

RESEARCH ARTICLE

# A comprehensive simulation study on classification of RNA-Seq data

Gökmen Zararsız<sup>1,2\*</sup>, Dincer Goksuluk<sup>1,3</sup>, Selcuk Korkmaz<sup>1,3</sup>, Vahap Eldem<sup>4</sup>, Gozde Erturk Zararsız<sup>2</sup>, Izzet Parug Duru<sup>5</sup>, Ahmet Ozturk<sup>2</sup>

**1** Turcosa Analytics Solutions Ltd Co, Erciyes Teknopark, 38039, Kayseri, Turkey, **2** Department of Biostatistics, Erciyes University, Kayseri, Turkey, **3** Department of Biostatistics, Hacettepe University, Ankara, Turkey, **4** Department of Biology, Istanbul University, Istanbul, Turkey, **5** Department of Physics, Marmara University, Istanbul, Turkey

\* [gokmenzararsiz@erciyes.edu.tr](mailto:gokmenzararsiz@erciyes.edu.tr)



**OPEN ACCESS**

**Citation:** Zararsız G, Goksuluk D, Korkmaz S, Eldem V, Zararsız GE, Duru IP, et al. (2017) A comprehensive simulation study on classification of RNA-Seq data. PLoS ONE 12(8): e0182507. <https://doi.org/10.1371/journal.pone.0182507>

**Editor:** Christian Schönbach, Kumamoto University, JAPAN

**Received:** March 18, 2017

**Accepted:** July 19, 2017

**Published:** August 23, 2017

**Copyright:** © 2017 Zararsız et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** We used both simulated and real datasets. Real datasets are publicly available and only used for performance comparison of the classifiers. We confirm that these data are not owned by the authors of this study. Cervical cancer dataset is from [30] study whose authors may be contacted at [dwitten@u.washington.edu](mailto:dwitten@u.washington.edu). Alzheimer dataset is from [31] study whose authors may be contacted at [andreas.keller@ccb.uni-saarland.de](mailto:andreas.keller@ccb.uni-saarland.de). Lung and renal cell cancer datasets are available from the TCGA database [32] whose authors may be contacted at [saleem@informatik.uni-leipzig.de](mailto:saleem@informatik.uni-leipzig.de). All simulation

## Abstract

RNA sequencing (RNA-Seq) is a powerful technique for the gene-expression profiling of organisms that uses the capabilities of next-generation sequencing technologies. Developing gene-expression-based classification algorithms is an emerging powerful method for diagnosis, disease classification and monitoring at molecular level, as well as providing potential markers of diseases. Most of the statistical methods proposed for the classification of gene-expression data are either based on a continuous scale (eg. microarray data) or require a normal distribution assumption. Hence, these methods cannot be directly applied to RNA-Seq data since they violate both data structure and distributional assumptions. However, it is possible to apply these algorithms with appropriate modifications to RNA-Seq data. One way is to develop count-based classifiers, such as Poisson linear discriminant analysis and negative binomial linear discriminant analysis. Another way is to bring the data closer to microarrays and apply microarray-based classifiers. In this study, we compared several classifiers including PLDA with and without power transformation, NBLDA, single SVM, bagging SVM (bagSVM), classification and regression trees (CART), and random forests (RF). We also examined the effect of several parameters such as overdispersion, sample size, number of genes, number of classes, differential-expression rate, and the transformation method on model performances. A comprehensive simulation study is conducted and the results are compared with the results of two miRNA and two mRNA experimental datasets. The results revealed that increasing the sample size, differential-expression rate and decreasing the dispersion parameter and number of groups lead to an increase in classification accuracy. Similar with differential-expression studies, the classification of RNA-Seq data requires careful attention when handling data overdispersion. We conclude that, as a count-based classifier, the power transformed PLDA and, as a microarray-based classifier, vst or rlog transformed RF and SVM classifiers may be a good choice for classification. An R/BIOCONDUCTOR package, MLSeq, is freely available at <https://www.bioconductor.org/packages/release/bioc/html/MLSeq.html>.

process is conducted from the authors and available in [S3 File](#).

**Funding:** This work was supported by the Research Fund of Erciyes University [TDK-2015-5468]. The authors declare that they have no competing interests.

**Competing interests:** The authors have declared that no competing interests exist.

**Abbreviations:** RNA, Ribonucleic acid; NGS, Next-generation sequencing; PLDA, Poisson linear discriminant analysis; NBLDA, Negative binomial linear discriminant analysis; DE, Differential expression; vst, Variance stabilizing transformation; rlog, Regularized logarithmic transformation; SVM, Support vector machines; bagSVM, Bagging support vector machines; CART, Classification and regression trees; RF, Random forests; BWA, Burrows-Wheeler algorithm; TMM, Trimmed mean of M values; RPKM, Reads per kilobase per million mapped reads; NB, Negative binomial; ADAS-Cog, Alzheimer Disease Assessment Scale-cognitive subscale; WMS, Wechsler Memory Scale; MMSE, Mini-Mental State Exam; CDR, Clinical Dementia Rating; RCC, Renal cell cancer; TCGA, The Cancer Genome Atlas; KIRP, Kidney renal papillary cell; KIRC, Kidney renal clear cell; KICH, Kidney chromophobe carcinomas; LUAD, lung adenocarcinoma; LUSC, lung squamous cell with carcinoma.

## Introduction

Transcriptome sequencing (RNA-Seq), with the advent of high-throughput NGS technologies, has become a popular experimental approach for generating a comprehensive catalog of protein-coding genes and non-coding RNAs and examining the transcriptional activity of genomes. Furthermore, RNA-Seq is a promising tool with a remarkably wide range of applications such that (i) discovering novel transcripts, (ii) detecting/quantifying the spliced isoforms, (iii) fusion detection and (iv) revealing sequence variations (e.g. SNPs, indels) [1]. In addition, beyond these common applications, RNA-Seq can be a method of choice for gene-expression-based classification to identify the significant transcripts, distinguishing biological samples and predicting the outcomes from large-scale gene-expression data which can be generated in a single run. This classification is widely used in medicine for diagnostic purposes and refers to the detection of a small subset of genes that achieves the maximal predictive performance. These genes are used afterwards for the classification of new observations into one of the disease classes (or tumor classes, cancer subtypes, cancer stage, etc.).

Microarray-based gene-expression classification has been widely used during the last decades. Recently, RNA-Seq replaced microarrays as the technology of choice in quantifying gene expression due to certain advantages such as providing less noisy data, detecting novel transcripts and isoforms or not requiring prearranged transcripts of interest [2–5]. Although microarray and RNA-Seq technologies can be used for measuring the expression levels of genes, there are differences in the resulting gene-expression data. Microarray technology produces continuous data while it is obtained in discrete scale from RNA-Seq technology, which are related with the abundance of mRNA transcripts [6]. Hence, the algorithms which are proposed for microarray-based gene-expression data cannot be directly applied to RNA-Seq data since they violate both data structure and distributional assumptions. In addition, RNA-Seq generates gene-expression data with overdispersion where the variance exceeds the mean [7]. One should take overdispersion into account since it has a significant effect on model performances. Various studies have been conducted to deal with the overdispersion problem for the differential-expression (DE) analysis of RNA-Seq data [8–12].

Several alternatives have been proposed for the classification and clustering of RNA-Seq data. One alternative, perhaps the preferable option, is to use discrete probability distributions (e.g. Poisson, negative binomial) for both classification and clustering tasks. Witten et al. [6] proposed sparse Poisson linear discriminant analysis (PLDA) by extending the popular microarray classifier called the nearest shrunken centroids algorithm to discrete RNA-Seq data. The authors also suggested applying a power transformation within PLDA algorithm in order to handle overdispersion problem. Dong et al. [13], on the other hand, proposed negative binomial linear discriminant analysis (NBLDA) by extending Poisson distribution to negative binomial distribution. There are few methods based on discrete distributions compared to those on continuous distributions. Hence, another choice may be to use some transformation approaches (e.g. vst: variance stabilizing transformation, or rlog: regularized logarithmic transformation) to bring RNA-seq samples closer to microarrays and apply microarray-based algorithms for classification applications [7–9].

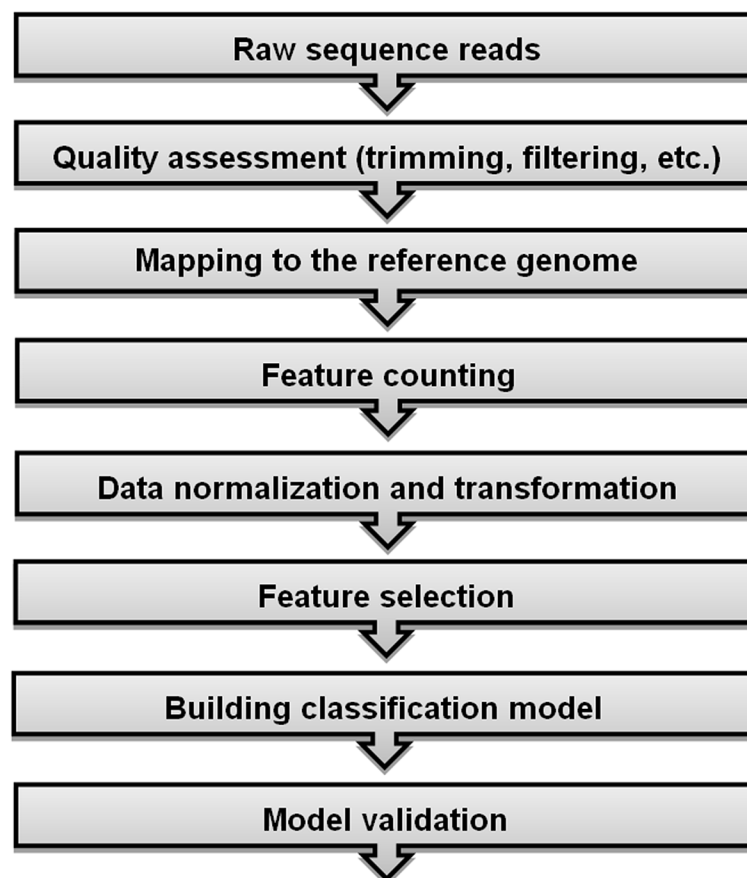
In this study, we applied algorithms based on both discrete and continuous distributions to RNA-Seq data. The NBLDA and PLDA are applied to discrete gene-expression data (i.e. no transformation on counts), while support vector machines (SVM), bagging support vector machines (bagSVM), random forests (RF) and classification and regression trees (CART) are applied to transformed gene-expression data. A comprehensive simulation study is conducted to measure the effect of several parameters on model performances, such as overdispersion, sample size, number of genes, number of classes, DE rate and the transformation method.

Four publicly available gene-expression datasets were also analyzed and the results were compared to the simulation results. An R/BIOCONDUCTOR package, called MLSeq, is developed to analyze RNA-Seq data using the proposed algorithms in this paper.

## Materials and methods

### A workflow for RNA-Seq classification

Providing a comprehensive and easy-to-understand workflow for RNA-Seq studies and its related algorithms helps researchers to find out the background of such studies. We outlined the count-based classification pipeline for RNA-Seq data in Fig 1 for providing a quick snapshot view of handling large-scale transcriptome data and establishing robust inferences by using well-known computer-aided learning algorithms. NGS platforms generate millions of raw sequence reads along with quality scores which correspond to each base-call. The very first step in RNA-Seq data analysis is to assess the quality of the sequenced data for further analysis. A number of pre-processing steps such as removal of the low-quality sequences, exclusion of the poor-quality reads with more than five unknown bases and trimming the sequencing adapters and primers should be taken into consideration to obtain a clean and ready to use RNA-Seq data for downstream analysis. Several tools/packages such as FASTQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), HTSeq [14], R ShortRead package [15], PRINSEQ (<http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi>), FASTX Toolkit ([http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)) and QTrim [16] are available for quality



**Fig 1. RNA-Seq classification workflow.**

<https://doi.org/10.1371/journal.pone.0182507.g001>

assessment and filtering. Following pre-processing steps, high-quality reads are aligned to a reference genome or transcriptome. The number of reads mapped to the reference genome is reported to be linearly related to transcript abundance [6]. Hence, transcript quantification, which are calculated from the total number of mapped reads, is a prerequisite for further analysis. Splice-aware aligners such as Tophat2 [17], MapSplice [18] or Star [19] can be preferred rather than unspliced aligners (BWA, Bowtie, etc.) for aligning short reads. Number of mapped reads to each transcript is counted after the alignment process is completed. This count matrix can be accomplished by using several tools such as HTSeq [14], bedtools [20] and FeatureCounts [21]. The mapped counts are used as the expression levels of corresponding genes and the obtained count matrix is transferred to following steps for downstream analysis. However, these counts cannot be directly used for further analysis since there exists between sample differences in the count matrix. One should normalize the counts to adjust between-sample differences using one of the proposed normalization techniques. Although there are several methods in the literature, there is no standard or state-of-art method for the normalization task. Some of the proposed methods, which are popular and frequently used in RNA-Seq studies, are the deseq median ratio [8], trimmed mean of M values (TMM) [22], reads per kilobase per million mapped reads (RPKM) [23] and quantile [24] methods. The normalized gene-expression data can be used for classification and/or clustering tasks. As previously mentioned, there are two strategies for modeling RNA-seq data. Firstly, normalized counts can be directly modeled using algorithms based on discrete distributions such as the PLDA [6] and NBLDA [13]. Secondly, the counts are transformed to a continuous scale and microarray-based classification and clustering algorithms are performed on the transformed gene-expression data. Some of the popular transformation methods are vst [8], rlog [10] and voom [25]. Apart from these approaches, power transformation is considered to decrease the dispersion of data before applying the PLDA classifier [6].

The crucial steps of classification can be considered as feature selection, model fitting and validation. RNA-Seq data has a large dimensions, i.e there are a large number of genes in the count data. In the feature selection step, we aim to work with an optimal subset of data in order to overcome the curse of dimensionality. This process is also crucial to reduce the computational cost, decrease noise, improve the accuracy for classification of phenotypes and work with more interpretable features to better understand the domain [26]. As the number of variables increases, the model becomes more complex and the classification results are more likely to be over/under-estimated. Various feature selection methods have been reviewed in details and compared in [27]. After an optimal set of genes is selected, the next step is fitting the best classification model to the RNA-Seq data. Several machine-learning algorithms are applied to training data to find the best model which best discriminates cases among classes. Next, the trained model can be used to predict the class memberships of new biological samples. Some of the commonly used classifiers are SVM, RF and other tree-based classifiers, artificial neural networks and k-nearest neighbors.

In many real life problems, it is possible to find classification algorithms which perform well and perfectly classify training samples. However, it may perform poorly when classifying new samples. This problem is called overfitting/overestimation as previously mentioned. This problem arises during the model training steps. In order to overcome overfitting problem and obtain generalized results, the appropriate model is selected by validating the selected model on independent test samples. Holdout, *k*-fold cross-validation, leave-one-out cross-validation and bootstrapping are among the recommended approaches for model validation.

## Simulation study

**Simulation setup.** A comprehensive simulation study is conducted to investigate the effect of several parameters. Gene-expression data are simulated under 864 different scenarios

using a negative binomial model as follows:

$$X_{ij}|y_i = k \sim NB(s_i g_j d_{kj}, \phi) \tag{1}$$

where,  $g_j$  is the total number of mapped counts per gene (i.e. gene total),  $s_i$  is the number of mapped counts per sample,  $d_{kj}$  is the differential-expression parameter of the  $j^{th}$  gene between classes  $k$  and  $\phi$  is the dispersion parameter. The datasets contain all possible combinations of the following:

- different dispersion parameters such as  $\phi = 0.01$  (very slightly overdispersed),  $\phi = 0.1$  (substantially overdispersed), and  $\phi = 1$  (highly overdispersed);
- number of biological samples ( $n$ ) changing as 40, 60, 80, 100;
- number of differentially-expressed genes ( $p'$ ) as 25, 50, 75, 100;
- differentially expressed gene rates as ( $d_{kj}$ ) 1%, 5% and 10%;
- number of classes ( $k$ ) as 2, 3 and 4;
- method of transformation as *rlog* and *vst*.

In the simulation setup,  $s_i$  and  $g_j$  are distributed identically and independently. Simulated datasets are generated using the CountDataSet function in the PoiClaClu package of R software [28] and manipulated based on the details given above. The seed number for random number generation is set to '10072013' in all analysis steps.

**Evaluation process.** All datasets are initially simulated for  $p = 10,000$  genes. Next, the data are split into training (70%) and test sets (30%). All the model building processes are performed on training datasets and model performances are evaluated in test sets.

**Size factor estimation.** The size factors are estimated using deseq median ratio approach [10]. Let  $x_{ij}$  the mapped read counts to  $j^{th}$  gene for  $i^{th}$  sample. Size factor of the  $i^{th}$  sample ( $\hat{s}_i$ ) can be estimated as below:

$$m_i = median_i \left\{ \frac{x_{ij}}{\left(\prod_{i=1}^n x_{ij}\right)^{1/n}} \right\} \tag{2}$$

$$\hat{s}_i = \frac{m_i}{\sum_{i=1}^n m_i} \tag{3}$$

Size factor of a test sample can be estimated using the same parameters as for the training datasets. In more detail, the size factors of the test datasets are calculated based on the geometric means of the training data. Therefore, we guarantee that the training and test datasets are in the same scale and homoscedastic to each other. Let,  $x_*$  is a count vector of a new test observation, whose class label  $y_*$  will be predicted. Size factor of the test sample ( $\hat{s}_*$ ) is estimated as follows:

$$m_* = median_g \left\{ \frac{x_{g^*}}{\left(\prod_{i=1}^n x_{gi}\right)^{1/n}} \right\} \tag{4}$$

$$\hat{s}_* = \frac{m_*}{\sum_{i=1}^n m_i} \tag{5}$$

**Dispersion function estimation.** After size factor estimation, the datasets are transformed using either rlog or vst transformation for the SVM, bagSVM, RF and CART algorithms. The logarithmic transformation approach transforms the data into a less skewed distribution with less extreme values as well; however, the genewise variances are still unstabilized [10]. In vst transformation, a local dispersion function is fit to the training data. This function is frozen to reapply for the test samples. In rlog transformation, the log fold changes of the counts for each gene are regularized over an intercept. Rlog transformation is applied as follows:

$$rlog(q_{ij}) = \beta_{0j} + \beta_{ij} \quad (6)$$

The dispersion function, beta prior variance and the intercept which are calculated from the training data are stored and directly used for the test dataset. More details can be found in DESeq2 paper [10].

**Filtering.** Next, we applied near-zero variance filtering to training data to filter the genes with low counts. The effect of the filtered genes is eliminated for further analysis [29]. Genes are filtered based on two criteria: (i) the frequency ratio of the most frequent value to the second most frequent value is higher than 19 (95/5), and (ii) the ratio of the number of unique values to the sample size is less than 10%. Filtered genes are removed from both the training and test sets. Next, the DESeq2 method is applied to detect the most DE 25, 50, 75 and 100 genes [10]. The same genes are selected for both test and training sets.

**Normalization and transformation.** After selecting the DE genes, training data are normalized with the estimated size factors to adjust sample specific differences [10]. After normalization, the datasets are transformed with the estimated dispersion functions using either rlog or vst transformation for the SVM, bagSVM, RF and CART algorithms. The normalized count datasets are directly used for the PLDA and NBLDA algorithms since both algorithms use discrete probability distributions to fit the models. Although the NBLDA takes overdispersion into account, the PLDA does not estimate the overdispersion parameter. Hence it assumes that there is no overdispersion in the data. Witten [6] suggested the use of power transformation on the raw counts when there is slight to moderate overdispersion in the data. Power transformation is useful to remove overdispersion in the data in such cases. However, one should explicitly estimate and consider overdispersion when data are highly overdispersed. In this paper, we performed power transformation for slightly or moderately overdispersed data. The results are given under PLDA<sub>2</sub>.

**Model building.** After the normalization and transformation processes, the parameters of each classifier are optimized to avoid overfitting and underfitting. A five-fold cross-validation is applied to the training data and the parameters that achieve the highest accuracy rate are selected as optimal parameters. Cross-validation folds are fixed for each classifier to make the results comparable. Each classifier is fitted with the optimal parameters. Fitted models are used in the test datasets for prediction and performance evaluation. The sample sizes are very low relative to the number of genes since we mimic the real datasets. Thus, the model performances may vary depending on the split ratio of the training and test sets. To overcome this limitation, we repeated the entire process 50 times and summarized the results in a single statistic, i.e. accuracy rates.

## Application to real datasets

In addition to the simulated data, four real datasets, including both miRNA and mRNA datasets were also used as real life examples (Table 1).

**Cervical dataset.** The cervical dataset is an miRNA sequencing dataset obtained from [30]. miRNAs are small non-coding RNA molecules with an average length of 21–23 bp. These

**Table 1. Description of real RNA-Seq datasets used in this study.**

Dataset	Number of groups	Sample size	Number of features
Cervical cancer [39]	2	58 (29 cervical cancer, 29 control)	714 miRNAs
Alzheimer [40]	2	70 (48 Alzheimer, 22 control)	416 miRNAs
Renal cell cancer [41]	3	1,020 (606 KIRP, 323 KIRC, 91 KICH)	20,531 mRNAs
Lung cancer [41]	2	1,128 (576 LUAD, 552 LUSC)	20,531 mRNAs

<https://doi.org/10.1371/journal.pone.0182507.t001>

small molecules regulate the gene expression levels. The objective of this study was to identify the novel miRNAs and detect the differentially expressed ones between normal and tumor cervical tissue samples. For this purpose, 58 small RNA libraries are constructed (29 with tumor and 29 without tumor). Among the 29 tumor samples, 21 were diagnosed as squamous cell carcinoma, 6 of them were adenocarcinomas and 2 were unclassified. In our analysis, we used the gene-expression levels of 714 miRNAs belonging to 58 human cervical tissue samples.

**Alzheimer dataset.** This dataset is another miRNA dataset provided by Leidinger et al. [31]. The authors aimed to discover potential miRNAs from blood in diagnosing Alzheimer and related neurological diseases. For this purpose, the authors obtained gene-expression data from 48 Alzheimer patients who were evaluated after undergoing some tests and 22 age-matched control samples. RNA sequencing is performed using an Illumina HiSeq2000 platform. The miRNAs with less than 50 counts in each group are filtered. We used the data including 416 miRNA read counts of 70 samples, where 48 Alzheimer and 22 control samples are considered as two separate classes for classification.

**Renal cell cancer dataset.** The renal cell cancer (RCC) dataset is an RNA-Seq provided by The Cancer Genome Atlas (TCGA) [32]. The TCGA is a comprehensive community resource platform for researchers to explore, download, and analyze datasets. RCC data contain 20,531 known human RNA transcript counts belonging to 1,020 RCC samples. These RNA-Seq data include 606 kidney renal papillary cell (KIRP), 323 kidney renal clear cell (KIRC) and 91 kidney chromophobe carcinoma (KICH) samples. These three classes are known as the most common subtypes of RCC and treated as three separate classes in our analysis [33].

**Lung cancer dataset.** Lung cancer is another RNA-Seq dataset provided by the TCGA platform. This dataset contains the read counts of 20,531 transcripts of 1,128 samples. Samples are separated into two distinct subclasses. These subclasses are lung adenocarcinoma (LUAD) and lung squamous cell with carcinoma (LUSC) with 576 and 552 class sizes, respectively. These two classes are used as class labels in our analysis.

**Evaluation process.** Real datasets are analyzed using similar procedures to those in the simulation study. Model building is performed on the training set (70%) and the test set (30%) is used to evaluate model performance. Size factors and dispersion functions are estimated for training datasets. Similar to the simulation experiments, the size factors and dispersion functions of test datasets are directly estimated from the training data to make them in the same scale and homoscedastic to each other. Near-zero variance filtering is applied to the training set. Filtered genes are also removed from the test set. The renal cell and lung cancer datasets include 20,531 features which dramatically increase the computational cost. Hence, we initially selected 5,000 genes with the highest variances to eliminate the effect of non-informative mRNAs and decrease the computational cost. All of the miRNAs are used in the model building process for the cervical and Alzheimer datasets. Differential expression was performed on the training data using the DESeq2 method and genes are ranked from the most significant to the least significant with increasing number of genes in steps of 25 up to 250 genes. The differentially expressed genes selected in the training data are also selected in the test datasets. Differentially expressed genes in the training data are normalized using the median ratio approach and transformed

using either the vst or rlog approaches. Since the sample sizes of the cervical and Alzheimer miRNA datasets are relatively small, the entire process is applied 50 times. The other model building processes applied are similar to those in the simulation study.

### Implementation of classifiers

Both simulated and real data are modeled using support vector machines (SVM), bagging support vector machines (bagSVM), random forests (RF), classification and regression trees (CART), Poisson linear discriminant analysis without power transformation (PLDA<sub>1</sub>), Poisson linear discriminant analysis with power transformation (PLDA<sub>2</sub>) and negative binomial linear discriminant analysis (NBLDA). In this section, we will summarize the background and use of each method.

**SVM.** SVM is among popular classification methods based on the statistical learning theory [34]. It has attracted great attention because of its strong mathematical background, learning capability, good generalization ability and wide range of application area such as computational biology, text classification, image segmentation and cancer classification [34,35]. SVM is capable of linear/nonlinear classification and deals with high-dimensional data.

Let  $x_i$  denotes the training data points,  $w$  denotes the weight vector and  $b$  denotes the bias term. The decision function that correctly classifies the data points by their true class labels in a linearly separable space is represented as follows:

$$f_{w,b} = \text{sign}(w \cdot x_i + b) \quad i = 1, 2, \dots, n \tag{7}$$

In a binary classification, the SVM aims to find an optimal separating hyperplane in the feature space which maximizes the margin and minimizes the misclassification rate by choosing the optimum value of  $w$  and  $b$  in Eq (7). When the cases are not linearly separable, “slack variables”  $\{\xi_1, \dots, \xi_n\}$ , a penalty term which is proposed by Cortes and Vapnik [36] can be used to allow misclassified data points where  $\xi_i > 0$ . In most of the classification problems, the separation surface is not linear. In this case, the SVM uses an implicit mapping  $\Phi$  of the input vectors to a high-dimensional space defined by a kernel function ( $K(x,y) = \Phi(x_i)\Phi(x_j)$ ) and the linear classification then applied in this high-dimensional space. Some of the most widely used kernel functions are linear:  $K(x,y) = x_i x_j$ , polynomial:  $K(x,y) = (x_i x_j + 1)^d$ , radial basis function:  $K(x,y) = \exp(-\gamma \|x_i - x_j\|^2)$  and sigmoidal:  $K(x,y) = \tanh(k(x_i x_j) - c)$  where  $c$  is a constant,  $d$  is the degree and  $\gamma > 0$  is sometimes parametrized as  $\gamma = 1/2\sigma^2$ . Normalized and transformed (either using vst or rlog) datasets are used as input to the SVM classifier. The radial basis kernel function is used in the analysis.

**bagSVM.** bagSVM is a bootstrap ensemble extension of SVM which creates individuals for its ensemble by training each SVM classifier on a random subset of the training set. For a given data set,  $k$  random bootstrap samples are drawn with replacement. SVM classifiers are trained independently on each randomly selected subsets and aggregated via an aggregation technique. A test set is predicted on each of the SVM classifiers and the predicted class labels are determined using aggregated results likely in training sets. Normalized and transformed datasets are used as input to the bagSVM classifier. The number of bootstrap samples were set to 101 since small changes were observed over this number.

**CART.** CART, which was introduced by Breiman [37], is one of the most popular tree classifiers with a wide range of applications. It uses the Gini index, which maximizes the decrease in impurity at each node, to find the optimal path. If  $p(i|j)$  is the probability of class  $i$  at node  $j$ , the Gini index is calculated using the equation  $1 - \sum_i p^2(i|j)$ . It is possible to obtain very large CART trees in large data sets, i.e. very large number of genes and samples. When



CART grows a maximal tree, this tree is pruned upward to get a decreasing sequence of sub-trees. Furthermore, pruning is preferred to overcome overfitting problem. The optimal tree that has the lowest misclassification rate is selected using a cross-validation. The assignment of each terminal node to a class is performed by choosing the class that minimizes the resubstitution estimate of the misclassification probability [37, 38]. Normalized and transformed datasets are used as input to the CART classifier.

**RF.** An RF is a collection of many CART trees combined by averaging the predictions of individual trees in the forest [39]. RF aims to combine many weak classifiers to produce a significantly better and strong classifier. First, training set is generated by drawing a bootstrap sample from the original data. This bootstrap sample includes 2/3 of the original data. The remaining part is used as a test set to predict the out-of-bag error of the classification. A subset of features are randomly selected at each node and the best split is used to split the corresponding nodes. If there are  $m$  features, for example,  $m_{try}$  out of  $m$  features is randomly selected at each node while growing the forest. Different splitting criterias can be used such as the Gini index, information gain and node impurity. The value of  $m_{try}$  is approximately equal to either  $\frac{\sqrt{m}}{2}$ ,  $\sqrt{m}$  or  $2\sqrt{m}$  and is constant during the forest growing. Unlike CART, an unpruned tree is grown for each of the bootstrap samples. Finally, class labels of new cases are predicted by aggregating (i.e. majority voting) the predictions from all trees [40, 41]. Normalized and transformed datasets are used as input to the RF classifier. The number of trees was set to 500 in the analysis.

**PLDA<sub>1</sub> and PLDA<sub>2</sub>.** Let  $X$  be an  $n \times p$  matrix of the sequencing data where  $n$  is the number of observations and  $p$  is the number of features. For sequencing data,  $X_{ij}$  indicates the total number of reads mapping to gene  $j$  in observation  $i$ . The observed counts are fitted to the Poisson log-linear model as given in Eq (8),

$$X_{ij} \sim \text{Poisson}(N_{ij}), \quad N_{ij} = s_i g_j \tag{8}$$

where  $s_i$  is the total number of reads per sample and  $g_j$  is the total number of reads per region of interest. For RNA-seq data, Eq (8) can be extended as follows:

$$X_{ij} | y_i = k \sim \text{Poisson}(N_{ij} d_{jk}), \quad N_{ij} = s_i g_j \tag{9}$$

where  $y_i \in \{1, \dots, K\}$  is the class label of the  $i^{th}$  observation. The  $d_{1j}, \dots, d_{Kj}$  terms allow the  $j^{th}$  feature to be differentially expressed between classes.

Let  $(x_i, y_i), i = 1, \dots, n$  be a training set and  $x^* = (X_1^*, \dots, X_p^*)^T$  be a test set. A new sample  $x^*$  is assigned to one of the classes with highest probability (or discrimination score) using the Bayes' rule as follows:

$$P(y^* = k | x^*) \propto f_k(x^*) \pi_k \tag{10}$$

where  $y^*$  denotes the unknown class label,  $f_k$  is the probability density of an observation in class  $k$  and  $\pi_k$  is the prior probability that an observation belongs to class  $k$ . If  $f_k$  is a Gaussian density with a class-specific mean and common variance, then a standard LDA is used to assign a new observation to the class [42]. When the model uses class-specific mean and a common diagonal matrix, then diagonal LDA is used for the classification [43]. However, the normality and common covariance matrix assumptions are not appropriate for sequencing data. Witten [6] assumes that the data follow a Poisson model as given in Eq (11),

$$X_{ij} | y_i = k \sim \text{Poisson}(N_{ij} d_{kj}), \quad N_{ij} = s_i g_j \tag{11}$$

where  $y_i$  is the class of the  $i^{th}$  observation and the features are independent. Eq (9) specifies

that  $X_j^* | y^* = k \sim \text{Poisson}(s^* g_j d_{kj})$ . First,

$$\log P(y^* = k | x^*) = \log \hat{f}_k(x^*) + \log \hat{\pi}_k + c = \sum_{j=1}^p X_j^* \log \hat{d}_{kj} - s^* \sum_{j=1}^p \hat{g}_j \log \hat{d}_{kj} + \log \hat{\pi}_k + c \quad (12)$$

where  $c$  and  $c'$  are constants and do not depend on the class label. A new observation is assigned to one of the classes for which Eq (12) is the largest [6].

Normalized count data are used as input to the PLDA<sub>1</sub> classifier. After normalization, a power transformation ( $X'_{ij} = \sqrt{X_{ij} + 3/8}$ ) is applied to reduce the overdispersion effect and make genes have constant variance. These normalized and power transformed datasets are used as input to the PLDA<sub>2</sub> classifier. To optimize the tuning parameter, a grid search (30 searches) is applied and the sparsest model with the highest accuracy rates is selected for classification.

**NBLDA.** Dong et al. [13] generalized the PLDA using an extra dispersion parameter ( $\phi$ ) of negative binomial distribution and called the method negative binomial linear discriminant analysis (NBLDA). This extra dispersion parameter is estimated using a shrinkage approach detailed in [44]. A new test observation will be assigned to its class based on the following NBLDA discriminating function:

$$\begin{aligned} \log P(y^* = k | x^*) = & \sum_{j=1}^p X_j^* [\log \hat{d}_{kj} - \log(1 + s^* \hat{g}_j d_{kj} \phi_j)] \\ & - \sum_{j=1}^p \phi_j^{-1} \log(1 + s^* \hat{g}_j d_{kj} \phi_j) + \log \hat{\pi}_k + c' \end{aligned} \quad (13)$$

As the dispersion decreases, NBLDA approximates to PLDA. More details on NBLDA can be found in [13].

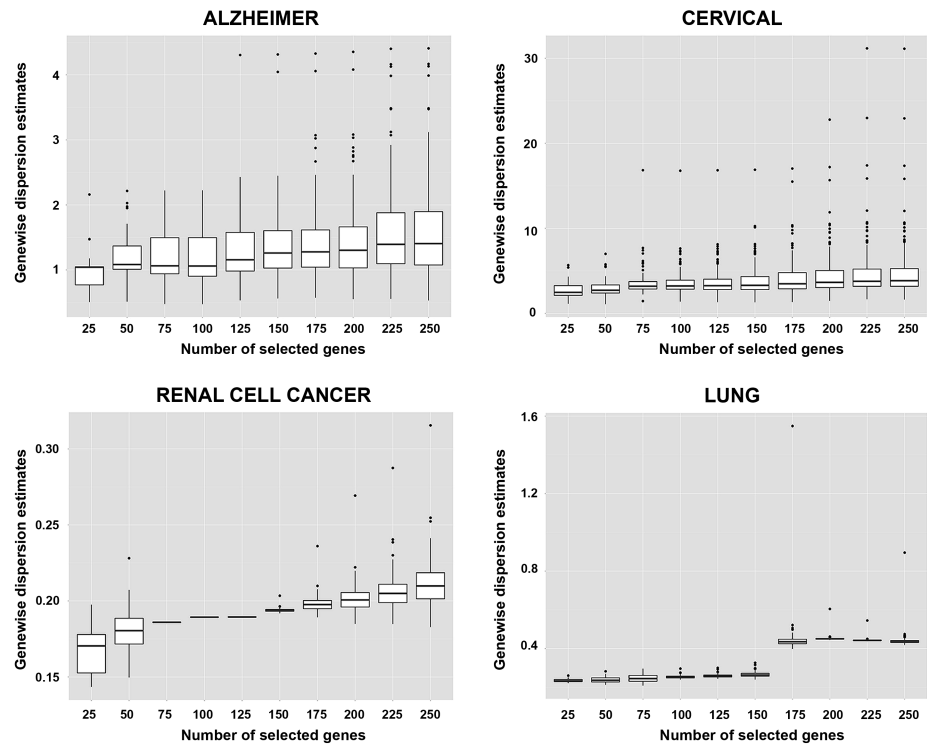
**Evaluation criteria.** To validate each classifier model, five-fold cross-validation is used. It is repeated 10 times and accuracy rates are calculated to evaluate the performance of each model. Cross-validation folds are fixed for all classifiers to make the results comparable to each other. Accuracy rates are calculated as  $(TP + TN)/n$  based on the confusion matrices of test set class labels and test set predictions. For multiclass scenarios, these measures are calculated via the one-versus-all approach. Since, the class sizes are unbalanced in the Alzheimer and renal cell cancer datasets, accuracies are balanced using the following formula:  $(\text{Sensitivity} + \text{Specificity})/2$ .

## MLSeq R/BIOCONDUCTOR package

We presented an R package in the BIOCONDUCTOR network for classification of the RNA-seq data. The MLSeq package accepts gene expression data which can be obtained from feature counting tools (e.g. HTSeq [14], bedtools [20] and FeatureCounts [21] etc.). It also has the ability to normalize and transform the gene-expression data. Finally, data are fitted to the selected model such as SVM, bagSVM, RF and CART. Users can access the MLSeq package from <https://www.bioconductor.org/packages/release/bioc/html/MLSeq.html>.

## Results and discussion

Genewise dispersion parameters are estimated for each classifier using the method of moments. Distribution of the estimated overdispersions is given in Fig 2. It is seen from the figure that the cervical and Alzheimer miRNA datasets are very highly overdispersed ( $\phi > 1$ ), while the lung and renal cell cancer datasets are substantially overdispersed. The simulation results for  $k = 2$  and  $k = 3$ ,  $d_{kj} = 10\%$  for rlog transformations are given in Fig 3 and Fig 4. All other



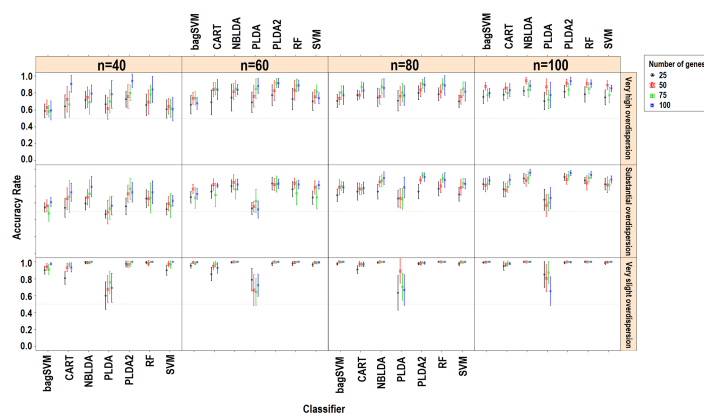
**Fig 2. Genewise dispersion estimations for real datasets.**

<https://doi.org/10.1371/journal.pone.0182507.g002>

simulation results are given in <http://www.biosoft.hacettepe.edu.tr/MLSeqSupplementary/> and the [S1 File](#). Results for real datasets are given in [Fig 5](#).

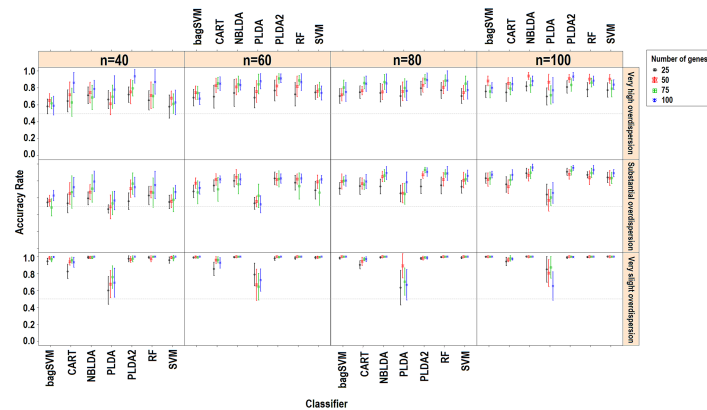
### Effect of simulation parameters

Since combining each significant gene on class conditions is equivalent to combining their predictive abilities, an increased number of differentially expressed genes leads to an increase in the classification accuracy ([Fig 2](#) and [Fig 3](#)). Similarly, in most scenarios, working with



**Fig 3. Simulation results for  $k = 2$ ,  $d_{kj} = 10\%$ , transformation:  $rlog$ .** Figure shows the performance results of classifiers with changing parameters of sample size ( $n$ ), number of genes ( $p$ ) and type of dispersion ( $\varphi = 0.01$ : very slight,  $\varphi = 0.1$ : substantial,  $\varphi = 1$ : very high).

<https://doi.org/10.1371/journal.pone.0182507.g003>



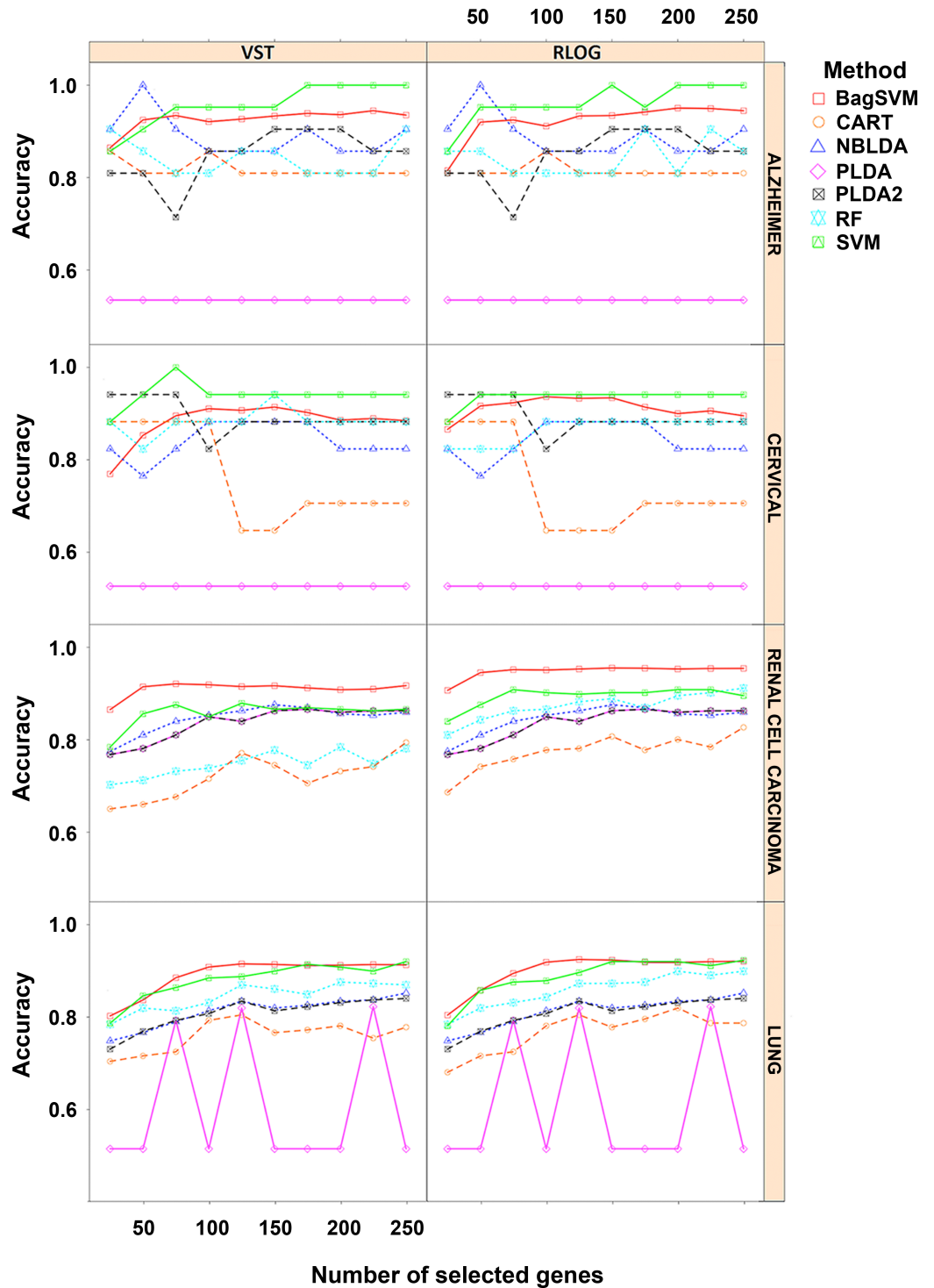
**Fig 4. Simulation results for  $k = 3$ ,  $d_{kj} = 10\%$ , transformation: rlog.** Figure shows the performance results of classifiers with changing parameters of sample size ( $n$ ), number of genes ( $p$ ) and type of dispersion ( $\varphi = 0.01$ : very slight,  $\varphi = 0.1$ : substantial,  $\varphi = 1$ : very high).

<https://doi.org/10.1371/journal.pone.0182507.g004>

more samples and genes has a positive impact on the overall model accuracies. This relationship between the number of genes and accuracy is mostly available when  $d_{kj} = 10\%$ . Likewise, classification accuracies slightly increase in real datasets since increasing the number of genes leads to an increase in the probability of a differentially expressed gene being included in the classification model. However, this may not be true for all cases. Since, the classification accuracy may decrease from certain point as more genes were included in the model. Since, increasing the number of genes will lead to an increase in the model complexity, we expect that the classification error of test samples will be reduced. For the PLDA classifier, a high number of selected genes provides alternative options for the lasso shrinkage method to test more genes in classification models. The RF algorithm builds trees with the bagging approach. As the number of genes increases, the RF algorithm uses more genes and may perform better to specify the optimal tree. Increasing the sample size, on the other hand, improves the discrimination power and classification accuracy. Conversely, overall accuracies decrease as the number of classes increases. This is due to the fact that the probability of misclassifying an observation may arise depending on the complexity of the model. As the number of classes increases, the samples become less likely to be linearly (or non-linearly) separable.

### Effect of dispersion on classification accuracies

The overall performance of each model increases as the data become less dispersed. Decreasing the dispersion parameter makes a significant contribution to classification accuracy, as expected, even for the same data and the same scenario. This is more obvious when  $k = 2$  and  $d_{kj} = 10\%$ . As the data become more spread out (i.e, increasing overdispersion), the variance increases; thus, we need more samples to better learn from data and achieve the same discrimination power. When we stabilize the sample size and increase the dispersion parameter, this leads to a decrease in the discrimination power and classification accuracies. Nagalakshmi et al. [7] reported that using biological replicates instead of technical replicates leads to an increase in the dispersion of the data. Based on this idea, increasing the biological variance of the observations results in an increase in the dispersion, which decreases the discrimination power. In DE studies of RNA-Seq data, overdispersion is one of the major problems to be carefully handled. Many studies are available in the literature which focus on the overdispersion problem [9–10, 25, 45–46]. When we look at the classification accuracy results, overdispersion also seems to be a major challenge in classification studies. Unless we work with technical replicates, RNA-Seq data are overdispersed and the mapped



**Fig 5. Results obtained from real datasets.** Figure shows the performance results of classifiers for datasets with changing number of most significant number of genes. Note that PLDA and NBLDA methods are not performed on the transformed data. However, the results for both transformed and non-transformed data are given in the same figure for the comparison purpose.

<https://doi.org/10.1371/journal.pone.0182507.g005>

counts from different biological replicates on the same gene have variance exceeding the mean [7]. This overdispersion can be seen in other studies [9, 46–49]. In conclusion, the results of our study revealed that overdispersion has a significant and negative effect on classification accuracies and should be taken into account before model building.

### Microarray-based classifiers and transformation effect on classification accuracies

Hundreds of microarray-based classifiers have been developed and are able to work in large  $p$  and small  $n$  settings. However, technological improvements make RNA-Seq a state-of-the-art approach for quantified transcriptomics. Currently, much of these microarray-based classifiers cannot be directly applied to RNA-Seq data because of the discrete nature of RNA-Seq data. Microarray data consist of the continuous log-intensities of probes while RNA-Seq data consist of the discrete and overdispersed mapped read counts of sequencing technologies. The results of this study revealed that, transforming the RNA-Seq data and bringing them hierarchically closer to microarrays (e.g. through  $rlog$  and  $vst$ ) might be a suitable approach to make the microarray-based classifier applicable for RNA-Seq data.

Witten et al. [6] stated that the normalization strategy has little impact on classification performance but may be important in differential expression analysis. However, data transformation has a direct effect on classification results since it changes the underlying distribution of the data. In this study, we used  $deseq$  normalization with  $vst$  and  $rlog$  transformations and obtained satisfactory classification performances. Love et al. [10] reported that  $vst$  transformation does not consider the size factors during transformation. However, in both simulated and real datasets, there were no substantial differences between  $rlog$  and  $vst$  transformation approaches in terms of classification accuracies. Both transformations can be applied to RNA-Seq data.

### Power transformed PLDA and other count-based classifiers

Without transformation, the PLDA seemed to perform well in very slightly overdispersed datasets. This can be seen in both simulated and real datasets. For instance, in the renal cell carcinoma dataset, the dispersion parameter is very low and the data seem to follow Poisson distribution. In this case, the overdispersion is negligible and no power transformation is needed. Hence,  $PLDA_1$  and  $PLDA_2$  show similar performances (Fig 5). However, the performance of this method decreases when the data become more overdispersed. The reason is that the PLDA classifies the data assuming that the underlying distribution is Poisson even though it should be negative binomial. Although the Poisson distribution assumption might be valid after power transformation for moderately overdispersed data, it is invalid for highly overdispersed data. Therefore, based on the results for both simulated and real datasets, we suggest that this transformation is very useful and should be applied for moderately overdispersed data before building the model. When power transformation fails, one should use negative binomial distribution rather than Poisson distribution. The NBLDA extends this classifier using a negative binomial model. We expect that the NBLDA should give better predictive performances compared to the PLDA. However, the classification accuracy of this method is not as high that of the PLDA with power transformation. This might be for several reasons. First, the NBLDA algorithm is not sparse. Hence, it uses more features than the PLDA; as a result of overdispersion, even these genes are not differentially expressed among classes. Second, there are several methods for estimating the overdispersion parameter. The selected method for overdispersion estimation might be another reason for lower predictive accuracy. In conclusion, novel or improved count-based classifiers are still needed for accurate and robust classification of RNA-Seq data.

In lung cancer dataset, the odd behavior of PLDA may appear based on its own built-in variable selection algorithm. In this data set, when PLDA is able to select some of the features in the classification task, it gives higher classification accuracy. However, if this algorithm was not able to select any feature in the discrimination, it classifies all the observations into first class. Hence the accuracy is the ratio of the samples in the corresponding class. However, when PLDA is used with power transformation, the accuracy of the model significantly increases and the variable selection algorithm works well in most of the model fitting process.

## Overall performances of classifiers

In simulated datasets, the power transformed PLDA was found to be the best classifier. The RF and NBLDA performed in a moderately similar manner. On the other hand, the SVM and bagSVM had the highest classification accuracies in real datasets. The PLDA<sub>2</sub>, RF and NBLDA give comparable and high classification accuracies, but they are lower than SVM and bagSVM. These slight differences may arise from the differences between negative binomial distribution, which is used in the simulation settings, and the exact distributions of real RNA-Seq data. In real datasets, SVM and bagSVM classifiers display their classification abilities. Moreover, it can be seen from the simulated and real datasets that the performance of the bagSVM classifier increases as the sample size increases. A possible explanation for such an observation is that the bagSVM uses the bootstrap technique and trains better models in datasets with a high number of samples. We also observed that PLDA<sub>2</sub> and NBLDA performed more accurate in cervical and alzheimer datasets. Similar with the simulation results, we can say that these two algorithms are more efficient in highly overdispersed datasets. The performances of CART and PLDA<sub>1</sub> were found to be lower than those of the other classifiers. This result is consistent with the results of simulated data.

All assessments in this study are made based on the classification accuracies. Another important measure may be the sparsity of classifiers. Since we included mostly the non-sparse classifiers in this study, we will leave the effect of the dispersion parameter on sparsity as a topic for further research.

## Conclusions

A considerable amount of evidence collected from genome-wide gene-expression studies suggests that the identification and comparison of differentially expressed genes have been a promising approach for diagnosis and prognosis purposes. Although microarray-based gene-expression studies have been widely used for discovering potential biomarkers related to disease status [50–53], it has limitations in terms of novel transcript discovery and abundance estimation with a large dynamic range. Thus, one choice is to utilize the power of RNA-Seq techniques in the analysis of the transcriptome for diagnostic classification to overcome the limitations of microarray-based experiment. As mentioned in earlier sections, working with less noisy data may improve the predictive performance of classifiers and novel transcripts may be discovered as a new biomarker in the studied disease or phenotypes.

Hundreds of studies have been published on microarray-based classification. The goal of these studies was to develop or adapt novel approaches to identify a small subset of genes and predict the class labels of a new observation. This has particular importance in biomedical studies for the molecular diagnosis of diseases. In this study, we demonstrated how researchers can classify the RNA-Seq data which is the state-of-the-art technique for the quantification of gene expression. We conducted a comprehensive simulation study and also used four real experimental miRNA/mRNA datasets.

Besides RNA-Seq's advantages over microarrays, the gene-expression data from RNA-Seq are overdispersed due to inherent variability. This overdispersion seemed to be a drawback for differential expression studies of RNA-Seq data. In this study, we showed that this overdispersion is also a drawback for classification studies since an increase in the variance will lead to a decrease in the discrimination power. We reached the conclusion that three solutions are available to handle the classification of overdispersed RNA-Seq data: (i) increasing the sample size, (ii) transforming the data to bring RNA-Seq data hierarchically closer to microarrays using variance stabilizers, e.g. vst and rlog transformations and (iii) using count-based classifiers such as the PLDA<sub>2</sub> and NBLDA. Our simulation study revealed that both microarray-based classifiers after rlog/vst transformations and count-based classifiers (that deal with overdispersion) can be efficiently used for the classification of RNA-Seq data.

To make an overall assessment for the performances of classifiers, the PLDA after a power transformation may be a good choice as a count-based classifier. Furthermore, its sparsity seems to be an advantage for researchers. However, further studies are needed. Surprisingly, the performance of the NBLDA was not satisfactory as a count-based classifier. Dong et al. [13] reported that the NBLDA performs better than the PLDA in moderate and highly overdispersed data. However, these comparisons were made with the same number of genes. Our analyses were performed based on the sparse PLDA classifiers where the best subset of genes is used in classification. The sparse PLDA classifier after a power transformation performed more accurately in all dispersion settings. We believe that extending the NBLDA algorithm into a sparse classifier may improve its classification performance by selecting the most significant genomic features.

Moreover, an alternative option may be to bring the data closer to microarrays and use microarray-based classifiers. Our results revealed that the RF, SVM and bagSVM may give accurate results after an rlog or vst transformation. Moreover, the efficiency of the bagSVM is improved markedly with increasing sample size.

We conclude that data with less overdispersion, highly differentially expressed genes, a lower number of groups and large sample size may improve the accuracy of the classifiers. Finally, we developed an R/BIOCONDUCTOR package, called MLSeq, for the classification of RNA-Seq data. This package can be accessed and downloaded through <https://www.bioconductor.org/packages/release/bioc/html/MLSeq.html>.

## Supporting information

**S1 File. All figures for simulation results.**

(ZIP)

**S2 File. MLSeq package source.**

(ZIP)

**S3 File. Simulation R codes.**

(ZIP)

**S4 File. Computational infrastructure.**

(DOCX)

**S5 File. Computational costs of classifiers.**

(DOCX)



## Acknowledgments

We would like to thank A. Keller for sharing the Alzheimer data, and also B. Klaus, S. Anders and M.I. Love for insightful discussions on the simulation settings of this paper.

## Author Contributions

**Conceptualization:** Gökmen Zararsız, Ahmet Ozturk.

**Data curation:** Gökmen Zararsız.

**Formal analysis:** Dincer Goksuluk, Selcuk Korkmaz, Vahap Eldem, Gozde Erturk Zararsız.

**Methodology:** Gökmen Zararsız, Dincer Goksuluk, Selcuk Korkmaz.

**Project administration:** Gökmen Zararsız, Ahmet Ozturk.

**Resources:** Selcuk Korkmaz, Vahap Eldem, Gozde Erturk Zararsız, Izzet Parug Duru.

**Software:** Gökmen Zararsız, Dincer Goksuluk, Selcuk Korkmaz.

**Supervision:** Ahmet Ozturk.

**Validation:** Gökmen Zararsız, Dincer Goksuluk, Izzet Parug Duru, Ahmet Ozturk.

**Visualization:** Dincer Goksuluk.

**Writing – original draft:** Gökmen Zararsız, Dincer Goksuluk, Selcuk Korkmaz, Vahap Eldem.

**Writing – review & editing:** Gökmen Zararsız, Dincer Goksuluk, Selcuk Korkmaz, Vahap Eldem.

## References

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009, 10(1): 57–63. <https://doi.org/10.1038/nrg2484> PMID: 19015660
2. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics.* 2000, 16(10):906–914. PMID: 11120680
3. Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert JP. Classification of microarray data using gene networks. *BMC Bioinformatics.* 2007, 8(1):35.
4. Uriarte RD, de Andres SA. Gene selection and classification of microarray data using random forest. *BMC Bioinformatic.* 2006, 7(1):3.
5. Zhu J, Hastie T. Classification of gene microarrays by penalized logistic regression. *Biostatistics.* 2004, 5(3):427–443. <https://doi.org/10.1093/biostatistics/5.3.427> PMID: 15208204
6. Witten DM. Classification and clustering of sequencing data using a poisson model. *Ann Appl Stat.* 2011, 5(4): 2493–2518.
7. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science.* 2008, 320(5881):1344–9. <https://doi.org/10.1126/science.1158441> PMID: 18451266
8. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biology.* 2010, 11(10):R106. <https://doi.org/10.1186/gb-2010-11-10-r106> PMID: 20979621
9. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010, 26(1):139–140. <https://doi.org/10.1093/bioinformatics/btp616> PMID: 19910308
10. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Biorxiv* <https://doi.org/https://doi.org/10.1101/002832>
11. Sonesson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics.* 2013, 14(91). <https://doi.org/10.1186/1471-2105-14-91> PMID: 23497356

12. Di Y, Schafer DW, Cumbie JS, Chang JH. The NBP negative binomial model for assessing differential gene expression from RNA-seq. *Stat Appl Genet Mol Biol*. 2011, 10:Article 24.
13. Dong K, Zhao H, Wan X, Tong T. NBLDA: Negative Binomial Linear Discriminant Analysis for RNA-Seq Data. *BMC Bioinformatics*. 2016, 17(369). <https://doi.org/10.1186/s12859-016-1208-1> PMID: 27623864
14. Anders S, Pyl PT, Huber W. HTSeq – A Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015, 31(2):166–169. <https://doi.org/10.1093/bioinformatics/btu638> PMID: 25260700
15. Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, Gentleman R. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*. 2009, 25:2607–2608. <https://doi.org/10.1093/bioinformatics/btp450> PMID: 19654119
16. Shrestha RK, Lubinsky B, Bansode VB, Moinz M, McCormack GP, Travers SA. QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform. *BMC Bioinformatics*. 2014, 15:15–33.
17. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*. 2013, 14:R36. <https://doi.org/10.1186/gb-2013-14-4-r36> PMID: 23618408
18. Wang K, Sing D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*. 2010, 38(18):e178. <https://doi.org/10.1093/nar/gkq622> PMID: 20802226
19. Dobin A, Davis JA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013, 29(1): 15–21. <https://doi.org/10.1093/bioinformatics/bts635> PMID: 23104886
20. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010, 26(6):841–842. <https://doi.org/10.1093/bioinformatics/btq033> PMID: 20110278
21. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014, 30(7):923–930. <https://doi.org/10.1093/bioinformatics/btt656> PMID: 24227677
22. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*. 2010, 11:R25. <https://doi.org/10.1186/gb-2010-11-3-r25> PMID: 20196867
23. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods*. 2008, 5(7):621–628. <https://doi.org/10.1038/nmeth.1226> PMID: 18516045
24. Bullard JH. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments, *BMC Bioinformatics*. 2010, 11:94. <https://doi.org/10.1186/1471-2105-11-94> PMID: 20167110
25. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*. 2014, 15:R29. <https://doi.org/10.1186/gb-2014-15-2-r29> PMID: 24485249
26. Ding C, Peng H. Minimum redundancy feature selection from microarray. *Proceedings of the Computational Systems Bioinformatics (CSB'03) 2003*
27. Xing EP, Jordan MI, Karp RM. Feature selection for high-dimensional genomic microarray data. *Proceedings of the Eighteenth International Conference on Machine Learning 2001*, 601–608
28. Witten D. PoiClaClu: Classification and clustering of sequencing data based on a Poisson model. R package version 1.0.2. 2013, <https://CRAN.R-project.org/package=PoiClaClu>
29. Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Softw*. 2008, 28(5).
30. Witten D, Tibshirani R, Gu SG, Fire A, Lui WO. Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biology*. 2010, 8(58). <https://doi.org/10.1186/1741-7007-8-58> PMID: 20459774
31. Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, et al. A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biology*. 2013, 14:R78. <https://doi.org/10.1186/gb-2013-14-7-r78> PMID: 23895045
32. Saleem M, Shanmukha A, Ngonga Ngomo AC, Almeida JS, Decker HF, Deus HF. Linked cancer genome atlas database. I-SEMANTICS '13 - Proceedings of the 9th International Conference on Semantic Systems: 04-06 September 2013 – Graz 2013 (p. 129-134).
33. Goyal R, Gersbach E, Yang XJ, Rohan SM. Differential Diagnosis of Renal Tumors with Clear Cytoplasm. *Clinical Relevance of Renal Tumor Subclassification in the Era of Targeted Therapies and Personalized Medicine. Arch Patol Lab Med*. 2013, 137:467–80.

34. Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer; 2000.
35. Korkmaz S, Zararsiz G, Goksuluk D. MLViS: A Web Tool for Machine Learning-Based Virtual Screening in Early-Phase of Drug Discovery and Development. *PLoS One*. 2015, 10(4): e0124600. <https://doi.org/10.1371/journal.pone.0124600> PMID: 25928885
36. Cortes C, Vapnik V. Support vector network, *Mach Learn*. 1995, 20:73–97.
37. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software; 1984.
38. Dudoit S, Fridlyand J. Classification in microarray experiments. *Statistical analysis of gene expression microarray data*. 2003, 1: 93–158.
39. Breiman L. Random Forests. *Mach Learn*. 2001, 45:5–32.
40. Liaw A, Wiener M. Classification and Regression by randomForest. *R news*. 2002, 2(3), 18–22.
41. Okun O, Priisalu H. Random forest for gene expression based cancer classification: overlooked issues. In *Pattern Recognition and Image Analysis*. 2007, (pp. 483–490). Springer Berlin Heidelberg.
42. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning*. New York: Springer; 2009.
43. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Amer Statist Assoc*. 2001, 96:1151–1160.
44. Yu D, Huber W, Vitek O. Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. *Bioinformatics*. 2013, 29(10): 1275–82. <https://doi.org/10.1093/bioinformatics/btt143> PMID: 23589650
45. Anders S, Huber W. Differential expression of RNA-Seq data at the gene level – the DESeq package. (2012)
46. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*. 2007, 23(21):2881–2887. <https://doi.org/10.1093/bioinformatics/btm453> PMID: 17881408
47. Bloom JS, Khan Z, Kruglyak L, Singh M, Caudy AA. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*. 2009, 10:221. <https://doi.org/10.1186/1471-2164-10-221> PMID: 19435513
48. Zhou YH, Xia K, Wright FA. A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics*. 2011, 27(19):2672–2678. <https://doi.org/10.1093/bioinformatics/btr449> PMID: 21810900
49. Auer PL, Doerge RW. A two-stage poisson model for testing RNA-Seq data. *Stat Appl Genet Mol*. 2011, 10(1):26.
50. Lee ZJ. An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer. *Artif Intell Med*. 2008, 42(1): 81–93. <https://doi.org/10.1016/j.artmed.2007.09.004> PMID: 18006289
51. Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*. 2008, 9(1):319.
52. Anand A, Suganthan PN. Multiclass cancer classification by support vector machines with class-wise optimized genes and probability estimates. *J Theor Biol*. 2009, 259(3): 533–540. <https://doi.org/10.1016/j.jtbi.2009.04.013> PMID: 19406131
53. George G, Raj VC. Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile. *arXiv preprint 2011*, arXiv:1109.1062.