International Journal of Mathematics and Statistics Invention (IJMSI)
E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759
www.ijmsi.org Volume 11 Issue 1 || Jan. - Feb., 2023 || PP 37-49

Performance of Permutation Tests Using Simulated Genetic Data

Ibrahim Soumare[1], Curt Doetkott[2], Rhonda Magel[3]
[1]*Department of Mathematics and Statistics, St. Cloud State University, St. Cloud, MN 56301*
[2]*Information Technology, North Dakota State University, Fargo, ND 58108*
[3]*Department of Statistics, North Dakota State University, Fargo, ND 58108*
Corresponding Author: Rhonda Magel

ABSTRACT:*A common challenge in RNA-seq data analysis is to identify genes whose mean expression levels change across different groups of samples, or, more generally, are associated with one or more variables of interest. Such analysis is called differential expression analysis. Many tools have been developed for analyzing differential gene expression (DGE) for RNA-seq data. RNA-seq data are represented as counts. Typically, a generalized linear model with a log link and a Negative Binomial response is fit to the count data for each gene, and DE genes are identified by testing, for each gene, whether a model parameter or linear combination of model parameters is zero. We conducted a simulation study to compare the performance of our proposed Permutation test to traditional parametric methods when applied to RNA-seq data. We considered different combinations of sample sizes and underlying distributions. In this simulation study, we simulated data using Monte Carlo simulation in SAS and assessed Type I error rate, Power rate, True Positive rate and False Positive rate for each model involved. The simulation results suggest that Permutation tests are a viable alternative with a comparable power and good control of Type I error and False Positive rate*.
KEYWORDS*:*

Date of Submission: 28-01-2023 Date of Acceptance: 10-02-2023

I. INTRODUCTION

Disease statuses and biological conditions are known to be greatly impacted by differences in gene expression levels [1]. The recent rise of RNA-seq technology has now supplanted microarrays as the technology of choice for genome- wide Differential Gene Expression (DGE) experiments. As described by [1], in each experiment, messenger ribonucleic acids (mRNAs) are shattered and reverse transcribed into complementary deoxyribonucleic acid (cDNA). These short pieces of cDNA are amplified by a polymerase chain reaction and sequenced by a sequencing machine, giving a list of short sequences called reads. These reads are then mapped to the reference genome using an appropriate algorithm, telling us which region each read comes from. Finally, for a set of regions of interest on the genome, such as genes, exons, or junctions, we count the number of reads mapped unambiguously to each of them and use this count as a measure of the expression of the region.

A common challenge in RNA-seq data analysis is to identify genes whose mean expression levels change across different groups of samples, or, more generally, are associated with one or more variables of interest. Such analysis is called differential expression analysis. Differential expression analysis usually involves carrying out a significance test for each gene. Because RNA-seq data generally contain thousands of genes, differential expression analysis involves testing thousands of hypotheses.

Many tools have been developed for Analyzing DGE for RNA-seq data. RNA-seq data are represented as counts and statistical methods that try to identify differential expression – enhanced ("up-regulated") or suppressed ("down-regulated") make assumptions about the statistical properties inherent to the data and they exploit a range of normalization and analysis techniques to compute the magnitude of a DGE result and estimate its significance.

Typically, a generalized linear model with a log link and a Negative Binomial response is fit to the count data for each gene, and DE genes are identified by testing, for each gene, whether a model parameter or linear combination of model parameters is zero. It is reported that data from technical replicates can often be well characterized by the Poisson distribution, while data from biological replicates have much larger variance and Negative Binomial models seem to be more appropriate.

Estimates obtained from inferential statistical methods are generally reliable when the underlying assumptions are met. However, when the underlying assumptions of the test statistic are not met, the sampling distribution of the test statistic may deviate substantially leading to inaccurate inferences. According to [2], the tendency of researchers to prefer the use of parametric statistics has led many to propose some transformation techniques to satisfy the underlying parametric assumptions. However, others such as [3] have shown that transforming data for certain designs can be dramatically non-robust and often produce poor power properties. This

controversy calls for the need to better understand statistical procedures available to researchers given an unknown or non-normal population distribution.

Simple Permutation tests use rearrangements of the original sample to build the sampling distribution of the test statistic so make minimal assumptions about the data. For clients with modest mathematical or statistical background, Permutation tests are often more intuitive than even basic parametric tests such as the two-sample t-test. Inferential methods associated with RNA-seq data are substantially more mathematically challenging than the two-sample t-test so may be even more difficult to comprehend. According to [4], Permutation tests can be applied to continuous, ordered and categorical data, and to values that are normal, almost normal, and non-normally distributed. For almost every parametric and nonparametric test, one may obtain a distribution-free Permutation counterpart. The resulting Permutation test is usually as powerful than alternative approaches. And Permutation methods can sometimes be made to work when other statistical methods fail [4].

Permutation tests can take multiple forms. Exact Permutation tests compile all possible combinations of treatment and control data for the chosen test statistic. They are called exact because the relevant properties are specifically determined, that is an exact level of significance is determined by a significance test [5]. The approximate randomization test focuses on a random subset of all possible Permutations[6]. In situations where the number of Permutations may be overwhelming due to a large sample size, an approximate randomization test can be a viable alternative. Several researchers suggest thatPermutation and randomization tests help to rehabilitate the power of parametric tests under conditions of non-normality [7,8]. And still, others offer Permutation tests as preferred alternatives to rank-based tests, citing that rank tests are less powerful than randomization tests on scores [9].

The goal of this study is twofold. First, to compare the performance of the Permutation test to comparable parametric tests in the two-sample differential expression setting using simulated data that mimic RNA-seq data. And secondly, to investigate how the sample sizes impact the Permutation results. Various scenarios will be explored, some in which the underlying assumptions on the data are met such that the parametric tests perform well and in others where the underlying assumptions on the data for the parametric tests are violated to varying degrees. Our general expectations are that the parametric tests will usually be more powerful, but if simple Permutation tests yield reasonably close results, clients due to their more intuitive nature may prefer them. A broad study outline follows.

The core of this study will be carried out in two phases. In the first phase, we will simulate RNA-seq data assuming various underlying distributions using Monte Carlo simulation in SAS to mimic real RNA-seq datasets for relatively small sample sizes ($n_1=n_2\leq10$). For our simulated data, we will consider four distributions, namely: Normal, Poisson, Negative Binomial (NB) and Zero Inflated Poisson (ZIP) distributions and then assess Type I error and Power rate for each combination of underlying distribution and sample sizes for the fitted models considered. In the second phase, we will repeat the simulation process as described above but this time we will set twenty percent (20%) of the simulated RNA-seq data to be differentially expressed (the DE genes are obtained at 0.5σ and 1σ effect sizes). We will then assess differential expression using Permutation, two-sample t-tests, Poisson regression and Negative Binomial regression onvarying numbers of replicates. For each combination of underlying distribution (Normal, Poisson, NB) and number of replicates, we will obtain the number of differentially expressed (DE) genes for each of the fitted models (T-test, Permutation, Poisson, Negative Binomial). From the DE genes obtained, we will assess the True Positive (TP) rate and the False Positive (FP) rate and compare these rates across the fitted models.

The process just described was conducted for distributions with mean $\mu = 30$ and standard deviation $\sigma = 5$ for samples simulated from Normal distribution and Lambda=30 for samples simulated from Poisson distribution. For Negative Binomial distribution, we increased the standard deviation from 5 to 8 and kept the mean the same as that of Normal and Poisson distribution at $\mu = 30$. However, it is important to note for the Permutation distribution that when a large number of replicates is considered we did not perform a full Permutation test. The number of possiblepermutations is overwhelming for large samples. Therefore, we take a random sample of all possible permutations of the data instead. We will use B to indicate the number of random permutations selected. In this study we decided to select 5000 random permutations (B=5000). For each of these randomly selected Permutation samples we will assess DE genes for each fitted model at varying numbers of replications and compare the True Positive and False Positive rates across all the fitted models.

Although we expect the TPR rate for the parametric tests to be better, we are interested in showing how much poorer the results are using the Permutation test. For some clients, the Permutation test may be preferable due to its intuitive nature *if* the loss of power is not too great.

Additionally, the two phases as described above were carried out under a balanced Two-Sample scenario (Treatment and Control). Moreover, we also applied the scenario described in phase 1 to unbalanced sample sizes and see how the results compared to that of balanced sample sizes.

We then attempt to answer the following research questions:

- How do Type I error rate, True Positive rate and False Positive (FP) rates compare across T-test, Permutation, Poisson, and Negative Binomial when applied to RNA-Seq data?
- How does the sample size impact the Permutation test results?

II. METHODOLOGY

Methodology

Data Simulation Overview

Our main goal for conducting this simulation study is to demonstrate that simple Permutation is a valid candidate for detecting differentially expressed genes in RNA-seq data sets when compared to the traditional parametric methods. We first started in our Initial research by assessing Type I error and Power rates under various conditions. Afterward, we estimate True Positive (TP) and False Positive (FP) rates when the parametric methods' underlying distribution assumptions are met as well as when the underlying distribution assumptions are violated.

We simulated 5,000 genes with two groups each with equal sample sizes ranging from 7 to 30. We then set twenty percent (20%) of the simulated genes to be differentially expressed (the two group means are different by a defined effect size) and the other eighty percent (80%) are equally expressed (the two groups have equal means). We used SAS for this simulation and considered four main data-generating distributions namely: Normal, Poisson, Zero Inflated Poisson (ZIP) and Negative Binomial. It is noteworthy that this study was motivated by an RNA-seq data set that has features similar to zero inflated data distributions. We are simulating data from ZIP to assess the performance of Permutation tests when there is an excess of zeros in an RNA-seq dataset.

Differential Gene Expression Assessment

Whenever the observed difference or change in read counts or expression levels between the two conditions of an RNA-seq data set (assuming two groups Control and Treatment) is statistically significant, the gene is declared to be differentially expressed (DE). Therefore, it is important to find the underlying distribution of the data when fitting a parametric method to identify differentially expressed genes. In practice, researchers do not always know the statistical distribution of the data and a violation could lead to an incorrect detection.

In this present study, the focus is to investigate the differential gene expression analysis based on the Permutation test and how it compares with traditional parametric methods used for gene expression analysis. The framework for this simulation study is as follow: We simulated 5,000 genes of two groups (control and Treatment) each of size n from the same underlying distribution. For our simulated data to exhibit the features of a true RNA-seq dataset, 80% of the data is set to be equally expressed while 20% is set to be differentially expressed. For the differentially expressed genes, we considered 0.5σ and 1σ effect sizes. By 0.5σ effect size we refer to the mean difference between a pair of genes in Control and Treatment is equal to half its standard deviation. For example, suppose that we have a $\mu_C = 30$ and $\sigma_C = 5$ for the Control group; a 0.5σ effect size will correspond to $\mu_T = \mu_C - 0.5\sigma = 30 - 2.5 = 27.5$ and standard deviation $\sigma_T = 5$ for the Treatment group such that $\mu_C - \mu_T = 2.5$. In general, as the effect sizes increase it becomes easier to detect any difference in means; namely, the detection rate of the test increases as well.

To detect DE genes, we then fit the models (T-test, Poisson, Negative Binomial, Permutation test and Zero Inflated Poisson regression) to each gene and count the number of times the null hypothesis for the test below was rejected:

$$H_0: \mu_C = \mu_T$$
$$H_a: \mu_C \neq \mu_T$$

The rejection rate (rejection rate or p-value refers to how often the null hypothesis was rejected) for each underlying distribution is computed by dividing the total count ofsamples in which the null hypothesis was rejected by 5,000 (number of simulated genes). For a 5% significance level we computed the rejection rate for each combination of underlying distributions and sampling efforts.

True Positive Rate

The True Positive rate (TPR; power), also called sensitivity, is the probability that a gene that is declared to be differentially expressed is actually differentially expressed. The rate is computing by tallying the true DE genes from the list of genes declared to be DE genes by a fitted model over the total number of the simulated True DE genes (1000). The TPR was calculated as follow:

$$TPR = \frac{TP}{TP + FN}$$

Where TP is true positive, FN is false negative

False Positive Rate

AFalse Positive, also often called a Type I error in statistics, occurswhen an equally expressed (EE) gene is falsely declared as a DE gene by a fitted model. The False Positive Rateis calculated as the ratio of the number of genes wrongly classified as DE genes over the total number of actual negative events (EE genes).

$$FPR = 1 - Specificity = \frac{FP}{FP + TN}$$

Where FP is false positive, TN is true negative

We use the True Positive and False Positive rates to measure the accuracy of our fitted model. If the difference in True Detection rate is not too large, the model that minimizes the False Positive Rate may be of interest. This may vary from one field to another. For some researchers and types of studies, controlling the False Positive Rate is a very important. Therefore, a model that consistently keeps the FPR very low is preferred.

Proposed Test

In this study, our main goal is to provide evidence that the simple Permutation test is a valid and viablealternative for analyzing RNA-seq data. Permutationtests and Randomization tests are often interchangeable; however, the distinction between the two varies among the statistical community [10]. Some authors consider a randomization test to bean approximate Permutation test that takes only a random sample of all possible Permutations[10]. Others however, differentiate Permutationtests as those based on the assumption of random sampling from two identical population distributions while a randomization test is based on the assumption of random assignment of group labels [11]. We refer to these asPermutation tests regardless of: (i) whether groups are obtained by random assignment or random sampling, and (ii) whether the testsare obtained from the full set of possible permutations or from a random sample of the possible permutations[10].

We are simulating data from Normal, Poisson, Negative Binomial and Zero Inflated Poisson distributions and then fitting traditional parametric methods used for the simulated genes and then compare these results to that of the simple Permutation test. We do not expect the Permutation test to be superior; if the Permutation test yields a power that is close enough to the gold standard then that is sufficient. Particularly, we are interested in how the Permutation test compares to the traditional methods with respect to controlling the False Positive Rate while yielding a competitive True Detection Rate.

For the simple Permutation test, we defined the test statistic as follows:

- When sampling from Normal, Poisson and Negative Binomialdistrbutions

Consider our hypothesis:

$$H_0: \mu_C = \mu_T$$
$$H_a: \mu_C \neq \mu_T$$

Where μ_C and μ_T represent the means of the control and treatment groups respectively and $\delta = \mu_C - \mu_T$ is the true difference in the means. We define $\bar{C}$ and $\bar{T}$as the means obtained from the samples from the control group (C) and the treatment group (T). The test statistic is given as:

$$D = \bar{C} - \bar{T}$$

Under the null hypothesis, the expected value of D, E(D) = 0. Moreover, suppose that the test statistic from our sample is defined as $d = \bar{c} - \bar{t}$. By definition, a two-tailed p-value based on $\bar{C} - \bar{T}$ is:

$$\text{p-value} = Pr(|\bar{C} - \bar{T}| \geq |\bar{c} - \bar{t}| \mid H_0 \text{ is true}) = Pr(|D| \geq |d| \mid H_0 \text{ is true})$$
$$= Pr(D \leq -|d| \mid H_0 \text{ is true}) + Pr(D \geq |d| \mid H_0 \text{ is true})$$

For our simulation study, we sampled 5,000 genes with two groups (Control and Treatment). Unlike parametric models assuming known distributions, such as the Gaussian or Student's t, to calculate the p-value for the

Permutation test, we first build the sampling distribution for our test statistic D = $\bar{C}$ - $\bar{T}$ by aggregating all (or a sample of) possible values of the test statistic obtained by rearranging the group labels associated with the observations. For small samples, less than or equal to 10, we did a full Permutation test whereas for larger sample sizes, greater than or equal to 12, we did a partial Permutation (or randomized Permutation) of B size. Here we set B at 5000 randomly selected Permutations. The p-value is then obtained by finding the proportion of the sampling distribution of the test statistic obtained from the Permutation samples that is at least as extreme as the actual test statistic. Specifically, we first compute the test statistic from the original sample as d = $\bar{c}$ - $\bar{t}$. And then, the samplingdistribution of the test statistic d is built by computing the B Permutation test statistics $d_1, \ldots, d_B$ where $d_i = \overline{c_i} - \overline{t_i}$, and $\overline{c_i}$ and $\overline{t_i}$ are the means of the control and treatment groups respectively when labels have been reassigned according to the i^{th}Permutation, or random shuffling, of the group labels. Our two-sided Permutation test p-value is then calculated as:

$$p - value = \frac{\sum_1^B I_{(|di| \geq |d|)}}{B}$$

$$P - value = \frac{\sum_1^B I_{(di \leq -|d|)}}{B} + \frac{\sum_1^B I_{(di \geq |d|)}}{B}$$

Where I, the indicator function, is equal to 1 when the condition is true and 0 otherwise.

- When sampling from Zero Inflated Poisson distribution

For genes simulated from the Zero inflated distribution, we had to find a way to incorporate the zero in the mean estimation. We sampled 5,000 genes from a mixture of Poisson and Uniform distributions. We set the probability of success to 20% for the Uniform distribution. In other words, there is a 20% chance of observing a 1 from the Uniform distribution. If the event is 1 then we set y = 0 otherwise y = Poisson (Lambda= λ). Now our random variable y follows a modified version of the Poisson (λ) distribution known as a Zero Inflated Poisson (ZIP) distribution with a density function defined as:

$$P(Y = k) = \begin{cases} \pi + (1 - \pi)\exp(-\lambda) & \text{if } k = 0 \\ (1 - \pi)\exp(-\lambda)\frac{\lambda^k}{k!} & \text{if } k \in \{1, 2, \ldots\} \end{cases}$$

Where $0 \leq \pi \leq 1$ and $\lambda \geq 0$.

The parameter π gives the probability at the value 0. When it vanishes, ZIP (π, λ) reduces to Poisson (λ). The mean and variance of the ZIP are:

$E(Y) = (1 - \pi)\,\lambda$

$V(Y) = (1 - \pi)\,(1 + \pi\,\lambda)\,\lambda$

We can easily derive λ from the expected value of the ZIP as follow:

$$\lambda = \frac{E(Y)}{(1 - \pi)}$$

From the simulated samples we can estimate the mean of the ZIP and the parameter π. Recall that π is the probability that y is 0. Therefore, we could estimate π by dividing the frequencyof 0 by the sample size.

The test statistic for the simple Permutation test remains the same as described previously with a slight modification in the Permutation process. We quickly realized that a full permutation or a partial permutation of the control and treatment groups led to a very poor Type I error. Both Control and Treatment contain zeros at approximately 20% of their size. Shuffling the two groups could cause the data to be skewed with all the zeros or most of the zeros to be in one group and no or very few zeros in the other. After muchtrial and error, we proposed a modified Permutation method. Instead of shuffling all the observations, we held the proportion of zeros constant in each group and permuted only the non-zero observations. Once we obtained all the permuted samples from the modified Permutation process, we then computed the means using the adjusted mean formula discussed above.

The test statistic is computed in the same fashion as described above. The only difference here is the Permutation procedure. When sampling from Normal, Poisson and Negative Binomial distributions we shuffled all the data in Control and Treatment groups. However, when sampling from the Zero Inflated Poisson, we modified the Permutation procedure by keeping the proportion of zeros constant in each group and permuting only the non-zero values.

III. RESULTS

In this section, we will first cover the results obtained from the first phase of our research comparing Type I Error and power performance, and then discuss the results from the second phase detection rate assessments (the TP and FP rate comparison for each combination of underlying distributions, fitted model and sample sizes for the case of two populations scenarios).

Phase I Research Results

In our Initial study, we assessed the Type I error as well as the Power of the fitted models for each underlying distribution we considered in the case of two populations scenarios.

Type I error Assessment: $\mu_C = \mu_T$

Type I error is defined as the probability of rejecting the null hypothesis when in fact it is true. For this simulation study we set our significance level alpha at $\alpha = 0.05$ and expect the estimated Type I error to be in the neighborhood of 5%.

We sampled two random groups (Control and Treatment) from an underlying distribution using three different models (Normal, Poisson and ZIP). The estimated Type I error was obtained based on these random samples assuming equals sample sizes for both Control and Treatment. Furthermore, we set the mean of the two groups to be equal to 30 and explored different sampling effortsof5, 7, 10, 12, 15, 20 and 30. For the Normal distribution we set the variance to be equal to 25 and for the Zero Inflated Poisson we set π=0.2; the mean remains the same for all underlying distributions (mean=30).

The estimated Type I error when sampling from Normal Distribution (Normal with mean 30 and standard deviation 5 for both Control and Treatment) is summarized in Table 1 below. The results suggest that, for all combinations of sampling effort and fitted models (T test, Permutation, Poisson and NB regressions), Type I error is maintained near the stated rate of $\alpha = 0.05$. Similar results are obtained when sampling from the Poisson distribution. As shown in Table 2, Type I error is maintained near 0.05 significance level with the exception of the Permutation test which isslightly conservative for sample sizes 5 and 7.

Table 1. Normal samples - rejection rates (%) for fitted models $\mu_C = \mu_T = 30$

Sampling Efforts	Fitted Models			
	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T = 5$	4.73	4.98	5.29	5.38
$n_C = n_T = 7$	4.37	4.70	4.60	4.71
$n_C = n_T = 10$	4.36	4.50	4.48	4.71

Table 2. Poisson samples - rejection rates (%) for fitted models $\mu_C = \mu_T = 30$

Sampling Efforts	Fitted Models			
	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T = 5$	4.37	4.02	4.81	4.81
$n_C = n_T = 7$	4.55	4.18	4.86	4.95
$n_C = n_T = 10$	5.21	4.85	5.37	5.44

Tables 3, 4, 5, 6, 7 and 8 below display estimated Type I error when sampling from Normal and Poisson distributions respectively, for samples sizes 12, 15 and 20, using different numbers of random Permutationsamples. For small samples above ($n_C = n_T \leq 10$) we performed a full permutation on the simulated data set. However, when the sample size is greater than 10 ($n_C = n_T \geq 12$) Permutation becomes overwhelming. In this case we randomly selected a sample of all possible permutations of the simulated data set (referred to subsequently as a partial permutation analysis). For our study, we considered three sizes (denoted using B in a manner similar to indicating the number of bootstrap samples) of these samples of possible permutations. B = 1000, 5000 and 10,000.

For all combinations of underlying distributions, sample sizes and number of permutationsamples (B) and fitted models, as shown from table 3 to table 8, Type I error is maintained near the stated significance level of 0.05. Let us note here that for B = 10,000 and sample size of 12 and 15 (Table 8) the Permutation test was slightly conservative.

Table 3. Normal samples - rejection rates (%) for fitted models $\mu_C = \mu_T = 30$,B=1000

Sampling Efforts	Fitted Models			
	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T = 12$	4.89	5.08	4.91	5.32
$n_C = n_T = 15$	4.87	5.01	4.93	5.20

$n_C = n_T = 20$	5.19	5.36	5.22	5.55

Table 4. Poisson samples - rejection rates (%) for fitted models $\mu_C = \mu_T = 30$, B=1000

Sampling Efforts	Fitted Models			
	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T = 12$	4.90	4.67	5.05	5.13
$n_C = n_T = 15$	5.04	4.76	5.06	5.30
$n_C = n_T = 20$	5.34	5.09	5.38	5.51

Table 5. Normal samples - rejection rates (%) for fitted models $\mu_C = \mu_T = 30$, B=5000

Sampling Efforts	Fitted Models			
	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T = 12$	5.08	5.08	5.14	5.41
$n_C = n_T = 15$	5.38	5.33	5.34	5.64
$n_C = n_T = 20$	5.28	5.28	5.28	5.58

Table 6. Poisson samples - rejection rates (%) for fitted models μC = μT = 30, B=5000

Sampling Efforts	Fitted Models			
	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T = 12$	4.83	4.42	4.83	4.93
$n_C = n_T = 15$	5.16	4.82	5.21	5.43
$n_C = n_T = 20$	4.81	4.38	4.82	5.12

Table 7. Normal samples - rejection rates (%) for fitted models $\mu_C = \mu_T = 30$, B=10K

Sampling Efforts	Fitted Models			
	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T = 12$	5.19	5.39	5.31	5.57
$n_C = n_T = 15$	4.82	4.96	4.83	5.26
$n_C = n_T = 20$	5.18	5.23	5.14	5.62

Table 8. Poisson samples - rejection rates (%) for fitted models $\mu_C = \mu_T = 30$, B=10K

Sampling Efforts	Fitted Models			
	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T = 12$	4.74	4.47	4.86	4.94
$n_C = n_T = 15$	4.75	4.46	4.78	4.99
$n_C = n_T = 20$	4.80	4.52	4.79	4.83

For genes from the ZIP distribution, we only looked at sample sizes greater than 10 and B = 1000 permutationsamples. From Table 9 below we can conclude that type I error is maintained near the stated alpha value of 0.05 for all fitted mode.

Table 9. ZIP samples - rejection rates (%) for fitted models $\mu_C = \mu_T = 30$, B=1000

Sampling Efforts	Fitted Models				
	T test	Perm	ZIP	Poisson	NB
$n_C = n_T = 12$	5.10	4.78	4.90	3.43	2.10
$n_C = n_T = 15$	4.41	4.59	5.07	3.41	2.43
$n_C = n_T = 20$	5.18	4.66	4.85	4.27	3.38
$n_C = n_T = 30$	5.32	4.94	5.14	4.66	4.04

The results obtained from the Type I error assessment for each combination of sampling efforts and underlying distributions, suggest that all the models fitted (T test, Poisson regression, Negative Binomial regression, Zero Inflated Poisson regressionand Permutation Test) control the Type I errors. Therefore, all of the above models are valid candidates to test the null hypothesis that the means are equal using RNA-seq data. We will next discuss the results from the power comparison to decide whether a particular model is preferred over the rest.

Power Comparison: $\mu_C \neq \mu_T$

After ensuring that all models maintained Type I error at below or near the stated significance level of α=0.05, we then conducted a power comparison under various conditions to check whether certain models performed better than others. We considered a 0.5σ effect size for the power comparison. By 0.5σ effect size, we refer to the mean difference between a pair of genes in Control and Treatment is equal to half its standard deviation. For example, suppose that we have a μ_C = 30 and σ_C=5 for the Control group; a 0.5σ effect size will correspond to $\mu_T = \mu_C$ - 0.5σ = 30 – 2.5 =27.5 and standard deviation σ_T=5 for the Treatment group such that $\mu_T - \mu_C$ = 2.5. In general, as the effect sizes increase it becomes easier to detect any difference in means; namely, the power of the test increases as well.

- **Effect size: half sigma (0.5σ)**

The power is defined as the probability of rejecting the null hypothesis when in fact it is false. Simulating our observations from two populations with different means and setting the null hypothesis as H_0: $\mu_C = \mu_T$makes the null hypothesis false. Tallying the number of times each of the models correctly detects the difference in the two groups (rejecting H_0 since there are in fact different and dividing it by 10,000 will give us our estimated powers). This process was repeated for each of the underlying distributions.

For small sample sizes ($n_C = n_T \leq 10$), when sampling from Normal and Poisson distribution, all fitted models (T test, Permutation, Poisson and NB regression) yielded a comparable power rate as shown in Tables 10-11.

Table 10. Normal samples - rejection rates (%) for fitted models $\mu_C \neq \mu_T$ (Effect size=0.5σ**,** μ=30)

Sampling Efforts	Fitted Models			
	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T$=5	9.89	10.45	10.68	10.88
$n_C = n_T$=7	13.47	14.02	14.04	14.32
$n_C = n_T$=10	18.20	18.55	18.37	18.98

Table 11. Poisson samples - rejection rates (%) for fitted models $\mu_C \neq \mu_T$ (Effect size=0.5σ**,** μ=30)

Sampling Efforts	Fitted Models			
	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T$=5	9.96	9.48	11.00	11.12
$n_C = n_T$=7	14.19	13.47	14.73	14.88
$n_C = n_T$=10	18.71	17.89	18.99	18.34

For samples sizes $n_C = n_T \geq$12, we performed a partial permutation analysis for the Permutation test (B=1000, 5000, 10000). When sampling from Normal and Poisson distribution, all fitted models (T test, Permutation, Poisson and NB regression) yielded a comparable power rate at all permutation sample sizes level as shown in Tables 12-17.

Table 12. Normal samples - rejection rates (%) for fitted models $\mu_C \neq \mu_T$ (Effect size=0.5σ**,** μ=30), B=1000

Sampling Efforts	Fitted Models			
	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T$=12	21.19	21.43	21.13	21.70
$n_C = n_T$=15	26.10	26.42	26.08	26.78
$n_C = n_T$=20	32.72	33.04	32.73	33.70

Table 13. Poisson samples - rejection rates (%) for fitted models $\mu_C \neq \mu_T$ (Effect size=0.5σ**,** μ=30), B=1000

Sampling Efforts	Fitted Models			
	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T$=12	22.43	21.57	22.73	23.02
$n_C = n_T$=15	27.49	26.63	27.73	28.01
$n_C = n_T$=20	35.09	34.36	35.17	35.52

Table 14. Normal samples - rejection rates (%) for fitted models $\mu_C \neq \mu_T$ (Effect size=0.5σ**,** μ=30), B=5000

Sampling Efforts	Fitted Models			
	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T$=12	20.99	21.35	21.16	21.75

$n_C = n_T = 15$	26.45	26.63	26.43	27.15
$n_C = n_T = 20$	34.11	34.49	34.05	35.26

Table 15. Poisson samples - rejection rates (%) for fitted models $\mu_C \neq \mu_T$ (Effect size=0.5σ, μ=30), B=5000

Sampling	Fitted Models			
Efforts	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T = 12$	22.37	21.59	22.70	22.87
$n_C = n_T = 15$	27.05	26.15	27.13	27.61
$n_C = n_T = 20$	34.89	33.98	34.97	35.49

Table 16. Normal samples - rejection rates (%) for fitted models $\mu_C \neq \mu_T$ (Effect size=0.5σ, μ=30), B=10K

Sampling	Fitted Models			
Efforts	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T = 12$	21.88	22.11	21.93	22.47
$n_C = n_T = 15$	25.59	25.88	25.65	26.49
$n_C = n_T = 20$	33.43	33.45	33.30	34.23

Table 17. Poisson samples - rejection rates (%) for fitted models $\mu_C \neq \mu_T$ (Effect size=0.5σ, μ=30), B=10000

Sampling	Fitted Models			
Efforts	T test	Permutation	Poisson	Negative Binomial
$n_C = n_T = 12$	22.28	21.49	22.46	22.72
$n_C = n_T = 15$	26.47	25.59	26.73	27.20
$n_C = n_T = 20$	35.04	33.78	35.11	35.28

Table 18 summarizes the power rate from all fitted models (T test, Permutation, Poisson and NB regression) when we sample from a ZIP distribution. We can see that the Permutation test and the ZIP model yielded comparable power whereas T test, Poisson, and Negative Binomial displayed a poor power.

Table 18. ZIP Samples - rejection rates (%) for fitted models $\mu_C \neq \mu_T$ (Effect size=0.5σ, μ=30, π=0.2), B=1000

Sampling	Fitted Models				
Efforts	T test	Perm	ZIP	Poisson	NB
$n_C = n_T = 12$	7.65	17.39	20.36	6.09	4.56
$n_C = n_T = 15$	7.63	21.28	23.56	6.48	5.36
$n_C = n_T = 20$	8.80	28.00	30.29	7.78	6.91
$n_C = n_T = 30$	10.71	40.54	42.18	9.99	9.26

Differential Expressed Genes Assessment

Whenever a fitted model detected a significant difference between the mean of the two conditions of our simulated RNA-seq data (Control vs Treatment), the gene wasdeclared to be differentially expressed. The detection rate wasthen obtained by tallying the total number of DE genes over the simulation size (5000).

We sampled two random groups (Control and Treatment) from each underlying distribution using three different models (Normal, Poisson and Negative Binomial). The estimated detection rate, True Positive and False Positive rate were obtained based on these random samples under two different designs: Balanced (equal sample sizes for both Control and Treatment; n1=n2) and Unbalanced (unequal sample sizes between Control and Treatment; n1≠n2). Furthermore, we set 80% of the simulated data to be equally expressed (μ1=μ2=30) while the other 20% are set to be DE genes ($\mu_1 \neq \mu_2$) with a 1σ effect size.

Various sample sizes were considered from 5 to 30. Our early results suggested that Permutation tests suffer from the granularity issue with relatively small sample sizes [12]. Note that the smallest possible value for the p-value is 1/N where N represents the number of permutation possible. We refer to 1/N as the granularity limit. For large sample sizes, since N is too large for the Permutation test to be computationally convenient, we take a partial permutationsample of size B for our simulation and therefore the smallest the p-value can beis 1/B. it is important to note that 1/B can be much larger than the permutation limit 1/N. To obtain a small p value, a larger number for B is required. Due to thisissue, the results are very poor compared to the parametric methods regardless of the design. For the rest of the study we will focus on the following sample sizes: 10, 15, 20, 25 and 30.

Balanced Design: $n_1=n_2$

When sampling from a Normal distribution (Normal with mean μ=30 and standard deviation σ=5) with a 1σ effect size, all fitted models (T test, Permutation, Poisson and NB regression) yielded a comparable detection rate as shown in Table 19. As the sample size increases, the number of genes declared to be differentially expressed increases as well. However, we are interested in the quality of the model to detect the True differentially expressed genes with minimal error. Negative Binomial regressions tend to detect more DE genes compared to the other models but the difference is very small. For instance, for sample sizes $n_C= n_T=25$, NB regression correctly detected 924 genes out of 1000, Poisson regression was second with 923 and the Permutation test was third with 922 genes. The difference is about two extra genes. It appears that all fitted models perform relatively well when it comes to detecting True DE genes when the effect size is relatively large.

The last column in Table 19 provides the False Positive rate which indicates the proportion of genes that were incorrectly declared to be DE genes. Overall, Permutationtests and Poisson regression consistently had lower False Positive rates for all sample sizes compared to Negative Binomial regression.Permutation and Poisson regression were somewhat similar, with Poisson regression having slightly lower FPR when the sample sizes are 15 and 20. Permutation however, had a lower FP rate when the sample sizes are 25 and 30. So there is not a consistent signal on whether Permutationtests or Poisson regression keeps lower FP rates. But clearly both controlled FP rate better than the Negative Binomial regression. The t-test on the other hand had the overall lowest FP rate. This is expected as the underlying distribution is Normal but the t-test also had the smallest True Positive rate (slightly lower than the others, which is surprising as we expected the t-test to be better when sampling from a Normal distribution).

Table 19. Normal samples - detection rates (%) for fitted models

Fitted Models	Sampling Efforts	Actual	Detected	TDE Gene	TPR(%)	FPR (%)
T-test	$n_C= n_T=10$	1000	722	546	54.60	4.40
	$n_C= n_T=15$	1000	968	763	76.30	5.13
	$n_C= n_T=20$	1000	1078	862	86.20	5.40
	$n_C= n_T=25$	1000	1119	921	92.10	4.95
	$n_C= n_T=30$	1000	1149	968	96.80	4.53
Permutation	$n_C= n_T=10$	1000	733	551	55.10	4.55
	$n_C= n_T=15$	1000	977	765	76.50	5.30
	$n_C= n_T=20$	1000	1080	866	86.60	5.35
	$n_C= n_T=25$	1000	1121	922	92.20	4.98
	$n_C= n_T=30$	1000	1146	967	96.70	4.48
Poisson	$n_C= n_T=10$	1000	725	543	54.30	4.55
	$n_C= n_T=15$	1000	968	761	76.10	5.18
	$n_C= n_T=20$	1000	1076	861	86.10	5.38
	$n_C= n_T=25$	1000	1124	923	92.30	5.03
	$n_C= n_T=30$	1000	1149	967	96.70	4.55
Negative Binomial	$n_C= n_T=10$	1000	748	557	55.70	4.78
	$n_C= n_T=15$	1000	1001	766	76.60	5.88
	$n_C= n_T=20$	1000	1091	864	86.40	5.68
	$n_C= n_T=25$	1000	1138	924	92.40	5.35
	$n_C= n_T=30$	1000	1180	966	96.60	5.35

Table 20 summarizes the simulation results when the underlying distribution is Poisson (Poisson with mean lambda=30). We see a similar pattern as with normal distribution samples. The difference in true DE genes is very small across fitted models with Poisson and Negative Binomial regression closely detecting about the same number. As the sample sizes increase, the models detected more True DE genes and the Permutation test becomes very close in True detection rate to Poisson and Negative Binomial regression. The performance of Permutationtests is very satisfying and competitive withits parametric counterparts. Interestingly, Permutationtests consistently had the lowest False Positive rate across all sample sizes with T-test second. This is in line with our early studies that showed that Permutation does a better job at controlling False Positive rates - lower than Poisson and Negative Binomial regression.

Table 20. Poisson samples - detection rates (%) for fitted models

Fitted Models	Sampling Efforts	Actual	Detected	True DE Gene	TPR (%)	FPR (%)
	$n_C = n_T = 10$	1000	768	578	87.80	4.75
	$n_C = n_T = 15$	1000	987	778	77.80	5.23
T-test	$n_C = n_T = 20$	1000	1089	888	88.80	4.93
	$n_C = n_T = 25$	1000	1167	958	95.80	5.23
	$n_C = n_T = 30$	1000	1158	976	97.60	4.55
	$n_C = n_T = 10$	1000	749	569	56.90	4.50
	$n_C = n_T = 15$	1000	959	767	76.70	4.80
Permutation	$n_C = n_T = 20$	1000	1078	888	88.80	4.75
	$n_C = n_T = 25$	1000	1152	956	95.60	4.90
	$n_C = n_T = 30$	1000	1151	974	97.40	4.43
	$n_C = n_T = 10$	1000	780	582	58.20	4.95
	$n_C = n_T = 15$	1000	990	775	77.50	5.38
Poisson	$n_C = n_T = 20$	1000	1090	891	89.10	4.98
	$n_C = n_T = 25$	1000	1170	959	95.90	5.28
	$n_C = n_T = 30$	1000	1162	976	97.60	4.65
	$n_C = n_T = 10$	1000	781	581	58.10	5.00
	$n_C = n_T = 15$	1000	988	775	77.50	5.33
Negative Binomial	$n_C = n_T = 20$	1000	1100	893	89.30	5.18
	$n_C = n_T = 25$	1000	1173	955	95.50	5.45
	$n_C = n_T = 30$	1000	1165	976	97.60	4.73

When we sample from Negative Binomial distribution (NB with mean $\mu=30$ and standard deviation $\sigma=8$) we see similar results as Poisson samples discussed above. As exhibited in Table 21 below, all fitted models yielded comparable True DE genes detection rates.Permutationtests however consistently kept the False Positive rate lower across all sample sizes. Negative Binomial regression tends to have slightly a higher False Positive rate probably due to a larger variation in the data. Thus Permutationtests not only appear to be competitive when compared to Poisson and Negative Binomial regression but most importantly it consistently had a better control of the False Positive rate for nearly all combinations of sample sizes and underlying distribution. A few exceptions occurred with Normal data where we saw Poisson regression having slightly lower FPR when the sample sizes were 15 and 20.

Table 21. Negative Binomial samples - detection rates (%) for fitted models

Fitted Models	Sampling Efforts	Actual	Detected	True DE Gene	TPR (%)	FPR (%)
	$n_C = n_T = 10$	1000	874	680	68.00	4.85
	$n_C = n_T = 15$	1000	1042	845	84.50	4.93
T-test	$n_C = n_T = 20$	1000	1143	937	93.70	5.08
	$n_C = n_T = 25$	1000	1164	969	96.90	4.88
	$n_C = n_T = 30$	1000	1205	989	98.90	5.40
	$n_C = n_T = 10$	1000	869	676	67.60	4.83
	$n_C = n_T = 15$	1000	1032	841	84.10	4.78
Permutation	$n_C = n_T = 20$	1000	1127	937	93.70	4.75
	$n_C = n_T = 25$	1000	1155	967	96.70	4.70
	$n_C = n_T = 30$	1000	1193	989	98.90	5.10
	$n_C = n_T = 10$	1000	881	677	67.70	5.10
	$n_C = n_T = 15$	1000	1036	834	83.40	5.05
Poisson	$n_C = n_T = 20$	1000	1143	939	93.90	5.10
	$n_C = n_T = 25$	1000	1167	967	96.70	5.00
	$n_C = n_T = 30$	1000	1203	989	98.90	5.35
	$n_C = n_T = 10$	1000	882	674	67.40	5.20
	$n_C = n_T = 15$	1000	1039	834	83.40	5.13
Negative Binomial	$n_C = n_T = 20$	1000	1143	937	93.70	5.15
	$n_C = n_T = 25$	1000	1168	966	96.60	5.05
	$n_C = n_T = 30$	1000	1204	989	98.90	5.38

Unbalanced Design: $n_1 \neq n_2$

Given the competitive performance of the Permutation test and its capability to control False Positive rate in the balanced design scenario, we decided to run a few unbalanced data sets and assess Permutationtest performance compared to Poisson and Negative Binomial regression. The results obtained are summarized in Tables 22, 23, 24 when sampling from Normal, Poisson and Negative Binomial distributions respectively. We see similar trends as for the balanced scenario. All models exhibited comparable True DE genes detection rate. Permutationtests kept the False Positive the lowest overall.

Table 22. Normal samples - detection rates (%) for fitted models

Fitted Models	Sampling Efforts	Actual	Detected	True DE Gene	TPR (%)	FPR (%)
T-test	n_C=15 n_T=10	1000	806	621	62.10	4.63
	n_C=20 n_T=10	1000	881	686	68.60	4.88
	n_C=20 n_T=15	1000	989	793	79.30	4.90
	n_C=30 n_T=15	1000	1081	872	87.20	5.23
Permutation	n_C=15 n_T=10	1000	811	628	62.80	4.58
	n_C=20 n_T=10	1000	894	707	70.70	4.68
	n_C=20 n_T=15	1000	989	794	79.40	4.88
	n_C=30 n_T=15	1000	1075	880	88.00	4.88
Poisson	n_C=15 n_T=10	1000	808	628	62.80	4.50
	n_C=20 n_T=10	1000	889	704	70.40	4.63
	n_C=20 n_T=15	1000	985	795	79.50	4.75
	n_C=30 n_T=15	1000	1074	877	87.70	4.93
Negative Binomial	n_C=15 n_T=10	1000	822	629	62.90	4.83
	n_C=20 n_T=10	1000	904	710	71.00	4.85
	n_C=20 n_T=15	1000	1002	796	79.60	5.15
	n_C=30 n_T=15	1000	1094	876	87.60	5.45

Table 23. Poisson samples - detection rates (%) for fitted models

Fitted Models	Sampling Efforts	Actual	Detected	True DE Gene	TPR (%)	FPR (%)
T-test	n_C=15 n_T=10	1000	929	723	72.30	5.15
	n_C=20 n_T=10	1000	1047	848	84.80	4.98
	n_C=30 n_T=15	1000	1086	892	89.20	4.85
Permutation	n_C=15 n_T=10	1000	900	706	70.60	4.85
	n_C=20 n_T=10	1000	1038	834	83.40	5.10
	n_C=30 n_T=15	1000	1070	893	89.30	4.43
Poisson	n_C=15 n_T=10	1000	910	711	71.10	4.98
	n_C=20 n_T=10	1000	1049	842	84.20	5.18
	n_C=30 n_T=15	1000	1076	894	89.40	4.55
Negative Binomial	n_C=15 n_T=10	1000	914	711	71.10	5.08
	n_C=20 n_T=10	1000	1049	834	83.40	5.38
	n_C=30 n_T=15	1000	1086	897	89.70	4.73

Table 24. Negative Binomial samples - detection rates (%) for fitted models

Fitted Models	Sampling Efforts	Actual	Detected	True DE Gene	TPR (%)	FPR (%)
T-test	n_C=15 n_T=10	1000	942	739	73.90	5.08
	n_C=20 n_T=15	1000	1071	885	88.50	4.65
	n_C=30 n_T=15	1000	1142	925	92.50	5.43
Permutation	n_C=15 n_T=10	1000	950	750	75.00	5.00
	n_C=20 n_T=15	1000	1069	890	89.00	4.48
	n_C=30 n_T=15	1000	1131	931	93.10	5.00
Poisson	n_C=15 n_T=10	1000	948	743	74.30	5.13
	n_C=20 n_T=15	1000	1073	886	88.60	4.68

	n_C=30 n_T=15	1000	1145	932	93.20	5.33
	n_C=15 n_T=10	1000	959	749	74.90	5.25
Negative Binomial	n_C=20 n_T=15	1000	1079	889	88.90	4.75
	n_C=30 n_T=15	1000	1149	934	93.40	5.38

IV. CONCLUSION

Our study using Monte Carlo simulation suggest that the Permutation test is a valid competitive model for analyzing RNA-seq data. The results are consistent for both balanced and unbalanced designs.Not only did Permutationtests yield similar True positive rates as Poisson and Negative Binomial regression, but theyconsistently controlled the False Positive rate better than parametric counterparts.

RNA-seq data are generally assumed to follow either a Poisson or Negative Binomial distribution. Traditional models developed for analyzing such data assume these distributions without providing a way to check whether the underlying assumptions are met. Our simulation results provide evidence that for both Poisson and Negative Binomial samples, the Permutation test is robust and offers good control of the False Positive rate.

REFERENCES

[1]. Li, J., &Tibshirani, R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Statistical methods in medical research, 22(5), 519-536.

[2]. Zimmerman, D. W. &Zumbo, B.D. (1990b). The relative power of the Wilcoxon-Mann-Whitney test and Student t-test under simple bounded transformations. The Journal of General Psychology, 117(4), 425-436

[3]. Sawilowsky, S.S., Blair, R. C., & Higgins, J. J. (1989). An investigation of the type 1 error and power properties of the rank transformation procedure in factorial ANOVA. Journal of Educational Statistics, 14(3), 255-267.

[4]. Good, P. (1994). Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. New York, NY: Springer-Verlag.

[5]. Walsh, E. O. (1968). An introduction to biochemistry. London, England: English Universities.

[6]. Mielke, P. W. & Berry, K. J. (2001). Permutation methods: A distance function approach. New York, NY: Springer.

[7]. Potvin, C. &Roff, D. (1993). Distribution-free and robust statistical methods: Viable alternatives to parametric statistics? Ecology, 74(6), 1617-1628.

[8]. Edgington, E. S. (1995). Randomization Tests. (3rd ed). New York, NY: Marcel Dekker.

[9]. May, R. B., Masson, E.J., & Hunter, M. A. (1989). Randomization tests: Viable alternatives to normal curve tests. Behavior Research Methods, Instruments, & Computers, 21(4), 482-483.

[10]. Christensen, W. F., & Zabriskie, B. N. (2021). When your Permutation test is doomed to fail. The American Statistician, 1-11.

[11]. Onghena, P. (2018). Randomization tests or Permutation tests? A historical and terminological clarification. In: Randomization, masking, and allocation concealment, pg. 209-228. Boca Raton, FL: CRC Press

[12]. He, H. Y., Basu, K., Zhao, Q., & Owen, A. B. (2019). Permutation $ p $-value approximation via generalized Stolarsky invariance. The Annals of Statistics, 47(1), 583-611.