonferroni (Simes, 1986) and Holm (Holm, 1979) methods. However, when FWER is used in gene expression experiments in which thousands of hypotheses are being tested simultaneously, the FWER generally results in extremely low statistical power for identifying DE genes. False Discovery Rate (FDR) introduced by (Benjamini and Hochberg, 1995) is a more powerful alternative to FWER.

Table 2.2. summarizes the hypothesis testing errors and the decision when null hypothesis is true or false. $\alpha = P(Type\ I\ error)$, the probability of committing an error of rejecting the null hypothesis when it is actually true, while $\beta = P(Type\ II\ error)$, the probability of committing an error of failing to reject the null hypothesis when it is actually false.

Table 2.2: Summary of hypothesis testing errors.

Decision	$H_0\ True$	$H_0\ False$
Do Not Reject H_0	Correct Decision $1 - \alpha$	Incorrect Decision Type II Error β
Reject H_0	Incorrect Decision Type I Error α	Correct Decision $1 - \beta$

2.6.2. Obtaining p-values from Gene Expression Data

2.6.2.1. edgeR

edgeR is a Bioconductor package for differential expression analyses of RNA-Seq read count data with biological replication. It is an implementation of statistical methodologies by (Robinson and Smyth, 2007; 2008) based on the negative binomial distributions. Empirical Bayes methods are used to moderate the degree of overdispersion across transcripts which leads to improving the reliability of inference (Robinson et al., 2010). Like RNA-seq, edgeR can also be used for analysis of differential signal of some other types of genomic data that produce count data such as ChIP-seq, ATAC-seq, Bisulfite-seq, SAGE and CAGE (Chen et al.). It was initially developed for SAGE (serial analysis of gene expression).

2.6.2.2. DESeq2

DESeq2 is another method for differential analysis of count data, using shrinkage estimation for dispersions and fold changes to improve stability and interpretability of estimates. It is an improvement from DESeq and uses the same median of ratios normalization method. The key difference between the two versions is that DESeq2 tests for strength of differential expression rather than just its presence (Love at al., 2014). DESeq2 also shrinks log fold changes toward zero, with shrinkage being stronger for genes with lower read counts because ratios can be noisier for weakly expressed genes. It uses a Wald test for testing for differential expression.

2.6.2.3. voom (Limma)

Linear models for microarray (Limma) is a package available in R for analyzing differential expression of microarray and RNA-seq experiments. It can be used for pre-processing of a two-color microarray data by fitting a linear model for each gene and then use an empirical Bayes approach to borrow information between genes to estimate gene-wise variance better (Ritchie et al., 2015). A moderated t-test is used for testing for differential expression by considering all genes when estimating the variance of each gene. This method makes inferences reliable due to increase of degrees of freedom regardless of small sample sizes. Because of its vast usage for different experimental designs, Limma is a popular choice amongst researchers for differential experiment analysis.

2.7. False Discovery Rate (FDR)

The false discovery rate (FDR) is a method of conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons. FDR-controlling procedures are designed to control the FDR, which is the expected proportion of "discoveries" (rejected null hypotheses) that are false (incorrect rejections of the null).

The FDR was developed by (Benjamini and Hochberg, 1995) and can be used as a more powerful alternative to family FWER. Consider the problem of testing m null hypotheses simultaneously, especially when m is large. m_0 is denoted as the number of true null hypotheses, **R** is the number of hypotheses rejected, **V** is the number of null hypotheses rejected from EE genes (false discoveries), and **S** is the number of null hypotheses rejected from DE genes (true discoveries). Table 2.3. summarizes notation for random variables associated with different scenarios in a multiple testing experiment. In table 2.3., $\boldsymbol{U}$ is the number of true non-discoveries (true negatives); V, the number false discoveries or false positives (Type I errors); m_0 represents the number of equivalently expressed genes; $\boldsymbol{T}$, the number of false non-discoveries or false negatives (Type II errors); S, the number of true discoveries or true positives; m_1, the number differentially expressed genes; $m - \boldsymbol{R}$, total number of non-discoveries or negatives; $\boldsymbol{R}$, total number of discoveries or positives; and m is the total number of tests/genes. m is known, but m_0 and $m_1 = m - m_0$ are unknown parameters.

Table 2.3: Random Variables Corresponding to the Number of Errors Committed when Testing m Hypothesis.

	Declared Non-significant	Declared Significant	Total
True Null Hypotheses	**U**	**V**	m_0
Non-True Null Hypothesis	**T**	**S**	m_1
	$m-\boldsymbol{R}$	**R**	m

Benjamini and Hochberg defined FDR as

$$FDR = E\left(\frac{V}{\max(R,1)}\right) = E\left(\frac{V}{R}\middle| R > 0\right) P(R > 0) \tag{2.15}$$

and proved by induction that the following procedure controls the FDR at level α when the *p*-values from true null hypotheses are independent and uniformly distributed.

2.7.1. Benjamini-Hochberg Procedure to Control the FDR

1. Specify k, the level at which to control FDR and compute the p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ for the m null hypothesis.
2. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered, observed p-values
3. Find $\hat{k}$, the largest k such that

$$p(k) \leq \alpha\frac{k}{m} \tag{2.16}$$

4. If $\hat{k}$ exists, then reject null hypotheses corresponding to $p_{(1)} \leq \cdots \leq p_{(\hat{k})}$. Otherwise, reject nothing.

CHAPTER 3: METHODS

3.1. Proposed Method for Estimating the Number of Equivalently Expressed Genes in Two Dependent Experiments

All methods reviewed in Chapter 2 are intended to analyze data from a single gene expression experiment or data from two independent experiments or analyses. Here, we propose a new method for analyzing gene expression data from two dependent experiments or analyses using p-values.

Again, consider the problem of testing m pairs of null hypotheses $(H_{11}, H_{21}), (H_{12}, H_{22}), \dots, (H_{1m}, H_{2m})$ with their corresponding p-value pairs $(p_{11}, p_{21}), (p_{12}, p_{22}), \dots, (p_{1m}, p_{2m})$. Within each pair, the p-values are assumed to be dependent. We propose applying the λ - estimator method and the histogram-based method to each set of p-values individually to obtain λ_1 and λ_2 values for each experiment (or analysis), and further consider the bivariate m^* p-values pairs such that $p_{1j} \geq \lambda_1$ and $p_{2j} \geq \lambda_2$. The m^* pairs of bivariate p-values are converted to z-values pairs $(z_{11}, z_{12}), (z_{12}, z_{22}), \dots, (z_{1m^*}, z_{2m^*})$, which are assumed to follow a truncated bivariate normal distribution with means $\mu_i = 0$ and variance $\sigma_i^2 = 1$ for $i = 1{,}2$. These means and variances follow from the assumption that, within each experiment or analysis, the p-values from equivalently expressed genes follow a Uniform (0,1) distribution. Thus, the joint probability density function for a z-value pair is given as

$$f(z_{1j}, z_{2j};\ \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{z_j}{2(1-\rho^2)}} \tag{3.1}$$

$\lambda_1^* \leq z_{1j} < \infty$, $\lambda_2^* \leq z_{2j} < \infty$. λ_i^* is the $p_{ij} \geq \lambda_i$ converted to z-values, $z_j \equiv z_{1j}^2 + z_{2j}^2 + 2\rho(z_{1j})(z_{2j})$, and $\rho \equiv cor(z_{1j}, z_{2j})$ is the correlation between the dependent $z_{ij}, i = 1{,}2$. The cumulative density function (CDF) is found to be

$$pA = \int_{\lambda_1^*}^{\infty} \int_{\lambda_2^*}^{\infty} f(z_{1j}, z_{2j}, \rho)\, dz_{2j}, dz_{1j}, \qquad \lambda_1^* \leq z_{1j} < \infty, \lambda_2^* \leq z_{2j} < \infty \tag{3.2}$$

The lower bounds for the bivariate z-values are obtained by converting the λ_i^* $(i = 1, 2)$ to bivariate z_i values. The joint density of the m^* pairs of the bivariate z-values are found, and using the maximum likelihood estimate, ρ, the correlation between the m^* pairs of dependent z-values from experiment 1 and experiment 2 were obtained. ρ was obtained using the optimize function in R.

Next, the estimate of m_{00} is calculated by dividing the number of p-values pairs, $p_{ij} \geq \lambda_i, i = 1,2$, the probability that a given pair … based on the CDF of z-values pairs with the estimated correlation, i.e.,

$$\widehat{m}_{00} = \frac{\sum_{j=1}^{m} \mathbf{1}\{(p_{1j}, p_{2j}) \in [\lambda_1, 1] \times [\lambda_2, 1]\}}{pA} \tag{3.3}$$

where

$$pA = \int_{\lambda_1^*}^{\infty} \int_{\lambda_2^*}^{\infty} f(z_{1j}, z_{2j}, \rho)\, dz_{2j}, dz_{1j} \tag{3.4}$$

Finally, the estimate of m_{11}, the number of genes that are differentially expressed in both experiments or analyses is,

$$\widehat{m}_{11} = m - \widehat{m}_0^{(1)} - \widehat{m}_0^{(2)} + \widehat{m}_{00} \tag{3.5}$$

The proposed method was compared with the intersection method by controlling FDR at 5% and 10% using Benjamini-Hochberg (BH) adjusted p-values so that $FDR \leq \alpha$, and the method by (Orr et. al., 2012) to evaluate its performances.

3.2. Simulation Studies

To compare the performance of the proposed method to other existing methods, we performed three separate simulation studies.

3.2.1. Simulation I – Analysis of Two Dependent Experiments

In the first simulation study, we use microarray data sets consisting of gene expressions from paired aliquots of lymphoblastoid cell lines (collected as a part of the HapMap project) treated with dexamethasone or vehicle (EtOH) for 8 hours. The data set is described by (Maranville JC, Luca F, Richards AL, Wen X et al., 2011) and available on GEO under the accession number GSE29342. Only D_trep1 and D_trep2 (experiments treated with dexamethasone) consisting of 22725 common genes and 112 subjects each were considered for this simulation. Paired t-test was conducted between the two experiments, and the mean difference was added to experiment 2 so that both experiments have the same means and p-values of 1. A total of m = 10,000 genes were randomly selected for the simulations. Each data set was simulated as follows:

1. For each experiment, treatment effects were generated following the settings in section 3.2. of (Orr et al., 2012). Simulations were done in a similar manner as in 3.2 of (Orr et al., 2012).
2. m_{10}, the number of DE genes in experiment 1, but EE in 2 are selected for each simulation setting. Treatment effects are added to m_{10}.
3. m_{01}, the number of DE genes in experiment 2, but EE in 1 are selected for each simulation setting. Treatment effects are added to m_{01}.
4. For each simulation setting, treatment effects are added to m_{11}, the selected DE genes in both experiments.
5. A total of $2n$ subjects are selected from each experiment and further split into 2 groups of n subjects each.
6. Limma package for moderated t-test in R was used to obtain p-values from each experiment and combined to forms m pairs of p-values.

7. m^* p-value pairs are found as discussed in 3.1.
8. Finally, $\widehat{m}_{00}$ and $\widehat{m}_{11}$ genes are estimated using the proposed method and the results are compared to (Orr et al., 2012) and intersection method.

3.2.2. Simulation II – Analysis of Two Dependent Experiments

In the second simulation study, the same data sets in simulation I are used, but for this simulation, E_trep1 and E_trep2 (experiments treated with vechicle (EtOH)) are considered, and the mean difference was not added to the experiments. The same simulation settings in section 3.2.1. are employed.

3.2.3. Simulation III – A Single Experiment with Two Dependent Analyses

In this simulation study, the gene expression data sets from a single two-group experiment were simulated using a real microarray data set. This data set consists of 462 patients diagnosed with chronic lymphocytic leukemia (CLL)/MBL patients that were prospectively enrolled from several Italian Institutions in an observational multicenter study (O-CLL1 protocol, clinical-trial.gov identifier NCT00917540) from January 2007 to May 2011.

The data set can be obtained from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51528). A total of m = 10,000 genes from this data set were randomly selected to be included in the simulations. Following a procedure similar to that of (Orr et al., 2012), each data set was simulated as follows:

1. A total of $2n$ experimental units were randomly selected from the samples and further split into 2 treatment groups of n subject each.
2. Treatment effects were generated following the settings in section 3.1. of (Orr et al., 2012), and added to one of the treatment groups chosen to contain DE genes.

3. Student t-test and moderated t-test (limma) were employed to obtain the m p-value pairs from the two treatment groups separately.
4. m^* p-value pairs are found as discussed in section 3.1.
5. m_{11}, differentially expressed genes are estimated using the proposed method.

3.2.4. Methods Compared

The proposed method was compared with (Orr et al., 2012) and the intersection method.

3.2.5. Performance Measures

The mean estimates of m_{11} and root mean squared errors (RMSEs) were reported for each setting of 100 simulated data sets. RMSE was used to evaluate the performances of the proposed method against other methods and used boxplots for visualization of the results.

3.3. Real Data Analysis

We analyzed the microarray data set from a real gene expression experiment. As described by (Glen et al., 2015), the data set was obtained from experiments in which Male CBA/Ca and BALB/c inbred mice were obtained from Harlan (Bicester, UK) and housed at the Division of Biomedical Services, University of Leicester. The tissue samples were taken from the liver and kidney of each mouse with the purpose of comparing or determining genes that are differentially expressed in both CBA/Ca and BALB/c mice. The focus of this analysis is the same as the one done by (Glen et al., 2015), to estimate the number of genes that are differentially expressed in both experiments. There are 42,575 genes common to both experiments, with experiment 1 and 2 having the RNA samples extracted from liver and kidney of treated BALB/c and CBA/Ca male mice respectively. Using the histogram-based method by (Liang and Nettleton, 2012) and the λ-estimator by (Storey and Tibshirani, 2003), we extracted p-value pairs in the upper-right quadrant of the histogram of p-values from both experiments. The p-values were obtained for each experiment separately using student t-test and moderated t-test for the hypothesis testing as described in section 2.5.1. and 2.5.2. The proposed method was then used to estimate genes that are differentially expressed in both experiments. The results are compared to (Orr et al., 2012) and the intersection method as shown in section 4.5.

CHAPTER 4: RESULTS

4.1. Simulation Results

For each simulation setting, 100 data sets were simulated, the mean and root mean squared error (RMSE) of the estimated m_{11} are reported for each of the methods considered in the analysis, and using the proposed method, the mean of the correlation between the two experiments are reported. The lowest RMSE is in bold font and parenthesis; n represents the number of subjects/samples in each treatment group; μ_δ is the relative effect size; m_{00} is the number of EE genes in both experiments; m_{11} is the number of DE genes (genes with treatment effects as discussed in the simulation settings) in both experiments; $\widehat{m}_{11}$, is the estimated number of DE genes in both experiments; $FDR \leq 0.05$ and $FDR \leq 0.1$ represent the level at which the FDR is controlled for the intersection method; $\bar{\rho}$ is the mean of the correlation between the experiments. Also, for every simulated data and simulation setting presented in the table, their boxplots are reported for visualization.

4.2. Results of Simulation I - Two Dependent Experiments

Table 4.1. presents the results of analyzing the data from the two dependent experiments involving D_trep1 and D_trep2 (experiments treated with dexamethasone) consisting of 22725 common genes and 112 subjects/samples described in Section 3.2.1.

In table 4.1., when the relative effect size is 1, the sample size is 4, and the differentially expressed genes are 500, we found that even though the intersection method seemed to have the lowest RMSE when the FDR is controlled at both 0.05 and 0.1, but was only able to estimate 0 differentially expressed genes in both experiments. There is a similar occurrence when the relative effect size is 2, the intersection method has the lowest RMSE, the proposed method estimated 1 differentially expressed gene in both experiments. This is evident from the boxplots as well.

Overall, aside from the two cases stated above, in other simulation settings, the proposed method has the lowest RMSE in four settings, two for the intersection method while (Orr et al., 2012) also has four. The mean of the correlation between the p-values pairs of the two dependent experiments is reported to be between -0.01 and 0.08 by the proposed method. In figure 4.1. and figure 4.2., there seems to be enough variation between the p-value pairs and appear to be normally distributed for most of the simulation settings as identified by the proposed method and (Orr et al., 2012). The intersection methods identified low number of DE genes.

Table 4.1: Results of Simulation I - Two Dependent Experiments as described in section 3.2.1. The mean and RMSE for the 100 simulated data sets are reported for estimated m_{11} for each simulation setting. The lowest RMSE amongst the methods is in bold font and parenthesis.

n	μ_δ	m_{00}	m_{11}	Proposed	Orr et al., 2012	Intersection Method		$\bar{\rho}$
				$\hat{m}_{11}$	$\hat{m}_{11}$	$FDR \leq 0.05$	$FDR \leq 0.1$	
4	1	8500	500	752(640)	876(654)	**0(500)**	**0(500)**	0.05
		7000	1000	928(485)	**1049(451)**	0(1000)	0(1000)	0.03
		1000	3000	1439(1678)	**1367(1669)**	1(2999)	1(2999)	-0.01
	2	8500	500	826(653)	957(738)	**1(499)**	**1(499)**	0.06
		7000	1000	**1229(553)**	1389(627)	1(999)	2(998)	0.05
		1000	3000	2421(756)	**2387(715)**	4(2996)	7(2993)	-0.01
10	1	8500	500	642(553)	769(594)	5(495)	**6(494)**	0.07
		7000	1000	**960(468)**	1122(491)	10(990)	13(987)	0.05
		1000	3000	**1856(1223)**	1788(1258)	28(2972)	39(2961)	-0.01
	2	8500	500	769(542)	932(670)	49(451)	**62(438)**	0.08
		7000	1000	**1197(543)**	1313(608)	98(902)	124(876)	0.07
		1000	3000	2622(528)	**2626(498)**	307(2693)	388(2613)	0.01

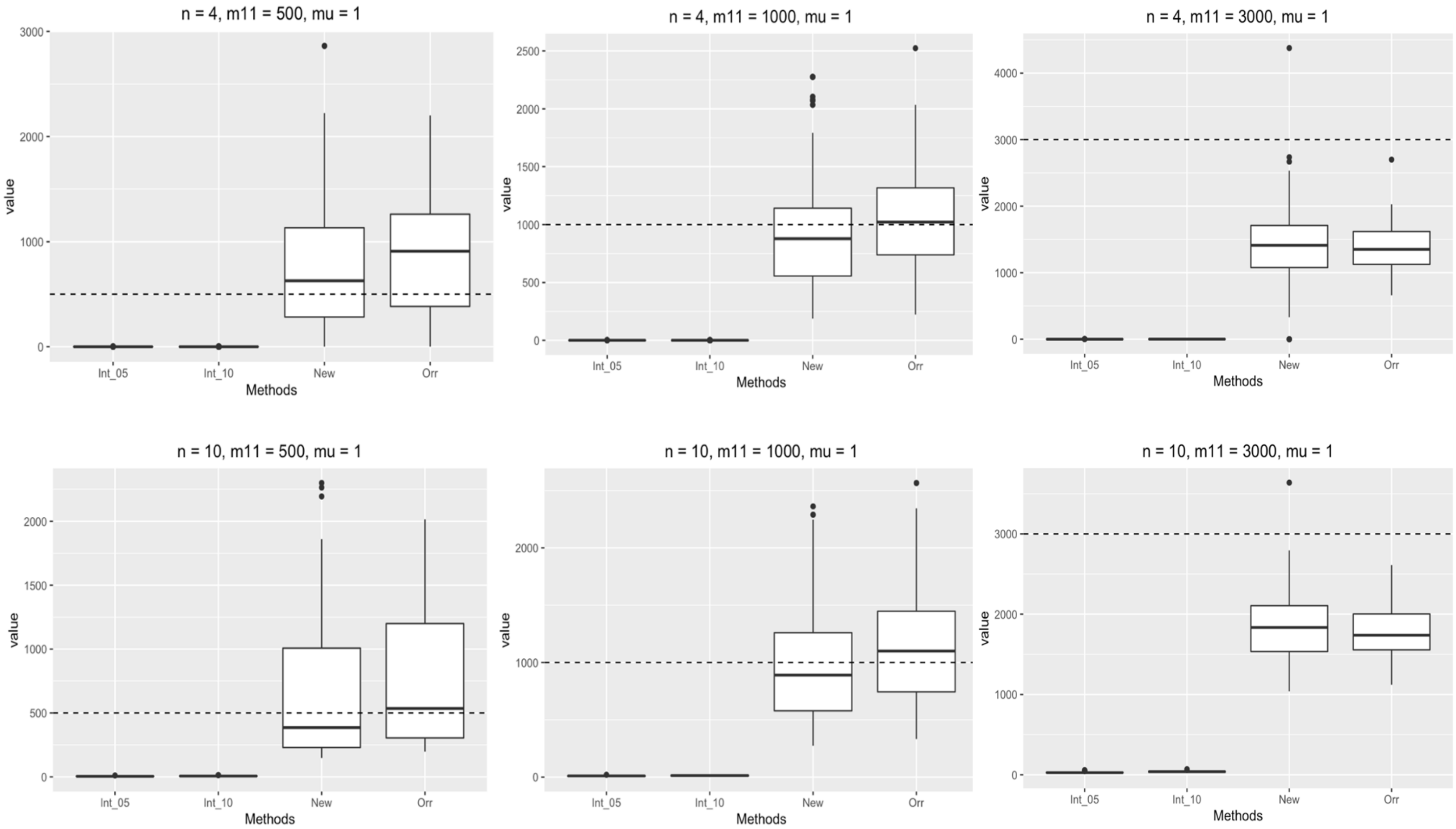

Figure 4.1: Boxplots of the results of Simulation I. The 100 simulated data for different simulation settings described in 3.2.1. with $n = 4\ and\ 10$ subjects/samples in each treatment group, and $\mu_\delta = 1$. New is the results of the proposed method, Orr represents the results for the method by (Orr et al., 2012) while Int_{05} and Int_{10} represent the results for the intersection method when FDR is controlled at 5% and 10% respectively.

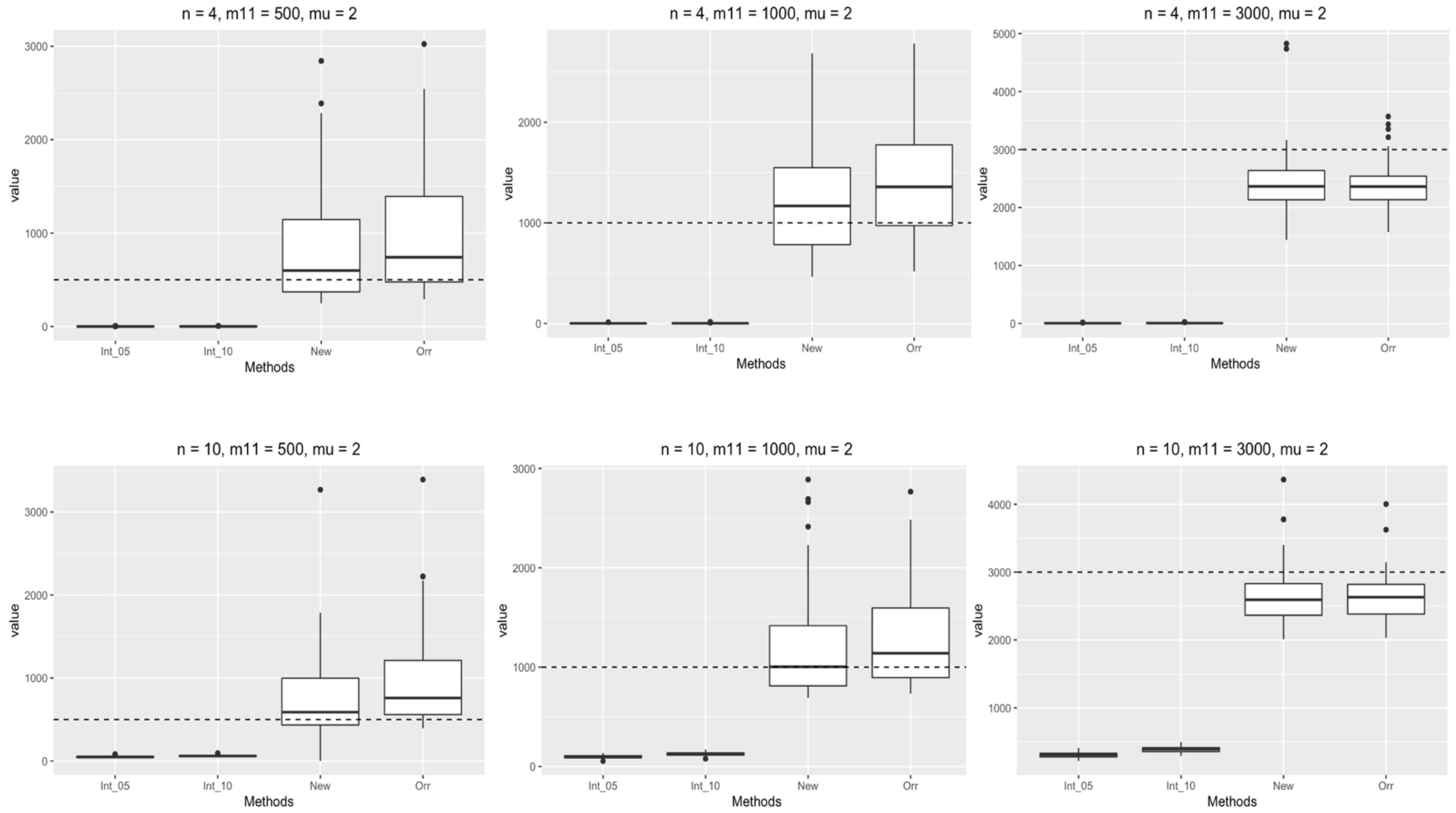

Figure 4.2: Boxplots of the results of Simulation I. The 100 simulated data for different simulation settings described in 3.2.1. with $n = 4\ and\ 10$ subjects/samples in each treatment group, and $\mu_\delta = 2$. New is the results of the Proposed method, Orr represents the results for the method by (Orr et al., 2012) while Int_{05} and Int_{10} represent the results for the intersection method when FDR is controlled at 5% and 10% respectively.

4.3. Results of Simulation II – Two Dependent Experiments

Table 4.2. presents the results of analyses of the data from the two dependent experiments involving E_trep1 and E_trep2 (experiments treated with vechicle (EtOH)) consisting of 22725 common genes and 112 subjects/samples described in Section 3.2.2. In this section, the report is similar to that of simulation I.

In table 4.2., similar to what we noticed in the result of simulation I, when the relative effect size is 1 and 2, the sample size is 4, and the differentially expressed genes are 500, the intersection method gives the same results as in simulation I. In general, the proposed method has the lowest RMSE in five simulation settings, (Orr et al., 2012) also has three while the intersection method has two for both when the FDR is being controlled at 5% and 10%, and one more at 10%. The mean correlation for the 100 simulated data sets between the p-values pairs of the two dependent experiments is reported to be between -0.01 and 0.09 by the proposed method. In figure 4.3. and 4.4., there seems to be enough variation between the p-value pairs for the 100 simulated data sets, and the p-value pairs appear to be normally distributed with a few outliers for most of the simulation settings as identified by the proposed method and (Orr et al., 2012). The intersection methods identified low number of DE genes.

Table 4.2: Results of Simulation II – Two Dependent Experiments as described in section 3.2.2. The mean and RMSE for the 100 simulated data sets are reported for estimated m_{11} for each simulation setting. The lowest RMSE amongst the methods is in bold font and parenthesis.

n	μ_δ	m_{00}	m_{11}	$Proposed$	Orr	Intersection Method		$\bar{\rho}$
				$\widehat{m}_{11}$	$\widehat{m}_{11}$	$FDR \leq 0.05$	$FDR \leq 0.1$	
4	1	8500	500	705(639)	743(623)	**0(500)**	**0(500)**	0.06
		7000	1000	925(522)	**1054(459)**	0(1000)	0(1000)	0.04
		1000	3000	**1481(1664)**	1346(1698)	1(2999)	1(2999)	-0.01
	2	8500	500	826(654)	934(700)	**1(499)**	**1(499)**	0.06
		7000	1000	**1229(533)**	1327(572)	1(999)	2(998)	0.04
		1000	3000	**2426(736)**	2331(759)	4(2996)	7(2993)	-0.03
10	1	8500	500	710(579)	830(640)	5(495)	**7(493)**	0.07
		7000	1000	994(511)	**1097(479)**	9(991)	13(987)	0.05
		1000	3000	**1793(1287)**	1750(1305)	29(2971)	41(2959)	-0.01
	2	8500	500	792(580)	945(726)	50(450)	**62(439)**	0.09
		7000	1000	**1248(565)**	1373(655)	98(902)	123(877)	0.07
		1000	3000	2614(547)	**2611(529)**	304(2697)	384(2616)	0.00

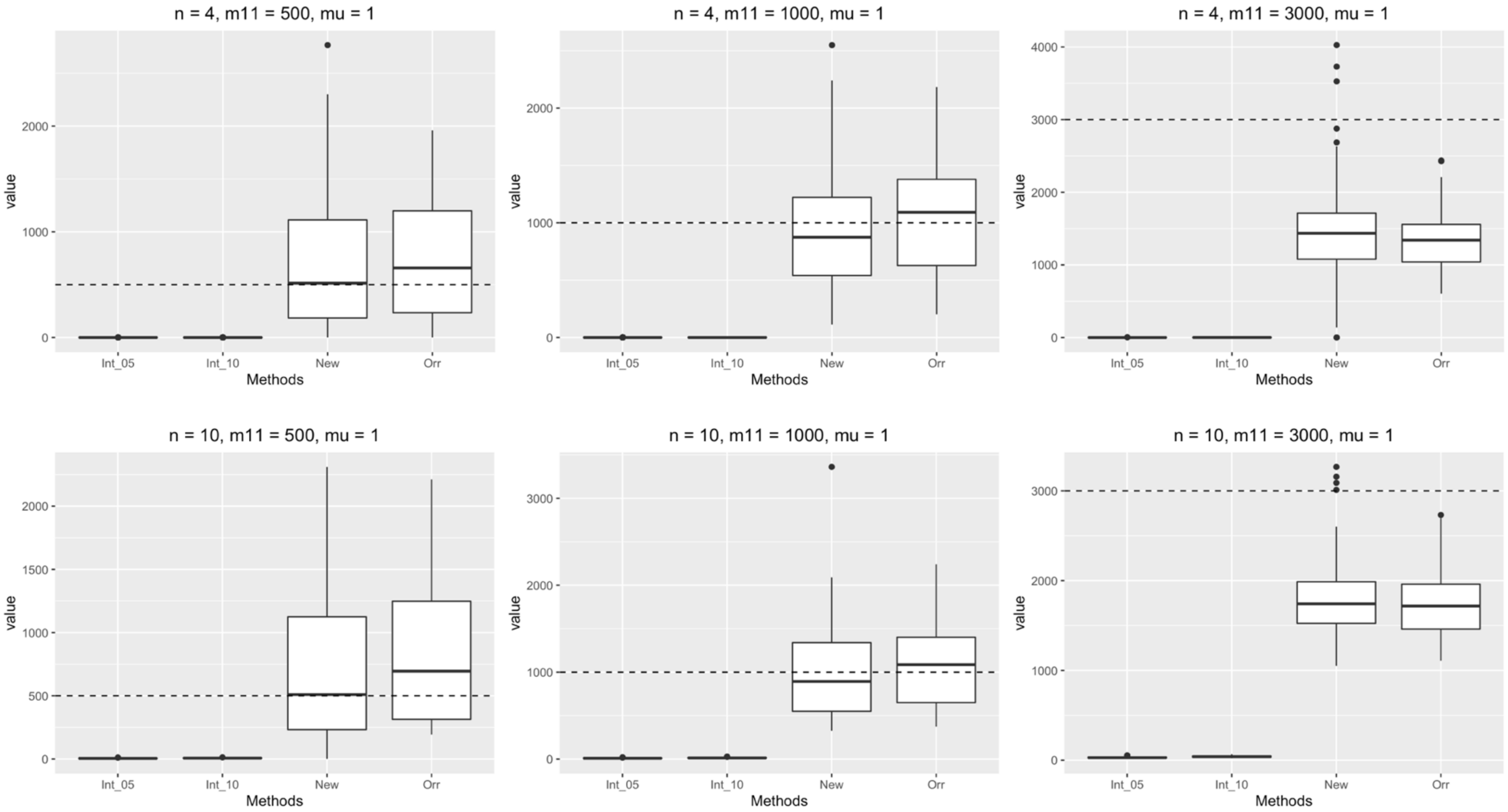

Figure 4.3: Boxplots of the results of Simulation II. The 100 simulated data for different simulation setting described in 3.2.2. with $n = 4\ and\ 10$ subjects/samples in each treatment group, and $\mu_\delta = 1$; New is the results of the proposed method; Orr represents the results for the method by (Orr et al., 2012); while Int_{05} and Int_{10} represent the results for the intersection method when FDR is controlled at 5% and 10% respectively

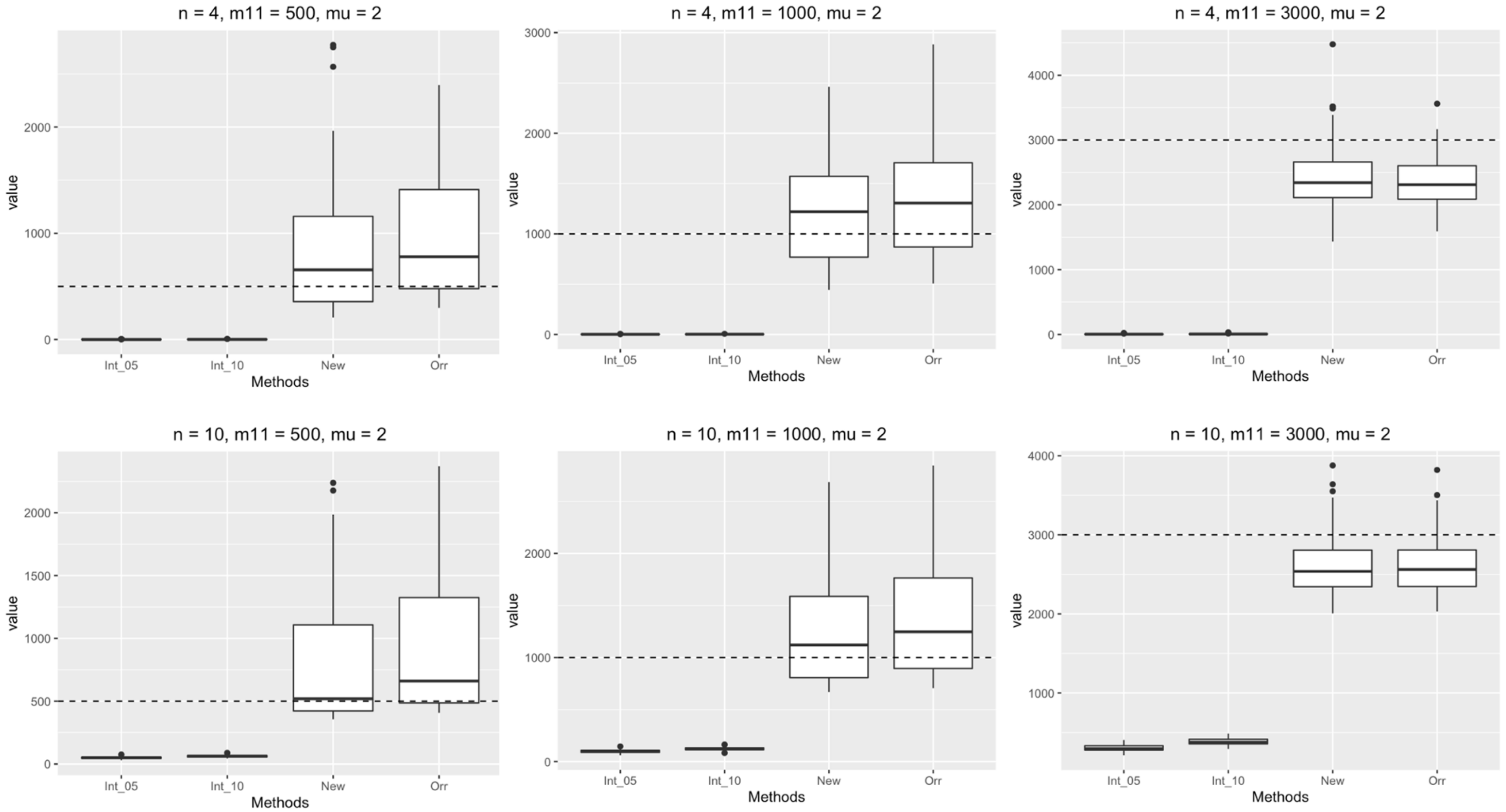

Figure 4.4: Boxplots of the results of Simulation II. The 100 simulated data for different simulation setting described in 3.2.2. with $n = 4 \; and \; 10$ subjects/samples in each treatment group, and $\mu_\delta = 2$; New is the results of the proposed method; Orr represents the results for the method by (Orr et al., 2012); while Int_{05} and Int_{10} represent the results for the intersection method when FDR is controlled at 5% and 10% respectively.

4.4. Results of Simulation III – A Single Experiment with Two Dependent Analyses

Table 4.3. presents the results of the analysis of the data sets from a single two-group experiment simulated using a real microarray data set. This data set consists of 462 patients diagnosed with chronic lymphocytic leukemia (CLL)/MBL patients described in Section 3.2.3.

In table 4.3., It is observed that for the 12 different simulation settings considered, the proposed method has the lowest RMSE in 7 when estimating the number of DE genes in both experiments, the intersection method has 5 cases when FDR is controlled at 10%. There appears to be a strong dependency between the experiments having a mean correlation ranging between 0.982 and 0.999. In figure 4.5. and 4.6., there seems to be enough variation between the p-value pairs for the 100 simulated data sets, and the p-value pairs appear to be normally distributed with a few outliers for most of the simulation settings when estimating DE genes using the proposed method.

Table 4.3: Results of Simulation III – A Single Experiment with Two Dependent Analyses as described in section 3.2.3. The mean and RMSE for the 100 simulated data sets are reported for estimated $m_{11}$for each simulation setting. The lowest RMSE amongst the methods is in bold font and parenthesis.

n	μ_δ	m_{00}	m_{11}	Proposed	Orr	Intersection Method		$\bar{\rho}$
				$\widehat{m}_{11}$	$\widehat{m}_{11}$	$FDR \leq 0.05$	$FDR \leq 0.1$	
4	1	9000	1000	886(1031)	0(1000)	3(997)	**19(984)**	0.999
		7000	3000	**1904(1527)**	35(2985)	67(2934)	230(2776)	0.993
		5000	5000	**2856(2320)**	0(5000)	229(4776)	638(4376)	0.994
	2	9000	1000	1137(1031)	15(996)	77(924)	**196(809)**	0.990
		7000	3000	**2452(1004)**	667(2673)	781(2238)	1273(1754)	0.982
		5000	5000	**4268(1052)**	1821(4265)	1807(3223)	2648(2388)	0.993
10	1	9000	1000	1178(1179)	14(996)	309(694)	**397(617)**	0.999
		7000	3000	**2196(1157)**	508(2684)	1196(1809)	1451(1558)	0.999
		5000	5000	**3706(1486)**	734(4655)	2212(2797)	2658(2356)	0.999
	2	9000	1000	1287(1009)	17(998)	698(308)	**787(233)**	0.999
		7000	3000	2884(771)	1654(2284)	2289(720)	**2513(510)**	0.999
		5000	5000	**4536(715)**	4379(3295)	3981(1028)	4294(722)	0.999

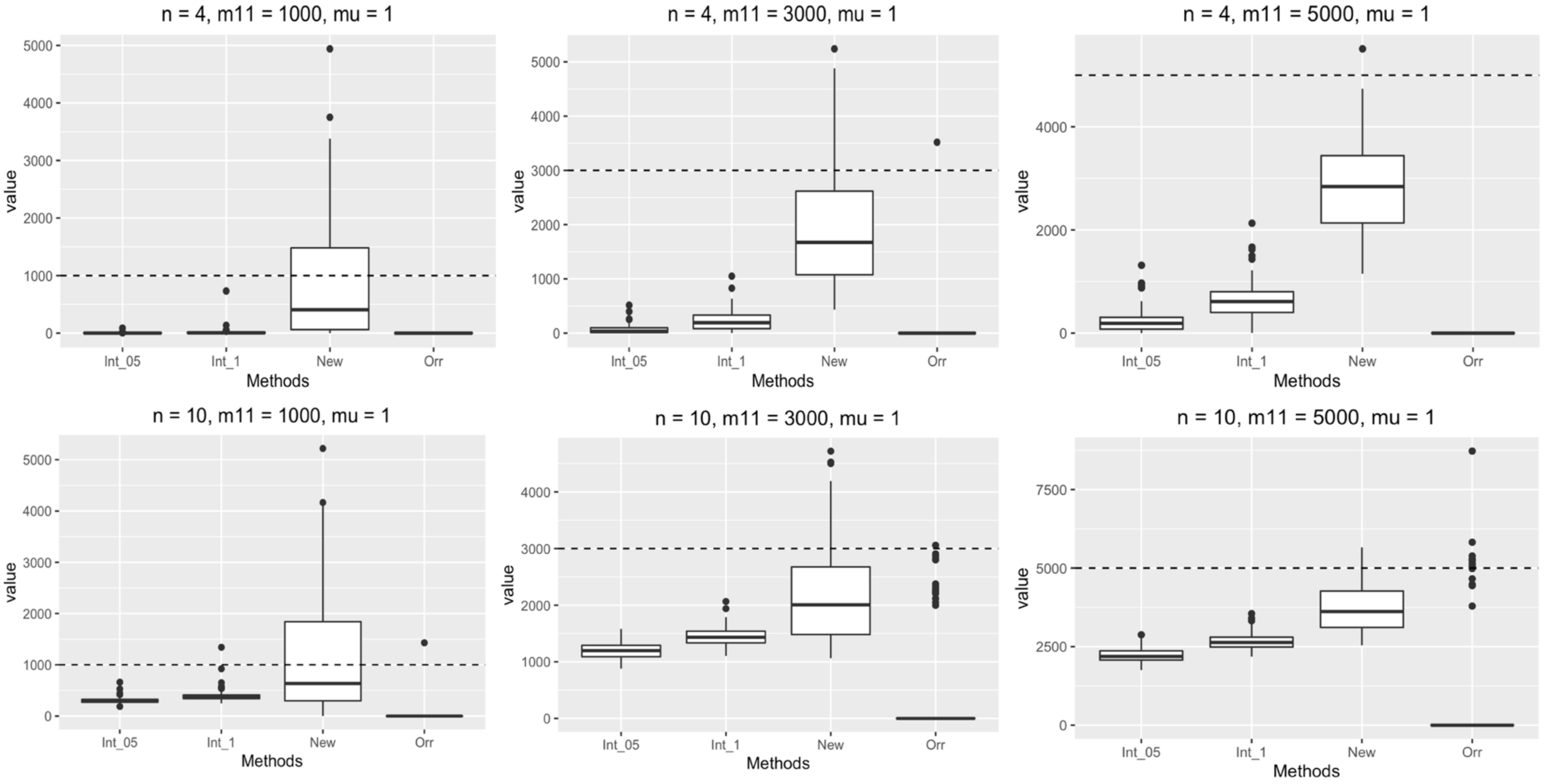

Figure 4.5: Boxplots of the results of Simulation III. The 100 simulated data for different simulation setting described in 3.2.3. with $n = 4\ and\ 10$ subjects/samples in each treatment group, and $\mu_\delta = 1$; New is the results of the proposed method; Orr represents the results for the method by (Orr et al., 2012); while Int_{05} and Int_{10} represent the results for the intersection method when FDR is controlled at 5% and 10% respectively.

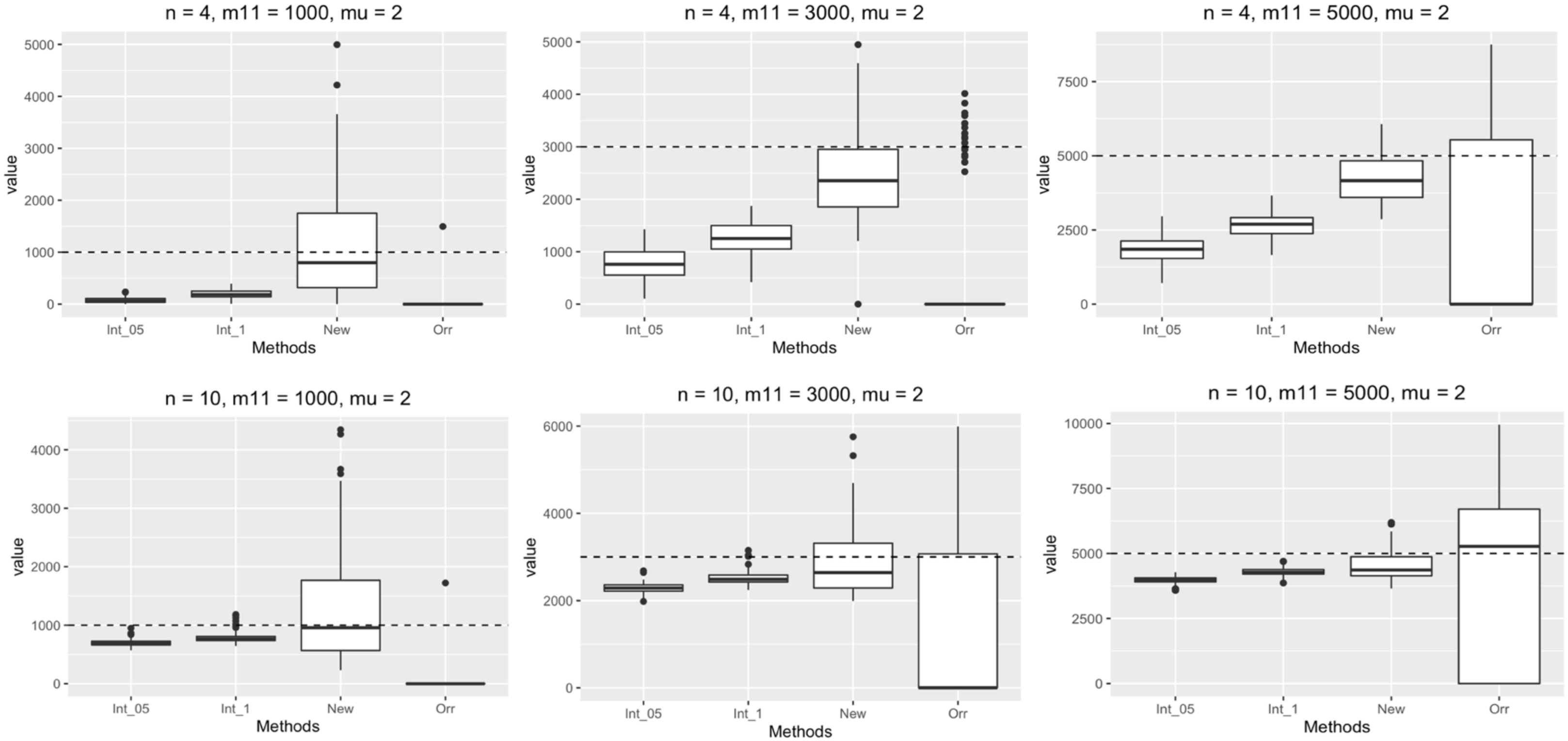

Figure 4.6: Boxplots of the results of Simulation III. The 100 simulated data for different simulation setting described in 3.2.3. with $n = 4\ and\ 10$ subjects/samples in each treatment group, and $\mu_\delta = 2$; New is the results of the proposed method; Orr represents the results for the method by (Orr et al., 2012); while Int_{05} and Int_{10} represent the results for the intersection method when FDR is controlled at 5% and 10% respectively.

4.5. Real Data Analysis Results

Table 4.4. and 4.5. present the results of analyzing the data from the two dependent experiments involving kidney and liver gene expression described in Section 3.3. using Student's t-tests and moderated t-test (limma) respectively.

In table 4.4., we found the cut-off points $\lambda_1 = 0.8 \; and \; \lambda_2 = 0.9$ for the kidney and liver experiments, respectively, while in table 4.5., the cut-off points are $\lambda_1 = 0.95 \; and \; \lambda_2 = 0.8$ for the kidney and liver experiments respectively. Both the proposed and (Orr et al., 2012) results suggest a high level of differential expression common to both experiments, with the proposed method having the highest estimate of 20,821 genes differentially expressed in both experiments using student t-test for the hypothesis testing and obtaining the p-value pairs for both experiments, but when moderated t-test was used, (Orr et al., 2012) has the highest estimate of 21,373 differentially expressed genes in both experiments. The intersection method identified a much smaller estimate of the number of genes that are differentially expressed in both experiments using Benjamini-Hochberg adjusted p-values for both experiments. The estimate of the correlations between the p-value pairs by the proposed method were -0.09 and 1.00 using student t-test and moderated t-test respectively.

Table 4.4: Result of the analysis from the two dependent experiments involving kidney and liver gene expression described in Section 3.3. using Student's t-test.

	Proposed	Orr et al. 2012	Intersection	
			$FDR \leq 0.05$	$FDR \leq 0.1$
Cut-off	$\lambda_1 = 0.8, \lambda_2 = 0.9, \rho = -0.09$			
m	42575			
m_{00}	3366	2650	40979	40346
m_{11}	20821	20105	98	124

Table 4.5: Result of the analysis from the two dependent experiments involving kidney and liver gene expression described in Section 3.3. using Moderated t-test (limma).

	Proposed	Orr et al. 2012	Intersection	
			$FDR \leq 0.05$	$FDR \leq 0.1$
Cut-off	$\lambda_1 = 0.95, \lambda_2 = 0.8, \rho = 1.0$			
m	42575			
m_{00}	660	3300	38624	37660
m_{11}	18735	21375	248	296

In table 4.6., we present the results of analyzing the data from a single experiment with two dependent analyses involving kidney gene expression and present the results of the analysis involving liver gene expression in table 4.7. as described in Section 3.3. Student t-test and moderated t-test are used for the hypothesis testing separately on each gene expression to obtain the m p-value pairs.

In table 4.6., we found the points $\lambda_1 = 0.8$ and $\lambda_2 = 0.95$ for the kidney gene expression involving analysis 1 and 2 respectively, while in table 4.7., the cut-off points are $\lambda_1 = 0.9$ and $\lambda_2 = 0.8$ for the liver gene expression involving analysis 1 and 2 respectively. The results suggest a high level of differential expression common to both analyses, with the proposed method having the highest estimate of 34,575 and 25,515 genes differentially expressed in kidney and liver gene expressions respectively. Using Benjamini-Hochberg adjusted p-values for each p-value from both analyses, the intersection method identified a much smaller estimate of the number of genes differentially expressed in both analyses. The estimate of the correlations between p-value pairs by the proposed method were 1.0 for both results in table 4.6. and 4.7. For both results presented in table 4.6. and 4.7., number of differentially expressed genes in both analyses is set to 0 for (Orr et al., 2012) because of the very large estimate of m_{00} $(m_{00} > m)$ by the method. This is due to the very high correlation between the m^* p-value pairs. The p-values are not spread out randomly throughout the entire scatterplot (as we would expect if the p-values in a pair were independent) but are concentrated in the lower left quadrant and upper right quadrant of the histogram of p-values from both analyses. Because the points are "overrepresented" in this upper right quadrant compared to what is expected under independence, the number of the equivalently expressed genes is also very large, much larger than m (the overall number of genes in both analyses).

Table 4.6: Results of analyzing the data from a single experiment with two dependent analyses involving kidney gene expression described in Section 3.3. using student t-test and limma to obtain m p-value pairs.

	Proposed	Orr et al. 2012	Intersection	
			$FDR \leq 0.05$	$FDR \leq 0.1$
Cut-off	$\lambda_1 = 0.8, \lambda_2 = 0.95, \rho = 1.0$			
m	42575			
m_{00}	7440	42575	39336	38500
m_{11}	34575	0	1097	1567

Table 4.7: Results of analyzing the data from a single experiment with two dependent analyses involving liver gene expression for the analysis described in Section 3.3. using student t-test and limma to obtain m p-value pairs.

	Proposed	Orr et al. 2012	Intersection	
			$FDR \leq 0.05$	$FDR \leq 0.1$
Cut-off	$\lambda_1 = 0.9, \lambda_2 = 0.8, \rho = 1.0$			
m	42575			
m_{00}	17120	42575	41472	41247
m_{11}	25515	0	454	594

CHAPTER 5: CONCLUSION AND RECOMMENDATION FOR FUTURE RESEARCH

5.1. Conclusion

In this research work, gene expressions involving two dependent experiments or analyses were analyzed to determine the number of differentially expressed genes in both experiments or analyses using different methods. Microarray data sets were used for both simulation studies and real data analysis. The mean estimates and RMSE were reported for each simulation setting of the 100 simulated data sets. RMSE is used to measure and compare the performances of the methods. The mean correlation between the two experiments or analyses was reported for each 100 simulated data set using the proposed method.

In the simulation settings involving two dependent experiments, the proposed method had the lowest RMSE in four and five different simulation settings for simulation 1 and 2 respectively; (Orr et al., 2012) had the lowest RMSE in three different settings for both simulation 1 and 2; the intersection method had the lowest RMSE in 4 settings when FDR is controlled at 10% in simulation 1 and 2. Although, the intersection method had the lowest RMSE when the FDR is controlled at 5% and 10% for simulation 1 and 2, it resulted in estimates 0 and 1 differentially expressed gene in both experiments. In addition, the high correlation between gene expressions from the same experimental units did not translate to high correlation between the p-value pairs.

In the simulation involving a single experiment with two dependent analyses, the proposed method had the lowest RMSE in seven of the twelve simulation settings while the intersection method had the lowest RMSE in the other five settings when FDR is controlled at 10% but failed completely when it is controlled at 5%. The performance of the intersection method depends on the choice or level of α. (Orr et al., 2012) failed completely amongst other methods due to the high correlation between the p-values of the two analyses. The proposed method is the clear winner when comparing two dependent analyses of the same data set having the highest number of lowest RMSE but did not perform as well as expected in the two dependent experiments case for simulated data sets.

In the real data analysis, the proposed method identified more differentially expressed genes than the other methods for the analyses involving two dependent experiments when using student t-test for the hypothesis testing, but when using moderated t-test for the hypothesis testing, (Orr et al., 2012) identified more differentially expressed genes than the other methods.

5.2. Future Research

Proposed method can be used in situations where experiments are independent because estimated correlation will be close to zero.

We assume that the transformed z-scores from equivalently expressed genes follow a bivariate normal distribution, but some other distribution may be more appropriate and investigated in the future.

The performance of the proposed method can be explored further using RNA-seq experiments.

REFERENCES

Aubert, J., Bar-Hen, A., Daudin, J. J. et al. (2004). Determination of the differentially expressed genes in microarray experiments using local FDR. BMC Bioinformatics 5, 125. https://doi.org/10.1186/1471-2105-5-125

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289–300. http://www.jstor.org/stable/2346101

Bull, J., Ellison, G., Patel, A. et al. (2001). Identification of potential diagnostic markers of prostate cancer and prostatic intraepithelial neoplasia using cDNA microarray. Br J Cancer 84, 1512–1519. https://doi.org/10.1054/bjoc.2001.1816

Chen Y., Lun A. A. T., Smyth G. K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, 5, 1438. doi: 10.12688/f1000research.8987.2.

Covshoff, S., Majeran, W., Liu, P., Kolkman, J. M., van Wilk, K. J., and Brutnell, T. P. (2008). Deregulation of Maize C4 Photosynthetic Development in a Mesophyll Cell-Defective Mutant, Plant Physiology, 146, 1469–1481.

Dudoit, S., Shaffer, J. & Boldrick, J. (2003). Multiple Hypothesis Testing in Microarray Experiments. Statistical Science. 18. 10.1214/ss/1056397487.

Fang, Z., Martin, J. & Wang, Z. (2012) Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. Cell Biosci 2, 26. https://doi.org/10.1186/2045-3701-2-26

Gao, H., Li, S., Wang, M., Yan, S., Cui, W., Ma, Z. ... Li, X. (2021). Screening and identification of differentially expressed microRNAs in diffuse large B-cell lymphoma based on microRNA microarray. Oncology Letters, 22, 753. https://doi.org/10.3892/ol.2021.13014

Giuseppe R, Charles Z. & Antonio G. (2003). Advantages, and limitations of microarray technology in human cancer, Oncogene volume 22, pages 6497–6507.

Glen, C. D., McVeigh, L. E., Voutounou, M., & Dubrova, Y. E. (2015). The effects of methyl-donor deficiency on the pattern of gene expression in mice. Molecular nutrition & food research, 59(3), 501–506. https://doi.org/10.1002/mnfr.201400660

Hong, M., Tao, S., Zhang, L. et al. (2020). RNA sequencing: new technologies and applications in cancer research. J Hematol Oncol 13, 166. https://doi.org/10.1186/s13045-020-01005-x

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. Nucleic acids research, 31(4), e15. https://doi.org/10.1093/nar/gng015

Lai, Y., Adam, B., Podolsky, R., and She, J. X. (2007). A Mixture Model Approach to the Tests of Concordance and Discordance Between Two Large-Scale Experiments with Two-Sample Groups. Bioinformatics, 23, 1243–1250.

Langaas, M., Lindqvist, B. & Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. Journal of the Royal Statistical Society Series B. 67. 555-572. 10.1111/j.1467-9868.2005.00515.x.

Liang, K., and Nettleton, D. (2012). Adaptive and Dynamic Adaptive Procedures for False Discovery Rate Control and Estimation. Journal of the Royal Statistical Society, Series B, 74, 163–182.

Liu, R., Cheng, W. J., Ye, F., Zhang, Y. D., Zhong, Q. P., Dong, H. F., Tang, H. B., & Jiang, H. (2020). Comparative Transcriptome Analyses of Schistosoma japonicum Derived from SCID Mice and BALB/c Mice: Clues to the Abnormality in Parasite Growth and Development. Frontiers in microbiology, 11, 274. https://doi.org/10.3389/fmicb.2020.00274

Lockhart D.J., Winzeler E.A. (2000). Genomics, gene expression and DNA arrays. Nature. 405(6788):827-36. doi: 10.1038/35015701. PMID: 10866209.

Love, M.I., Huber, W. & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Maura, F., Cutrona, G., Mosca, L., Matis, S., Lionetti, M., Fabris, S., Agnelli, L., Colombo, M., Massucco, C., Ferracin, M., Zagatti, B., Reverberi, D., Gentile, M., Recchia, A. G., Bossio, S., Rossi, D., Gaidano, G., Molica, S., Cortelezzi, A., Di Raimondo, F., Neri, A. (2015). Association between gene and miRNA expression profiles and stereotyped subset #4 B-cell receptor in chronic lymphocytic leukemia. Leukemia & lymphoma, 56(11), 3150–3158. https://doi.org/10.3109/10428194.2015.1028051

Maranville, J. C., Luca, F., Richards, A. L., Wen, X., Witonsky, D. B., Baxter, S., Stephens, M., & Di Rienzo, A. (2011). Interactions between glucocorticoid treatment and cis-regulatory polymorphisms contribute to cellular response phenotypes. PLoS genetics, 7(7), e1002162. https://doi.org/10.1371/journal.pgen.1002162

McCarthy D.J., Chen Y., Smyth G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10), 4288-4297. doi: 10.1093/nar/gks042.

Miyama, M. & Hanagata, N. (2007). Microarray analysis of 7029 gene expression patterns in Burma mangrove under high-salinity stress. Plant Science. 172. 948-957. 10.1016/j.plantsci.2007.01.004.

Nettleton, D., Hwang, J., Caldo, R., and Wise, R. (2006), “Estimating the Number of True Null Hypotheses from a Histogram of p Values,” Journal of Agricultural, Biological, and Environmental Statistics, 11, 337–356.

Orr, M., Liu, P., & Nettleton, D. (2012). Estimating the Number of Genes That Are Differentially Expressed in Both of Two Independent Experiments. Journal of Agricultural, Biological, and Environmental Statistics, 17(4), 583–600. http://www.jstor.org/stable/41724592

Orr, M. (2022). Empirical Bayes Analysis of Microarray Experiments, Introduction to Statistical Design and Analysis of Gene Expression Experiments, Department of Statistics, North Dakota State University.

Richard J. M. MBBS. (2020). Genetic Aspects of Perinatal Disease and Prenatal Diagnosis, FRACP, in Fanaroff and Martin's Neonatal-Perinatal Medicine.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43, e47. http://nar.oxfordjournals.org/content/43/7/e47

Robinson, M. D., & Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. Bioinformatics (Oxford, England), 23(21), 2881–2887. https://doi.org/10.1093/bioinformatics/btm453

Robinson M. D., McCarthy D. J., Smyth G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139-140. doi: 10.1093/bioinformatics/btp616.

Russo, G., Zegar, C. & Giordano, A. (2003). Advantages, and limitations of microarray technology in human cancer. Oncogene 22, 6497–6507. https://doi.org/10.1038/sj.onc.1206865

Sandford, E. E., Orr, M., Shelby, M., Li, X., Zhou, H., Johnson, T. J., Kariyawasam, S., Liu, P., Nolan, L. K., & Lamont, S. J. (2012). Leukocyte transcriptome from chickens infected with avian pathogenic Escherichia coli identifies pathways associated with resistance. Results in immunology, 2, 44–53. https://doi.org/10.1016/j.rinim.2012.02.003

Shankar K., Mehendale H. M. (2014). Encyclopedia of Toxicology (Third Edition).

Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments, Statistical Applications in Genetics and Molecular Biology, 3, 3.

Storey, J. D. (2002). A Direct Approach to False Discovery Rates, Journal of the Royal Statistical Society, Series B, 64, 479–498.

Storey, J. D., Taylor, J., and Siegmund, D. (2004). Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach, Journal of the Royal Statistical Society, Series B, 66, 187–205.

Storey, J. D., and Tibshirani, R. (2003). Statistical Significance for Genomewide Studies, Proceedings of the National Academy of Sciences of the United States of America, 100, 9440–9445.

te Velde, A. A., de Kort, F., Sterrenburg, E., Pronk, I., ten Kate, F. J., Hommes, D. W., & van Deventer, S. J. (2007). Comparative analysis of colonic gene expression of three experimental colitis models mimicking inflammatory bowel disease. Inflammatory bowel diseases, 13(3), 325–330. https://doi.org/10.1002/ibd.20079

Wikipedia contributors. (2022). False discovery rate. In Wikipedia, The Free Encyclopedia. Retrieved 05:51, April 13, 2022, from https://en.wikipedia.org/w/index.php?title=False_discovery_rate&oldid=1072358810