A comprehensive simulation study on classification of RNA-Seq data

Gokmen Zararsiz [Corresp., 1], **Dinçer Göksülük** [2], **Selçuk Korkmaz** [2], **Vahap Eldem** [3], **Gözde Ertürk Zararsız** [1], **İzzet Parug Duru** [4], **Turgay Unver** [5], **Ahmet Öztürk** [1]

[1] Biostatistics, Erciyes University, Faculty of Medicine, Kayseri, TURKEY
[2] Department of Biostatistics, Hacettepe University, Ankara, Turkey
[3] Department of Biology, Istanbul University, Istanbul, Turkey
[4] Department of Physics, Marmara University Istanbul, Istanbul, Turkey
[5] Genomics, Izmir International Biomedicine and Genome Institute, Izmir, Turkey

Corresponding Author: Gokmen Zararsiz
Email address: gokmenzararsiz@hotmail.com

Background RNA sequencing (RNA-Seq) is a powerful technique for transcriptome profiling of the organisms that uses the capabilities of next-generation sequencing (NGS) technologies. Recent advances in NGS let to measure the expression levels of tens to thousands of transcripts simultaneously. Using such information, developing 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 disease. Microarray based classifiers cannot be directly applied due to the discrete nature of RNA-Seq data. One way is to develop count-based classifiers, such as poisson linear discriminant analysis (PLDA) and negative binomial linear discriminant analysis (NBLDA). Other way is to transform the data hierarchically closer to microarrays and apply microarray-based classifiers. In most of the studies, the data overdispersion seems to be an another challenge in modeling RNA-Seq data. In this study, we aimed to examine the effect of dispersion parameter and classification algorithms on RNA-Seq classification. We also considered the effect of other parameters (i) sample size, (ii) number of genes, (iii) number of class, (iv) DE (differential expression) rate, (v) transformation method on classification performance.

Methods We designed a comprehensive simulation study, also used two miRNA and two mRNA experimental datasets. Simulated datasets are generated from negative binomial distribution under different scenarios and real datasets are obtained from publicly available resources. We compared the results of several classifiers including PLDA with and without power transformation, NBLDA, single SVM, bagging SVM (bagSVM), classification and regression trees (CART), and random forests (RF).

Results Results from the simulated and real datasets revealed that increasing the sample size, differential expression rate, number of genes and decreasing the dispersion parameter and number of groups lead to an increase in the classification accuracy. To make an overall assessment, power transformed PLDA, RF and SVM classifiers performed the highest classification accuracies.

Discussion Overdispersion seems to be an important challenge in RNA-Seq classification studies. Similar with differential expression studies, classification of RNA-Seq data requires careful attention on handling data overdispersion. We conclude that, as a count-based classifier, power transformed PLDA; as a microarray based classifier vst or rlog transformed RF and SVM (bagSVM for high sample sized data) classifiers may be a good choice for classification. However, there is still a need to develop novel classifiers or transformation approaches for classification of RNA-Seq data. An **R/BIOCONDUCTOR** package MLSeq with a vignette is freely available at

http://www.bioconductor.org/packages/2.14/bioc/html/MLSeq.html .

A Comprehensive Simulation Study on Classification of RNA-Seq Data

Gokmen Zararsiz[1], Dincer Goksuluk[2], Selcuk Korkmaz[2], Vahap Eldem[3], Gozde Erturk Zararsiz[1], Izzet Parug Duru[4], Turgay Unver[5], Ahmet Ozturk[1]

[1]Department of Biostatistics, Erciyes University, Kayseri, Turkey

[2]Department of Biostatistics, Hacettepe University, Ankara, Turkey

[3]Department of Biology, Istanbul University, Istanbul, Turkey

[4]Department of Physics, Marmara University, Istanbul, Turkey

[5]Izmir International Biomedicine and Genome Institute, İzmir, Turkey

Corresponding Author:

Gokmen Zararsiz

Erciyes University Biostatistics, Kayseri, 38039, Turkey

Email address: gokmenzararsiz@erciyes.edu.tr, gokmenzararsiz@hotmail.com

Abstract

Background

RNA sequencing (RNA-Seq) is a powerful technique for transcriptome profiling of the organisms that uses the capabilities of next-generation sequencing (NGS) technologies. Recent advances in NGS let to measure the expression levels of tens to thousands of transcripts simultaneously. Using such information, developing 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 disease. Microarray based classifiers cannot be directly applied due to the discrete nature of RNA-Seq data. One way is to develop count-based classifiers, such as poisson linear discriminant analysis (PLDA) and negative binomial linear discriminant analysis (NBLDA). Other way is to transform the data hierarchically closer to microarrays and apply microarray-based classifiers. In most of the studies, the data overdispersion seems to be an another challenge in modeling RNA-Seq data. In this study, we aimed to examine the effect of dispersion parameter and classification algorithms on RNA-Seq classification. We also considered the effect of other parameters (i) sample size, (ii) number of genes, (iii) number of class, (iv) DE (differential expression) rate, (v) transformation method on classification performance.

Methods

We designed a comprehensive simulation study, also used two miRNA and two mRNA experimental datasets. Simulated datasets are generated from negative binomial distribution under different scenarios and real datasets are obtained from publicly available resources. Data normalization is applied using deseq median ratio approach. A variance stabilizing transformation (vst) and regularized logarithmic transformation (rlog) methods are used before

applying microarray-based classifiers. We compared the results of several classifiers including PLDA with and without power transformation, NBLDA, single SVM, bagging SVM (bagSVM), classification and regression trees (CART), and random forests (RF).

Results

Results from the simulated and real datasets revealed that increasing the sample size, differential expression rate, number of genes and decreasing the dispersion parameter and number of groups lead to an increase in the classification accuracy. To make an overall assessment, power transformed PLDA, RF and SVM classifiers performed the highest classification accuracies.

Discussion

Overdispersion seems to be an important challenge in RNA-Seq classification studies. Similar with differential expression studies, classification of RNA-Seq data requires careful attention on handling data overdispersion. We conclude that, as a count-based classifier, power transformed PLDA; as a microarray based classifier vst or rlog transformed RF and SVM (bagSVM for high sample sized data) classifiers may be a good choice for classification. However, there is still a need to develop novel classifiers or transformation approaches for classification of RNA-Seq data. An **R/BIOCONDUCTOR** package MLSeq with a vignette is freely available at http://www.bioconductor.org/packages/2.14/bioc/html/MLSeq.html.

Introduction

With the advent of high-throughput NGS technologies, transcriptome sequencing (RNA-Seq) has become one of the central experimental approaches for generating a comprehensive catalog of protein-coding genes and non-coding RNAs and examining the transcriptional activity of genomes. Furthermore, RNA-Seq has already proved itself to be a promising tool with a remarkably diverse range of applications; (i) discovering novel transcripts, (ii) detection and quantification of spliced isoforms, (iii) fusion detection, (iv) reveal sequence variations (e.g., SNPs, indels) (Wang, Gerstein & Snyder, 2009).Additionally, beyond these general applications, RNA-Seq holds great promise for gene expression-based classification to identify the significant transcripts, distinguish biological samples and predict clinical or other outcomes due to large amounts of data, which can be generated in a single run. This classification is widely used in medicine for diagnostic purpose and refers to the detection of small subset of genes that achieves the maximal predictive performance. These genes are used afterwards for classification of new observations into the disease classes (or tumor classes, cancer subtypes, cancer stage, etc.).

Although microarray-based gene expression classification have become very popular during last decades, more recently, RNA-Seq replaced microarrays as the technology of choice in quantifying gene expression due to some advantages as providing less noisy data, detecting novel transcripts and isoforms, and unnecessary of prearranged transcripts of interest (Furey et al., 2000; Zhu & Hastie, 2004; Uriarte & de Andres, 2006; Rapaport et al., 2007).However, to measure gene expression, microarray technology provides continuous data, while RNA-Seq technology generates discrete count data, which corresponds to the abundance of mRNA transcripts (Witten, 2011).Another issue is the overdispersion problem, where the variance exceeds the mean (Nagalakshmi et al., 2008). Various studies have been employed to deal with

the overdispersion problem for differential expression (DE) analysis of RNA-Seq data (Anders & Huber, 2010; Robinson, McCarthy & Smyth, 2010; Di et al., 2011; Soneson & Delorenzi, 2013; Love, Huber & Anders, 2014).

One way is to use discrete probability distributions (e.g. poisson, negative binomial) to deal with huge amount of RNA-Seq data for expression-based classification purpose. Witten et al. (Witten, 2011)proposed the sparse Poisson linear discriminant analysis (PLDA) classifier by extending the popular microarray classifier, nearest shrunken centroids algorithm, to discrete RNA-Seq data. The authors also suggested applying a power transformation, since Poisson distribution underestimates the variation observed from the data. Dong et al. (Dong et al., 2016) proposed negative binomial distribution by extending PLDA with the use of negative binomial distribution. 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 hierarchically closer to microarrays and apply known algorithms for classification applications (Nagalakshmi et al., 2008; Anders & Huber, 2010; Robinson, McCarthy & Smyth, 2010).

In this study, we designed a comprehensive simulation study, also used four real datasets to examine the effect of dispersion parameter and classification algorithms on RNA-Seq classification. We also considered the effect of other parameters (i) sample size, (ii) number of genes, (iii) number of class, (iv) DE rate, (v) transformation method on classification performance. For each scenario, we performed PLDA and NBLDA as well as other machine learning algorithms i.e. support vector machines (SVM), bagging support vector machines (bagSVM), random forests (RF) and classification and regression trees (CART) algorithms.

Materials and Methods

A workflow for RNA-Seq classification

Providing a pipeline for classification algorithm of RNA-Seq data gives us a quick snapshot view of how to handle the large-scale transcriptome data and establish a robust inference by using computer-assisted learning algorithms. Therefore, we outlined the count-based classification pipeline for RNA-Seq data in Fig. 1. NGS platforms produce millions of raw sequence reads with quality scores corresponding to each base-call. The first step in RNA-Seq data analysis is to assess the quality of the raw sequencing data for meaningful downstream analysis. The conversion of raw sequence data into ready-to-use clean sequence reads needs a number of processes such as removing the poor-quality sequences, low-quality reads with more than five unknown bases, and trimming the sequencing adaptors and primers. In quality assessment and filtering, the current popular tools are FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), HTSeq (Anders, Pyl & Huber, 2015), R ShortRead package (Morgan et al., 2009), PRINSEQ (http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi), FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and QTrim (Shrestha et al., 2014). Following these procedures, next step is to align the high-quality reads to a reference genome or transcriptome. It has been reported that the number of reads mapped to the reference genome is linearly related to the transcript abundance. Thus, transcript quantification (calculated from the total number of mapped reads) is a prerequisite for further analysis. Splice-aware short read aligners such as Tophat2 (Kim et al., 2013), MapSplice (Wang et al., 2010) or Star (Dobin et al., 2012) can be prefered instead of unspliced aligners (BWA, Bowtie, etc.). After obtaining the mapped reads, next step is counting how many reads mapped to each transcript. In this way, gene expression levels can be inferred for each sample for downstream analysis. This

step can be accomplished with HTSeq (Anders, Pyl & Huber, 2015), bedtools (Quinlan & Hall, 2010) and FeatureCounts (Liao, Smyth & Shi, 2014) softwares. However, these counts cannot be directly used for further analysis and should be normalized to adjust between-sample differences. There is no standard tool for normalization, but the popular ones include deseq median ratio (Anders & Huber, 2010), trimmed mean of M values (TMM) (Robinson & Oshlack, 2010), reads per kilobase per million mapped reads (RPKM) (Mortazavi et al., 2008) and quantile (Bullard, 2010). For transformation, vst (Anders & Huber, 2010), rlog (Love, Huber & Anders, 2014) and voom (Law et al., 2014) methods can be a method of choice. Apart from these approaches, Witten considered power transformation to decrease the dispersion of data, before applying PLDA classifier (Witten, 2011). Once all mapped reads per transcripts are counted and normalized, we obtain gene-expression levels for each sample.

First way is to apply the count based classifiers, e.g. PLDA (Witten, 2011) and NBLDA (Dong et al., 2016) directly to the count data or to the power transformed data. Second way is to use the same workflow of microarray classification after transforming the data hierarchically to microarrays. The crucial steps of classification can be written as feature selection, building classification model and model validation. In feature selection step, we aim to work with an optimal subset of data. This process is crucial to reduce the computational cost, decrease of noise and improve the accuracy for classification of phenotypes, also to work with more interpretable features to better understand the domain (Ding & Peng, 2005). Various feature selection methods have been reviewed in detail and compared in (Xing, Jordan & Karp, 2001). Next step is model building, which refers to the application of a machine-learning algorithm and to learn the parameters of classifiers from training data. Thus, the built model can be used to predict class

memberships of new biological samples. The commonly used classifiers include SVM, RF and other tree-based classifiers, artificial neural networks and k-nearest neighbors.

In many real life problems, it is possible to experience that a classification algorithm may perform well and perfectly classify training samples, however perform poorly when classifying new samples. This problem is called as overfitting and independent test samples should be used to avoid overfitting and to generalize classification results. Holdout, k-fold cross-validation, leave-one-out cross-validation and bootsrapping are among the recommended approaches for model validation.

Implementation of classifiers

Simulation study

Simulation setup

A comprehensive simulation is conducted to investigate the effect of several parameters. Simulated datasets are generated under 864 different scenarios using a negative binomial model as follows:

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

where, s_i is the number of counts per sample, g_j is the number of counts per gene, d_{kj} is the differential expression probability of j^{th} gene between classes k and φ is the dispersion parameter. The datasets contain all possible combination of:

- different dispersion parameters as φ=0.01 (very slightly overdispersed), φ=0.1 (substantially overdispersed), φ=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;
- differential expression probability (d_{kj}) as 1%, 5% and 10%;

- number of classes (k) as 2, 3, 4;
- method of transformation as rlog and vst.

In simulation setup, s_i and g_j are distributed identically and independently as si and gj respectively. Simulated datasets are generated using the CountDataSet function of PoiClaClu package of R software (Witten, 2013) and manipulated based on the details given above. Seed number 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 model building processes are applied in training datasets, model performances are evaluated in test sets. We applied near-zero filtering to training data to filter the genes with low counts to eliminate the effect of this genes for further analysis (Kuhn, 2008).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), (ii) the ratio of the number of unique values to the sample size is less than 10%. Filtered genes are also excluded from the test datasets. Next, DESeq2 method is applied to detect the most DE 25, 50, 75 and 100 genes (Love, Huber & Anders, 2014). Selected genes are also selected in test datasets.

After selecting the DE genes, training data is normalized using median ratio approach to adjust sample specific differences (Love, Huber & Anders, 2014). After normalization, datasets are transformed using either rlog or vst transformation for SVM, bagSVM, RF and CART algorithms. Classical logarithmic transformation approach transforms the data into a less skewed distribution with less extreme values as well, however the genewise variances are still unstabilized (Love, Huber & Anders, 2014). Normalized count datasets are directly used for PLDA and NBLDA algorithms, since both algorithms use discrete probability distributions to fit

the models. In another scenario, a power transformation is applied to minimize the effect of overdispersion and PLDA algorithms is applied to this transformed data. This approach is defined as $PLDA_2$ in Results section. Note that, test datasets are normalized and transformed using the same parameters as training datasets. Since, training and test datasets should be in same scale and homoscedastic to each other. For instance, to normalize the test datasets, size factors of test datasets are calculated based on the geometric means of training data. Dispersion estimations are applied based on the training models as well.

After normalization and transformation processes, the parameters of each classifier are optimized to avoid overfitting and underfitting. A five-fold cross-validation is applied to training data and the parameters that achieves the highest accuracy rate are selected as optimal parameters. Same folds are used for each classifier to make the results comparable. Each classifier is fit with the optimal parameters. Fitted models are used in 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 of training and test sets. To overcome this limitation, we repeated the entire process 50 times and summarized the results in single statistics, 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).

Experimental datasets

Cervical dataset: Cervical dataset is a miRNA sequencing dataset obtained from (Witten et al., 2010). miRNAs are non-coding small RNA molecules with average 21-23 bp length and take

role in the regulation of gene expression. The objective of this study was to both identify the novel miRNAs and to detect the differentially expressed ones between normal and tumor cervical tissue samples. For this purpose, the authors constructed 58 small RNA libraries, prepared from 29 cervical cancer and 29 matched control tissues. After deep sequencing with Solexa/Illumina sequencing platform, they obtained a total of 25 Mb and 17 Mb RNA sequences from the normal and cancer libraries respectively. Of these 29 tumor samples, 21 of them had a diagnosis of squamous cell carcinomas, 6 of them had adenocarcinomas and 2 were unclassified. In our analysis, we used the data that contains the sequence read counts of 714 miRNAs belonging to 58 human cervical tissue samples, where 29 tumor and 29 non-tumor samples are treated as two distinct classes for prediction.

Alzheimer dataset: This dataset is another miRNA dataset provided by Leidinger et al. (Leidinger et al., 2013). The authors aimed to discover potential miRNAs from blood in diagnosing alzheimer and related neurological diseases. In this purpose, the authors obtained blood samples from 48 alzheimer patients that were evaluated after undergoing some tests including Alzheimer Disease Assessment Scale-cognitive subscale (ADAS-Cog), Wechsler Memory Scale (WMS), and Mini-Mental State Exam (MMSE) and Clinical Dementia Rating (CDR). A total of 22 age-matched control samples were obtained and all sample libraries were sequenced using 53 Illumina HiSeq2000 platform. After obtaining the raw read counts, the authors filtered the miRNAs with less than 50 counts in each group. 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: Renal cell cancer (RCC) dataset is an RNA-Seq dataset that is obtained from The Cancer Genome Atlas (TCGA) (Saleem et al., 2013). TCGA is a

comprehensive community resource platform for researchers to explore, download, and analyze datasets. We downloaded this dataset (with options level 3, RNASeqV2 data) from this database and obtained the raw 20,531 known human RNA transcript counts belonging to 1,020 RCC samples. This RNA-Seq data has 606, 323 and 91 specimens from kidney renal papillary cell (KIRP), kidney renal clear cell (KIRC) and kidney chromophobe carcinomas (KICH), respectively. These three classes are referred as the most common subtypes of RCC (account for nearly 90%-95% of the total malignant kidney tumors in adults) and treated as three separate classes in our analysis (Goyal et al., 2013).

Lung cancer dataset: Lung cancer is another RNA-Seq dataset provided from TCGA platform. Same options were used in the download process. The resulting count file contains the read counts of 20,531 transcripts of 1,128 samples. The dataset has two distinct classes including 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

A similar procedure is applied with the simulation study. Model building is applied in training (70%) and tested in the test (30%) sets. Near-zero filtering is applied to the training set. Filtered genes are also removed fromthe test set. For renal cell cancer and lung cancer datasets, 5,000 genes with highest variances are selected to eliminate the effect of non-informative mRNAs. All miRNA's are used in model building process for cervical and alzheimer datasets. Differential expression was performed to training data using DESeq2 method and genes are ranked from the most significant to the less with increasing number of genes in steps of 25 up to 250 genes. Selected differentially expressed genes in the training data are also selected in the test datasets. Differentially expressed genes in training data are normalized using median ratio approach and

transformed using either vst or rlog approaches. Similar to simulation experiments, test datasets are normalized based on the parameters obtained from the training data to make them in same scale and homoscedastic to each other. Since, the sample size of cervical and alzheimer miRNA datasets are relatively small, entire process is applied 50 times. Seed numbers in data selections are set between 1 to 50 and results are summarized based on these 50 repeats. Other model building process are applied as same as the simulation study.

Implementation of classifiers

Seven different algorithms are applied to both simulated and real datasets. In this section, we summarize the background and use of each method.

SVM:SVM is a classification method based on statistical learning theory, which is developed by Vapnik and his colleges, and has taken great attention because of its strong mathematical background, learning capability and good generalization ability (Vapnik, 2000). Moreover, SVM is capable of nonlinear classification and deal with high-dimensional data. Thus, it has been applied in many fields such as computational biology, text classification, image segmentation and cancer classification (Vapnik, 2000; Korkmaz, Zararsiz & Goksuluk, 2015).

In linearly separable cases, the decision function that correctly classifies the data points by their true class labels represented by:

$$f_{w,b} = sign(w.x_i + b) \quad (2)$$

$$i = 1,2,\ldots,n$$

In binary classification, SVM finds an optimal separating hyperplane in the feature space, which maximizes the margin and minimizes the probability of misclassification by choosing w and b in equation (2).For the linearly non-separable cases, slack variables $\{\xi_1,\ldots,\xi_n\}$, which is a penalty introduced by Cortes and Vapnik, can be used to allow misclassified data points, where

ξ_i>0(Cortes & Vapnik, 1995). In many classification problems, the separation surface is nonlinear. In this case, SVM uses an implicit mapping **Φ** of the input vectors to a high-dimensional space defined by a kernel function ($\boldsymbol{K(x,y)=\Phi(x_i)\Phi(x_j)}$) and the linear classification then takes place in this high-dimensional space. The most widely used kernel functions are linear :$\boldsymbol{K(x,y)=x_ix_j}$, polynomial: $\boldsymbol{K(x,y)=(x_ix_j+1)^d}$, radial basis function: $\boldsymbol{K(x,y)=\exp(-\gamma\|x_i-x_j\|^2)}$ and sigmoidal: $\boldsymbol{K(x,y)=tanh(k(x_ix_j)-c)}$, where d is the degree, γ>0 sometimes parametrized as $\boldsymbol{\gamma=1/2\sigma^2}$, and *c* is a constant. Normalized and transformed (either using vst or rlog) datasets are used as input to SVM classifier. Radial basis kernel function is used in the analysis.

BagSVM: BagSVM is a bootstrap ensemble method, which creates individuals for its ensemble by training each SVM classifier (learning algorithm) on a random subset of the training set. For a given data set, multiple SVM classifiers are trained independently through a bootstrap method and they are aggregated via an aggregation technique. To construct the SVM ensemble, k replicated training sets are generated by randomly re-sampling, but with replacement, from the given training set repeatedly. Each sample, $\boldsymbol{x_i}$, in the given training set, may appear repeated times, or not at all, in any particular replicate training set. Each replicate training set will be used to train a specific SVM classifier. Normalized and transformed (either using vst or rlog) datasets are used as input to BagSVM classifier. Number of bootstrap samples were set to 101, since small changes were observed over this number.

CART: CART, which is introduced by Breiman et al., is one of the most popular tree classifiers and applied in many fields (Breiman et al., 1986). It uses Gini index to choose the split which maximizes the decrease in impurity at each node. If $p(i|j)$ is the probability of class*i* at node *j*, then the Gini index is $1-\sum_i p^2(i|j)$. When CART grows a maximal tree, this tree is pruned upward to get a decreasing sequence of subtrees. Then, a cross-validation is used to identify the subtree

that having the lowest estimated misclassification rate. Finally, the assignment of each terminal node to a class is performed by choosing the class that minimizes the resubstitution estimate of the misclassification probability(Breiman et al., 1984; Dudoit & Fridlyand, 2003). Normalized and transformed (either using vst or rlog) datasets are used as input to CART classifier.

RF: A random forest is a collection of many CART trees combined by averaging the predictions of individual trees in the forest (Breiman, 2001).The idea behind the RF is to combine many weak classifiers to produce a significantly better strong classifier. For each tree, a training set is generated by bootstrap sample from the original data. This bootstrap sample includes 2/3 of the original data. The remaining of the cases are used as a test set to predict out-of-bag error of classification. If there are m features, m_{try} out of m features are randomly selected at each node and the best split is used to split the node. Different splitting criteria can be used such as Gini index, information gain and node impurity. The value of m_{try} is chosen to be approximately either $\frac{\sqrt{m}}{2}$or $\sqrt{m}$or $2\sqrt{m}$and constant during the forest growing. An unpruned tree is grown for each of the bootstrap sample, unlike CART. Finally, new data is predicted by aggregating, i.e. majority votes, the predictions of all trees (Liaw & Wiener, 2002; Okun & Priisalu, 2007). Normalized and transformed (either using vst or rlog) datasets are used as input to RF classifier. Number of trees was set to 500 in the analysis.

$PLDA_1$ and $PLDA_2$: Let $\boldsymbol{X}$ be an nxp matrix of sequencing data, where n is number of observations and p is number of features. For sequencing data, X_{ij} indicates the total number of reads mapping to gene j in observation i. Therefore, Poisson log-linear model can be used for sequencing data,

$$X_{ij} \sim Poisson(N_{ij}), \quad N_{ij} = s_i g_j \quad (3)$$

where s_i is total number of reads per sample and g_j is total number of reads per region of interest. For RNA-Seq data, equation (3) can be extended as follows,

$$X_{ij}|y_i = k \sim Poisson(N_{ij}d_{jk}), \qquad N_{ij} = s_i g_j \qquad (4)$$

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

Let $(x_i, y_i), i = 1,\ldots,n$, be a training set and $x^* = (X^*_1,\ldots,X^*_p)^T$ be a test set. Using the Bayes' rule as follows,

$$P(y^* = k|x^*) \propto f_k(x^*)\pi_k \qquad (5)$$

where y^* denotes the unknown class label, f_k is the density of an observation in class k and π_k is the prior probability that an observation belongs to class k. If f_k is a normal density with a class-specific mean and common variance then a standard LDA is used for assigning a new observation to the class (Hastie, Tibshirani & Friedman, 2009). In case of the observations are normally distributed with a class-specific mean and a common diagonal matrix, then diagonal LDA methodology is used for the classification (Dudoit, Fridlyand & Speed, 2001). However, neither normality nor common covariance matrix assumptions are not appropriate for sequencing data. Instead, Witten (Witten, 2011) assumes that the data arise from following: Poisson model,

$$X_{ij}|y_i = k \sim Poisson(N_{ij}d_{kj}), \qquad N_{ij} = s_i g_j \quad (6)$$

where y_i represents the class of the i^{th} observation and the features are independent. The equation (4) specifies that $X^*_j|y^* = k \sim Poisson(s^* g_j d_{kj})$. First, the size factors for the training data, $s_1,\ldots,s_n$, is estimated. Then s^*, g_j, d_{kj} and π_k are estimated as described in (Witten, 2011). Substituting these estimations into equation (4) and recalling independent features assumption, equation (5) produces,

$$logP\left(\widehat{y^* = k}|x^*\right) = \log \hat{f}_k(x^*) + \log \hat{\pi}_k + c$$

$$= \Sigma_{j=1}^{p} X^*_j \widehat{logd_{kj}} - s^* \Sigma_{j=1}^{p} \hat{g}_j log\hat{d}_{kj} + \log \hat{\pi}_k + c', \quad (7)$$

where c and c' are constants and do not depend on the class label. The classification rule that assigns a new observation to the one of the classes for which equation (7) is the largest and it is linear in x^* (Witten, 2011).

Normalized count data is used as input to $PLDA_1$ 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 $PLDA_2$ classifier. To optimize the tuning parameter, a grid search (30 searches) is applied and the sparsest model with the highest accuracy rates are selected for classification.

NBLDA: Dong et al. generalized that PLDA using an extra dispersion parameter (φ) of negative binomial distribution and named the method as negative binomial linear discriminant analysis (NBLDA)(Dong et al., 2016). This extra dispersion parameter is estimated using a shrinkage approach detailed in (Yu, Huber & Vitek, 2013). A new test observation will be assigned to its class based on the following NBLDA discriminating function:

$$logP\left(\widehat{y^* = k}|x^*\right) = \sum_{j=1}^{p} X^*_j \left[\widehat{logd_{kj}} - \log\left(1 + s^* \hat{g}_j d_{kj} \phi_j\right)\right] -$$

$$\Sigma_{j=1}^{p} \phi^{-1}_j \log\left(1 + s^* \hat{g}_j d_{kj} \phi_j\right) + \log \hat{\pi}_k + c', \quad (8)$$

Decreasing the dispersion parameter will approximate the data distribution from negative binomial to poisson, thus will approximate NBLDA to PLDA. More details about this method can be found in (Dong et al., 2016).

Evaluation criteria

To validate each classifier model, 5-fold cross-validation was used, repeated 10 times and accuracy rates were calculated to evaluate the performance of each model. Same folds are used 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 one-versus-all approach. Since, class sizes are unbalanced in alzheimer and renal cell cancer datasets, accuracies are balanced using the formula: $(Sensitivity + Specificity) / 2$.

MLSeq R/BIOCONDUCTOR Package

We presented an R package in BIOCONDUCTOR network to make RNA-Seq classification less complicated for researchers and allow users to fit classifiers using single functions. MLSeq package requires from users to upload their raw count data in which can be obtained from feature counting tools (e.g. HTSeq (Anders, Pyl & Huber, 2014), bedtools (Quinlan & Hall, 2010) and FeatureCounts (Liao, Smyth & Shi, 2014) etc.) and allow them to normalize, transform and build classifiers including SVM, bagSVM, RF and CART. Users can access MLSeq package from https://www.bioconductor.org/packages/release/bioc/html/MLSeq.html .

Results and Discussion

Datasets and Classifiers

A comprehensive simulation study is designed under 864 different scenarios. Negative binomial distribution is used in all simulation settings. Simulated datasets contain possible combinations of different dispersion parameters, number of biological samples, number of differentially expressed genes, differential expression rate, number of class and transformation method. Moreover, four real mRNA (lung and renal cell cancer) and miRNA (alzheimer and cervical cancer) datasets were used alongside the simulated datasets (Table 1).Support vector machines (SVM), bagging support vector machines (bagSVM), random forests (RF), classification and regression trees (CART), Poisson linear discriminant analysis without power transformation (PLDA1), Poisson linear discriminant analysis with power transformation (PLDA2) and negative binomial linear discriminant analysis (NBLDA) classifiers were applied to each simulated and real datasets. More detailed information about the datasets, classifiers and analysis settings can be found in Methods section.

Experimental Results and Discussion

Genewise dispersion parameters are estimated for each classifier with method of moments approach and given in Fig. 2. It is seen from the figure that cervical and alzheimer miRNA datasets are very highly overdispersed, while lung and renal cell cancer datasets are substantially overdispersed. Simulation results for k=2, d_{kj}=10% for vst and rlog transformations are given in Fig. 3 and Fig. 4. All other simulation results are given in http://www.biosoft.hacettepe.edu.tr/MLSeqSupplementary/ and in Supp. file-1. More detailed results are given in Supp. file-2. 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, increased number of differentially expressed genes leads to an increase in the classification accuracy (Fig. 4-5). Similarly, in most scenarios, working with more samples and genes has a positive impact on the overall model accuracies. This relationship between number of genes and accuracy is mostly available in d_{kj}=10% scenarios. Likewise, slight increases is observed in real dataset classification accuracies, since this leads to an increase in the probability of a differentially expressed gene to be included into classification model. For PLDA classifier, high number of selected genes provides alternative options for the lasso shrinkage method to test more genes in classification models. On the other hand, RF builds trees with bagging approach, thus using more genes, and enhances its probability to specify the optimal tree. Increasing sample size improves the discrimination power, as well as the classification accuracy. Conversely, overall accuracies decrease as the number of classes increases. This is due to the fact that the misclassification probability of an observation may be arised depending on the increase in class number.

Dispersion effect on classification accuracies

The performance of each method was increasing depending on the decrease in dispersion parameter. In fact, only decreasing the dispersion parameter makes a significant contribution to classification accuracy, even for the same data and the same scenario. This is mostly clear in k=2 and d_{kj}=10% scenarios. When the data is overdispersed, the variance increases; thus we need more sample sizes to achieve the same discrimination power. When we stabilize the sample size and increase the dispersion parameter, this will decrease the discrimination power and lead to a decrease in the classification accuracies. Nagalakshmi et al. mentioned that using biological replicates instead of technical replicates leads to an increase in the dispersion of the data

(Nagalakshmi et al., 2008). Based on this idea, increasing the biological variance of the observations will lead to an increase in the data dispersion, thus the classification of observations will be much harder. In differential expression studies of RNA-Seq data, overdispersion is one of the major problems in analysis settings. Many studies are made to overcome this problem (Robinson, McCarthy & Smyth, 2010; Robinson & Oshlack, 2010; Love, Huber & Anders, 2014; Anders & Huber, 2012; Law et al., 2014). When we look at the classification accuracy results, overdispersion seems to be a major challenge in classification studies as well. Unless we work with technical replicates, RNA-Seq data is overdispersed and that leads for same gene, counts from different biological replicates have variance exceeding the mean (Nagalakshmi et al., 2008). This overdispersion can be seen in other studies (Robinson & Smyth, 2007, Bloom et al., 2009; Robinson, McCarthy & Smyth, 2010; Zhou, Xia & Wright, 2011; Auer & Doerge, 2011). 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 are developed and able to work in large *p* and small *n* settings. However, the technological improvements makes RNA-Seq state-of-the-art approach for quantified transcriptomics. Currently, much of these microarray based classifiers are no longer to be applied to RNA-Seq data, because of the different data types of microarrays and RNA-Seq. Microarray data consists the continuous log-intensities of probes, while RNA-Seq data consists the discrete and overdispersed mapped read counts of sequencing technologies. Results of this study revealed that, transforming the data hierarchically to microarrays (e.g. through rlog and vst) will be a proper approach to recover these classifiers for RNA-Seq classification.

Witten et al. stated that normalization strategy has little impact on the classification performance but may be important in differential expression analysis (Witten, 2011). However, data transformation has a direct effect on classification results, by changing the distribution of data. In this study, we used deseq normalization with vst and rlog transformations and had satisfactory classification performances. Love et al. discussed that vst transformation does not consider the size factors during the transformation (Love, Huber & Anders, 2014). However, there were no substantial differences between rlog and vst transformation approaches on classification accuracies. Both transformations can be applied with microarray based classifiers.

Power transformed PLDA and other count based classifiers

Without transformation, PLDA seemed to perform well in very slightly overdispersed datasets. This can be seen in both simulated and real datasets (Fig. 5). For instance, in renal cell carcinoma dataset, the dispersion parameter is very low and the data seem to follow a Poisson distribution. In this dataset, $PLDA_1$ and $PLDA_2$ shows similar performances (Fig. 5). However, the performance of this method decreases, when the data becomes more overdispersed. The reason is that PLDA classifies the data using a model based on Poisson distribution. It minimizes the dispersion parameter and makes a significant improvement on classification accuracy using a power transformation (Witten, 2011). Therefore, we suggest that this transformation is very useful and should be applied to be used with PLDA classifier, even in very slightly overdispersed datasets. NBLDA extends this classifier using a negative binomial model. However, classification accuracies of this method is not as higher as PLDA with power transformation. Hence, we believe that this may be due to the dispersion parameter estimation or the unsparsed property of the classifier. We conclude that, novel count-based classifiers are still needed for accurate and robust classification of RNA-Seq data.

Overall performances of classifiers

In simulated datasets, power transformed PLDA performed to be the best classifier. RF and NBLDA performed moderately similar. On the other hand, SVM and bagSVM performed the highest classification accuracies in real datasets. $PLDA_2$, RF and NBLDA have still comparable and high classification accuracies, but lower than SVM and bagSVM. This slight differences may arise from the differences between negative binomial distribution which is used in simulation settings and exact distributions of real RNA-Seq data. In real datasets, SVM and bagSVM classifiers put forward their classification abilities. Moreover, it can be seen from the simulated and real datasets that, the performance of bagSVM classifier increases as the sample size increases. A possible explanation for such observation is that bagSVM uses bootstrap technique and trains better models in datasets with high number of samples. The performance of CART and $PLDA_1$ were seemed to be lower than the other classifiers.

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 unsparsed classifiers to this study, we leave the effect of 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 of cancer classification for diagnosis and prognosis purposes. Although microarray-based gene expression studies through a combination of classification algorithms such as SVM and feature selection techniques have recently been widely used for new biomarkers for cancer diagnosis (Lee, 2008; Statnikov, Wang & Aliferis, 2008; Anand & Suganthan, 2009; George & Raj, 2011), it has its own limitations in terms of novel transcript discovery and abundance estimation with large dynamic range. Thus, one choice is to utilize the power of RNA-Seq techniques in the analysis of transcriptome for diagnostic classification to surpass the limitations of microarray-based experiment. As mentioned in earlier sections, working with less noisy data can enhance the predictive performance of classifiers, and the novel transcripts may be a biomarker in interested disease or phenotypes.

Hundreds of studies are published for microarray based classification. The goal of these studies were to develop or adapt novel approaches to identify a small subset of genes and predict the class labels of a new observation. This has a particular importance in biomedical studies for 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 quantification of gene expression. We designed a comprehensive simulation study and also used four real experimental miRNA and mRNA datasets.

Besides its technological advantages of RNA-Seq as compared to microarrays, the data obtained from this method is overdispersed due to the inherent variability. This overdispersion seemed to be a drawback for differential expression studies of RNA-Seq data. In this study, we showed that

517 this overdispersion is also a drawback for classification studies, since an increase in the variance
518 will lead to a decrease in the discrimination power. We reach a conclusion that three solutions
519 are available to handle classification of overdispersed RNA-Seq data: (i) increasing the sample
520 size, (ii) transforming the data hierarchically closer to microarrays with variance stabilizers, e.g.
521 vst and rlog transformations, (iii)using count based classifiers, e.g. $PLDA_2$ and NBLDA.Our
522 simulation study revealed that both microarray based classifiers after an rlog/vst transformations
523 and count based classifiers (that are dealing with the overdispersion) can be efficiently used for
524 classification of RNA-Seq data.
525 To make an overall assessment for the performances of classifiers, PLDA after a power
526 transformation may be a good choice as a count based classifier. Furthermore, its sparsity seems
527 to be an advantage for researchers, however further researches are needed. Surprisingly, the
528 performance of the NBLDA was not satisfactory enough as a count based classifier. Dong et al.
529 mentioned that NBLDA has a better performance than PLDA in moderate and highly
530 overdispersed data (Dong et al., 2016). However, these comparisons are made with same number
531 of genes. Our analysis are performed based on the sparse PLDA classifiers, where the best subset
532 of genes are used in classification. Sparse PLDA classifier after a power transformation
533 performed more accurately in all dispersion settings. We believe that extending NBLDA
534 algorithm into a sparse classifier may improve its classification performance by selecting the
535 most significant genomic features.
536 Moreover, an alternative option may be to transform the data hierarchically closer to microarrays
537 and perform microarray based classifiers. Our results revealed that RF, SVM and bagSVM may
538 perform accurate results after an rlog or vst transformation. Moreover, the efficiency of the
539 bagSVM is improved observably with the increasing sample size.

We conclude that, the data with less overdispersion, highly differentially expressed genes, lower number of groups and large sample size may improve the accuracy of the classifiers. Finally, we developed an R/BIOCONDUCTOR package, MLSeq, to make the computation less complicated for researchers and allow them to learn a classification model using various classifiers with one single function. This package can be accessed and downloaded through https://www.bioconductor.org/packages/release/bioc/html/MLSeq.html.

Supplemental Information

Supp-1. All figures for simulation results

Supp-2. MLSeq package source

Supp-3. Simulation R Codes

Supp-4. Computational Infrastructure

Supp-5. Computational costs of classifiers

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

GZ developed the method's framework, DG and SK contributed to algorithm design and implementation. GZ, VE and IPD surveyed the literature for other available methods and collected performance data for the other methods used in the study for comparison. GZ, VE, IPD, DG carried out the simulation studies and data analysis. GZ, DG and SK developed MLSeq package. GZ, DG, SK and VE wrote the paper, TU and AO supervised the research process, revised and directed the manuscript and contributed statistical concepts and ideas. All authors read and approved the final manuscript.

Funding

This research was by the Research Fund of Erciyes University [TDK-2015-5468] and Istanbul University [29506].The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

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

References

Anand A, and Suganthan PN. 2009. Multiclass cancer classification by support vector machines with class-wise optimized genes and probability estimates. *J Theor Biol* 259:533-540. 10.1016/j.jtbi.2009.04.013

Anders S, and Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* 11:R106. 10.1186/gb-2010-11-10-r106

Anders S, and Huber W. 2012. Differential expression of RNA-Seq data at the gene level–the DESeq package. Heidelberg, Germany: European Molecular Biology Laboratory (EMBL).

Anders S, Pyl PT, and Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166-169. 10.1093/bioinformatics/btu638

Auer PL, and Doerge RW. 2011. A two-stage Poisson model for testing RNA-seq data. *Statistical Applications in Genetics and Molecular Biology* 10:1-26.

Bloom JS, Khan Z, Kruglyak L, Singh M, and Caudy AA. 2009. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics* 10:221. 10.1186/1471-2164-10-221

Breiman L. 2001. Random forests. Machine learning 45:5-32.

Breiman L, Friedman J, Stone CJ, and Olshen RA. 1984. Classification and regression trees: CRC press.

Bullard JH, Purdom E, Hansen KD, and Dudoit S. 2010.Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.*BMC Bioinformatics* 11:94. 10.1186/1471-2105-11-94

Cortes C, and Vapnik V. 1995. Support-vector networks. *Machine learning* 20:273-297.

591 Di Y, Schafer DW, Cumbie JS, and Chang JH. 2011. The NBP negative binomial model for
592 assessing differential gene expression from RNA-Seq. *Statistical Applications in Genetics and*
593 *Molecular Biology* 10.
594 Díaz-Uriarte R, and De Andres SA. 2006. Gene selection and classification of microarray data
595 using random forest. *BMC Bioinformatics* 7:3.
596 Ding C, and Peng H. 2005. Minimum redundancy feature selection from microarray gene
597 expression data.*J Bioinform Comput Biol* 3:185-205.
598 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and
599 Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15-21.
600 10.1093/bioinformatics/bts635
601 Dong K, Zhao H, Tong T, and Wan X. 2016. NBLDA: negative binomial linear discriminant
602 analysis for RNA-Seq data. *BMC Bioinformatics* 17:369. 10.1186/s12859-016-1208-1
603 Dudoit S, Fridlyand J, and Speed TP. 2001. Comparison of discrimination methods for the
604 classification of tumors using gene expression data.*Journal of the American statistical*
605 *association* 97:77-87.
606 Dudoit S, and Fridlyand J. 2003. Classification in microarray experiments.Statistical analysis of
607 gene expression microarray data 1:93-158.
608 Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, and Haussler D. 2000. Support
609 vector machine classification and validation of cancer tissue samples using microarray
610 expression data. *Bioinformatics* 16:906-914.
611 George G, and Raj VC. 2011. Review on feature selection techniques and the impact of SVM for
612 cancer classification using gene expression profile. *arXiv* preprint arXiv:11091062.

613 Goyal R, Gersbach E, Yang XJ, and Rohan SM. 2013. Differential diagnosis of renal tumors
614 with clear cytoplasm: clinical relevance of renal tumor subclassification in the era of targeted
615 therapies and personalized medicine. *Arch Pathol Lab Med* 137:467-480. 10.5858/arpa.2012-
616 0085-RA
617 Hastie T, Tibshirani R, and Friedman J. 2009. The elements of statistical learning 2:1. NY
618 Springer.
619 Kuhn M. 2008. Building Predictive Models in R Using the caret Package. *J Stat Softw*:28-5
620 Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL. 2013. TopHat2: accurate
621 alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome*
622 *Biol* 14:R36. 10.1186/gb-2013-14-4-r36
623 Korkmaz S, Zararsiz G, and Goksuluk D. 2015. MLViS: A Web Tool for Machine Learning-
624 Based Virtual Screening in Early-Phase of Drug Discovery and Development. *PLoS One*
625 10:e0124600. 10.1371/journal.pone.0124600
626 Law CW, Chen Y, Shi W, and Smyth GK. 2014. voom: Precision weights unlock linear model
627 analysis tools for RNA-seq read counts. *Genome Biol* 15:R29. 10.1186/gb-2014-15-2-r29
628 Lee ZJ. 2008. An integrated algorithm for gene selection and classification applied to microarray
629 data of ovarian cancer. *Artif Intell Med* 42:81-93. 10.1016/j.artmed.2007.09.004
630 Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, Haas J, Ruprecht K, Paul
631 F, Stahler C, Lang CJ, Meder B, Bartfai T, Meese E, and Keller A. 2013. A blood based 12-
632 miRNA signature of Alzheimer disease patients. *Genome Biol* 14:R78. 10.1186/gb-2013-14-7-
633 r78

634 Liao Y, Smyth GK, and Shi W. 2014.featureCounts: an efficient general purpose program for
635 assigning sequence reads to genomic features. *Bioinformatics* 30:923-930.
636 10.1093/bioinformatics/btt656
637 Liaw A, and Wiener M. 2002.Classification and regression by randomForest.*R new*s 2:18-22.
638 Love MI, Huber W, and Anders S. 2014.Moderated estimation of fold change and dispersion for
639 RNA-seq data with DESeq2.*Genome Biol* 15:550. 10.1186/s13059-014-0550-8
640 Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, and Gentleman R. 2009. ShortRead: a
641 bioconductor package for input, quality assessment and exploration of high-throughput sequence
642 data. *Bioinformatics* 25:2607-2608. 10.1093/bioinformatics/btp450
643 Mortazavi A, Williams BA, McCue K, Schaeffer L, and Wold B. 2008. Mapping and quantifying
644 mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621-628. 10.1038/nmeth.1226
645 Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, and Snyder M. 2008. The
646 transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320:1344-
647 1349. 10.1126/science.1158441
648 Okun O, and Priisalu H. 2007. Random forest for gene expression based cancer classification:
649 overlooked issues. Iberian Conference on Pattern Recognition and Image Analysis: *Springer*. p
650 483-490.
651 Quinlan AR, and Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
652 features. *Bioinformatics* 26:841-842. 10.1093/bioinformatics/btq033
653 Rapaport F, Zinovyev A, Dutreix M, Barillot E, and Vert JP. 2007. Classification of microarray
654 data using gene networks. *BMC Bioinformatics* 8:35. 10.1186/1471-2105-8-35

655 Robinson MD, McCarthy DJ, and Smyth GK. 2010. edgeR: a Bioconductor package for
656 differential expression analysis of digital gene expression data. *Bioinformatics* 26:139-140.
657 10.1093/bioinformatics/btp616
658 Robinson MD, and Oshlack A. 2010.A scaling normalization method for differential expression
659 analysis of RNA-seq data.*Genome Biol* 11:R25. 10.1186/gb-2010-11-3-r25
660 Robinson MD, and Smyth GK. 2007. Moderated statistical tests for assessing differences in tag
661 abundance. *Bioinformatics* 23:2881-2887. 10.1093/bioinformatics/btm453
662 Saleem M, Padmanabhuni SS, Ngomo A-CN, Almeida JS, Decker S, and Deus HF. 2013.
663 Linked cancer genome atlas database. *Proceedings of the 9th International Conference on*
664 *Semantic Systems*: ACM. p 129-134.
665 Shrestha RK, Lubinsky B, Bansode VB, Moinz MB, McCormack GP, and Travers SA. 2014.
666 QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454
667 sequencing platform. *BMC Bioinformatics* 15:33.
668 Soneson C, and Delorenzi M. 2013.A comparison of methods for differential expression analysis
669 of RNA-seq data.*BMC Bioinformatics* 14:91. 10.1186/1471-2105-14-91
670 Statnikov A, Wang L, and Aliferis CF. 2008. A comprehensive comparison of random forests
671 and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*
672 9:319. 10.1186/1471-2105-9-319
673 Vapnik VN. 2000. The nature of statistical learning theory, ser. Statistics for engineering and
674 information science. New York: *Springer* 21:1003-1008.
675 Wang Z, Gerstein M, and Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics.
676 *Nat Rev Genet* 10:57-63. 10.1038/nrg2484

677 Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm
678 SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, and Liu J. 2010. MapSplice: accurate
679 mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* 38:e178.
680 10.1093/nar/gkq622
681 Witten D. 2013. PoiClaClu: Classification and clustering of sequencing data based on a Poisson
682 model. R package version 1.0.2.,https://CRAN.R-project.org/package=PoiClaClu
683 Witten DM. 2011.Classification and clustering of sequencing data using a Poisson model. The
684 Annals of Applied Statistics:2493-2518.
685 Witten D, Tibshirani R, Gu SG, Fire A, and Lui WO. 2010. Ultra-high throughput sequencing-
686 based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical
687 tumours and matched controls. *BMC Biol* 8:58. 10.1186/1741-7007-8-58
688 Xing EP, Jordan MI, and Karp RM. 2001. Feature selection for high-dimensional genomic
689 microarray data.*ICML*: Citeseer. p 601-608.
690 Yu D, Huber W, and Vitek O. 2013. Shrinkage estimation of dispersion in Negative Binomial
691 models for RNA-seq experiments with small sample size. *Bioinformatics* 29:1275-1282.
692 10.1093/bioinformatics/btt143
693 Zhou YH, Xia K, and Wright FA. 2011. A powerful and flexible approach to the analysis of
694 RNA sequence count data. *Bioinformatics* 27:2672-2678. 10.1093/bioinformatics/btr449
695 Zhu J, and Hastie T. 2004. Classification of gene microarrays by penalized logistic regression.
696 *Biostatistics* 5:427-443. 10.1093/biostatistics/5.3.427

Figure legends

Fig. 1. RNA-Seq classification workflow

Fig. 2. Genewise dispersion estimations for real datasets

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

Fig. 4. 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 (φ=*0.01:* very slight, φ=*0.1:* substantial, φ=*1:* very high)

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2761v2 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017

Table 1(on next page)

Description of real RNA-Seq datasets used in this study

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

1 **Table**

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

Dataset	Number of groups	Sample size	Number of features
Cervical cancer (Witten et al., 2010)	2	58 (29 cervical cancer, 29 control)	714 miRNAs
Alzheimer (Leidinger et al., 2013)	2	70 (48 alzheimer, 22 control)	416 miRNAs
Renal cell cancer (Saleem et al., 2013)	3	1,020 (606 KIRP, 323 KIRC, 91 KICH)	20,531 mRNAs
Lung cancer (Saleem et al., 2013)	2	1,128 (576 LUAD, 552 LUSC)	20,531 mRNAs

3

Figure 1(on next page)

RNA-Seq classification workflow

Fig 1 - RNA-Seq classification workflow

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2761v2 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017

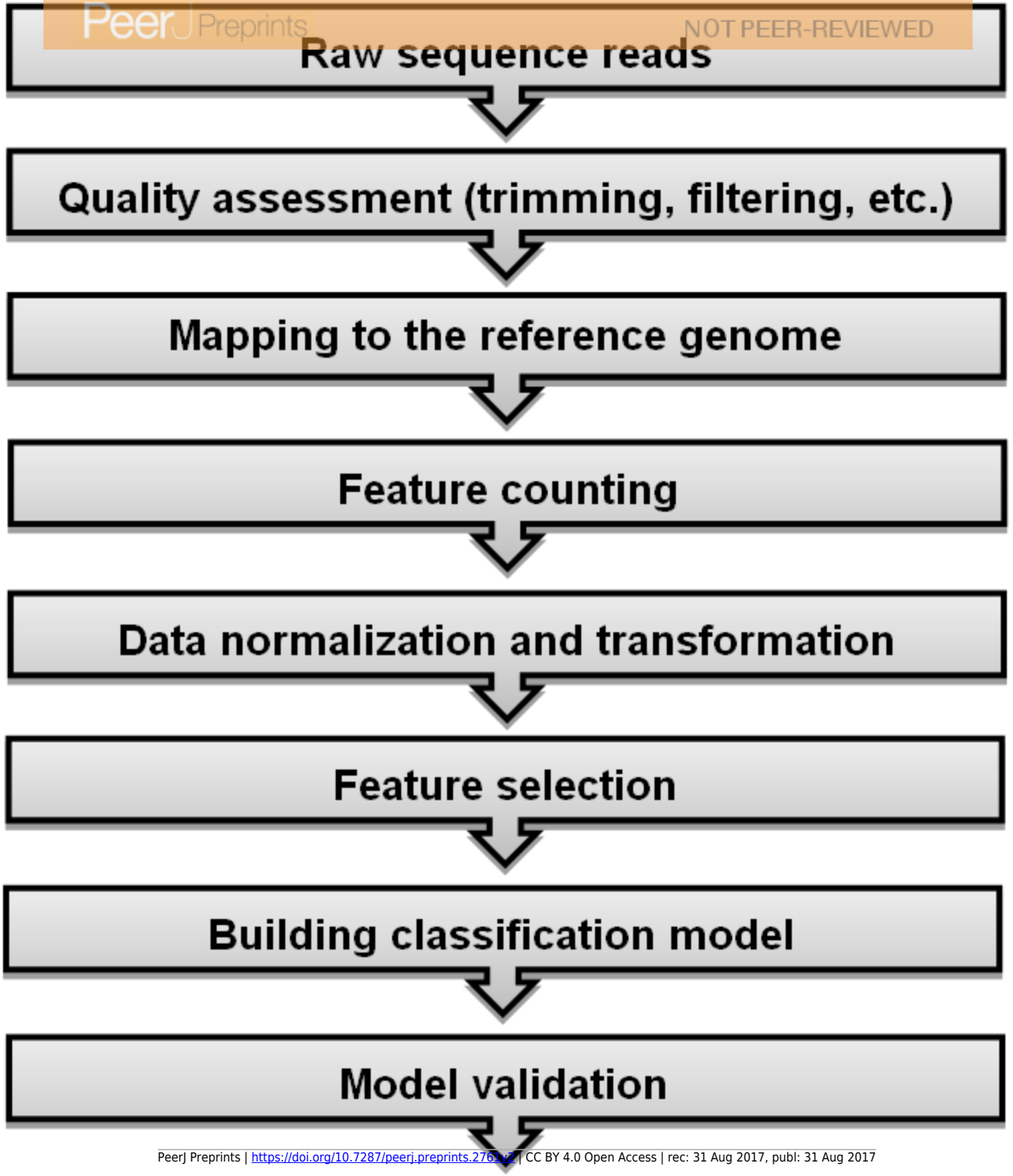


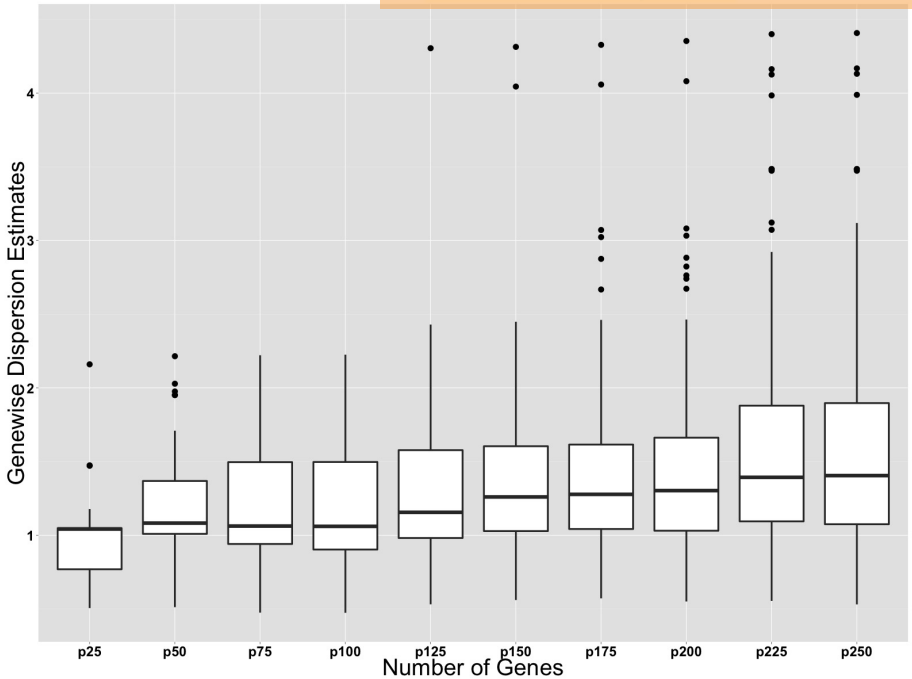
Figure 2(on next page)

Genewise dispersion estimations for real datasets

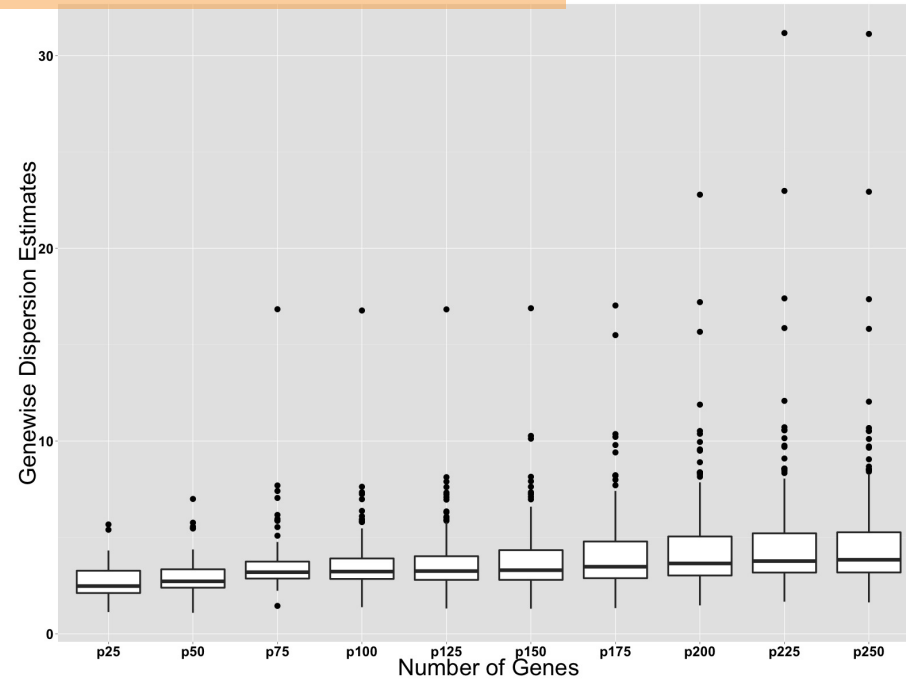
Fig 2 - Genewise dispersion estimations for real datasets

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2761v2 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017

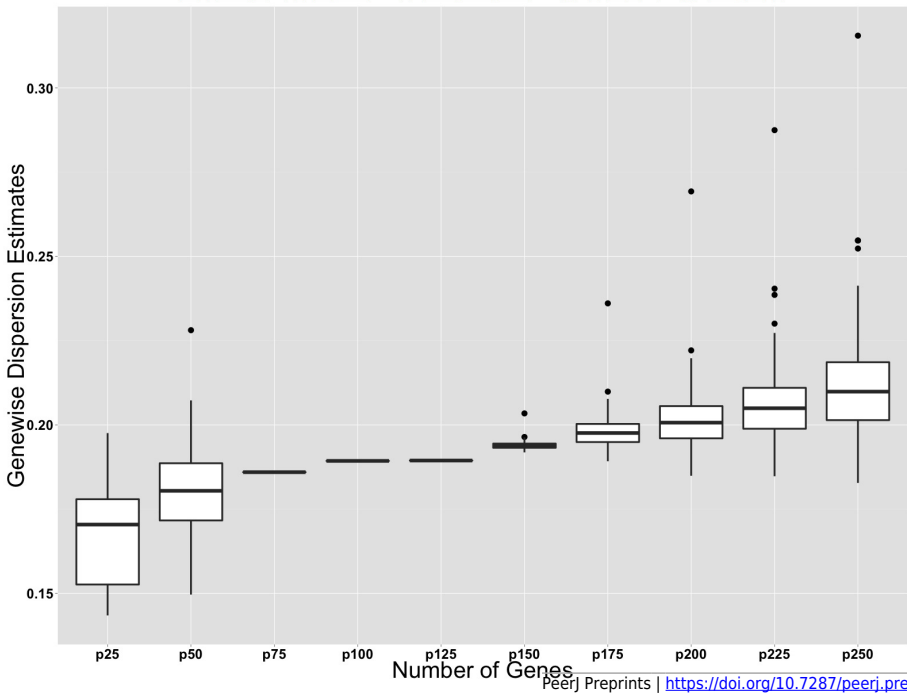
ALZHEIMER



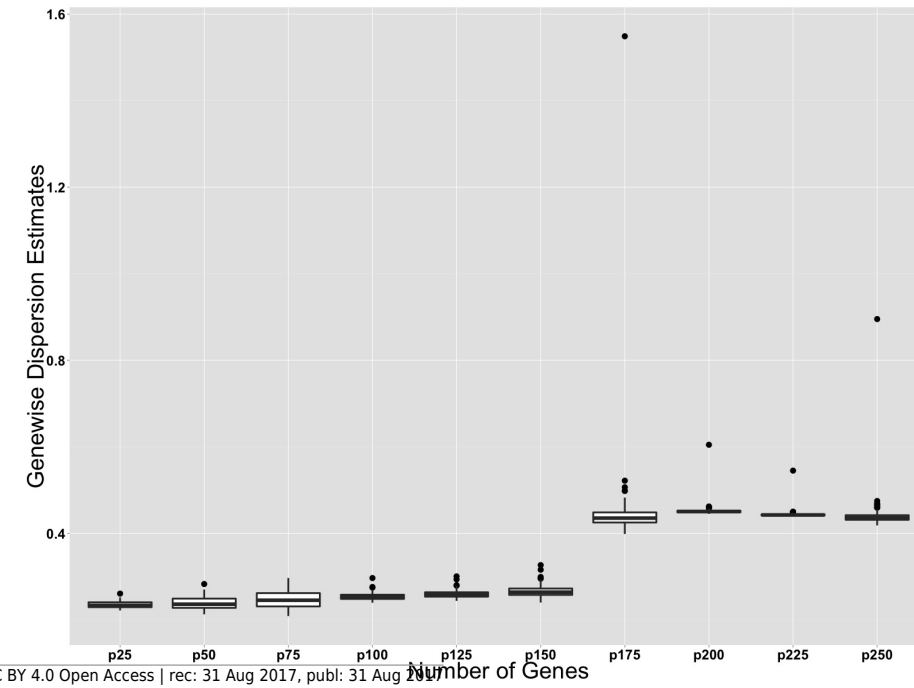
CERVICAL



RENAL CELL CANCER



LUNG

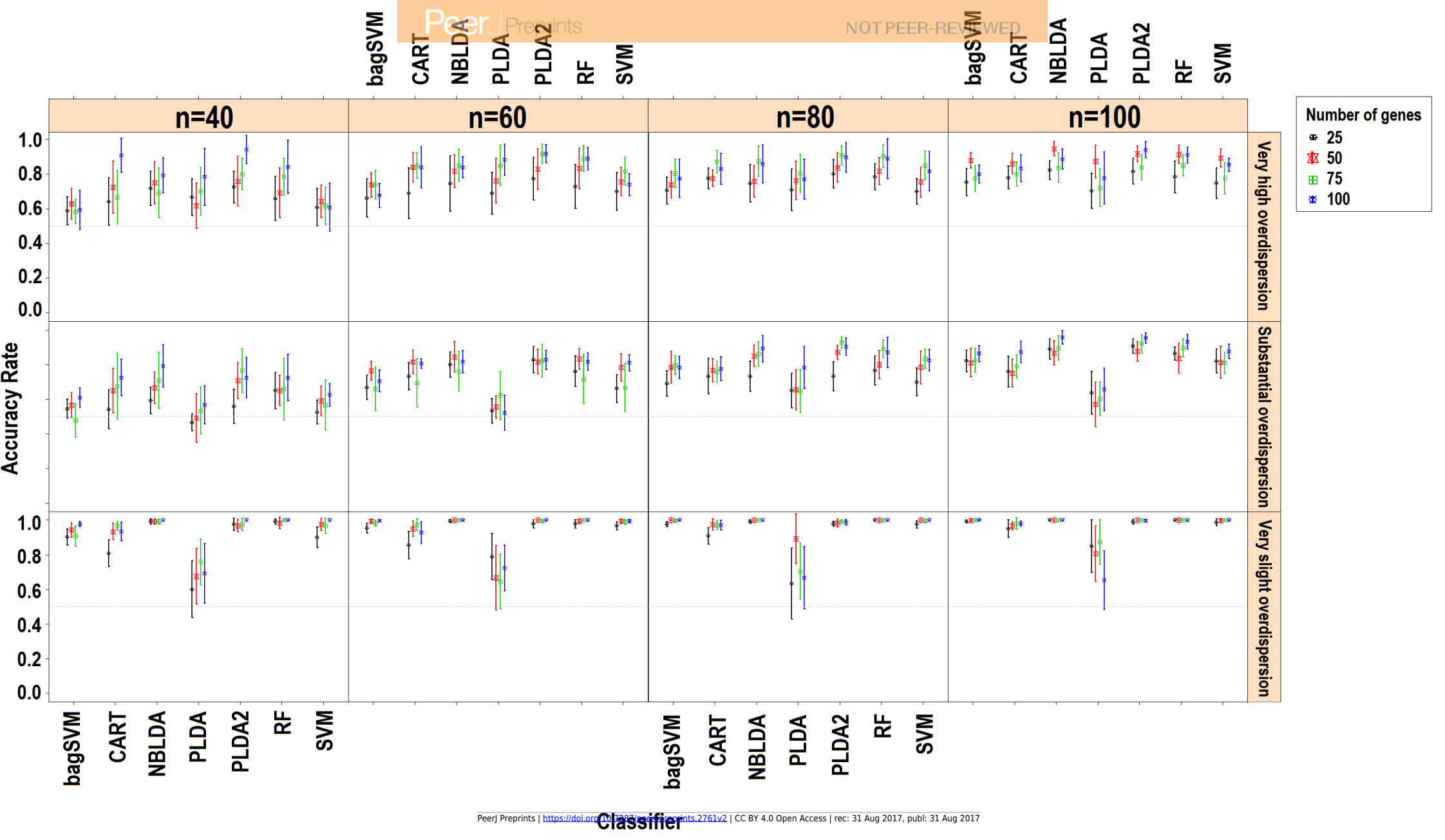


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2761v2 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017

Figure 3(on next page)

Simulation results for$k$=2,d_{kj}=10%, transformation: vst. Figure shows the performance results of classifiers with changing parameters of sample size (n), number of genes (p) and type of dispersion (φ=*0.01:* very sligh

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

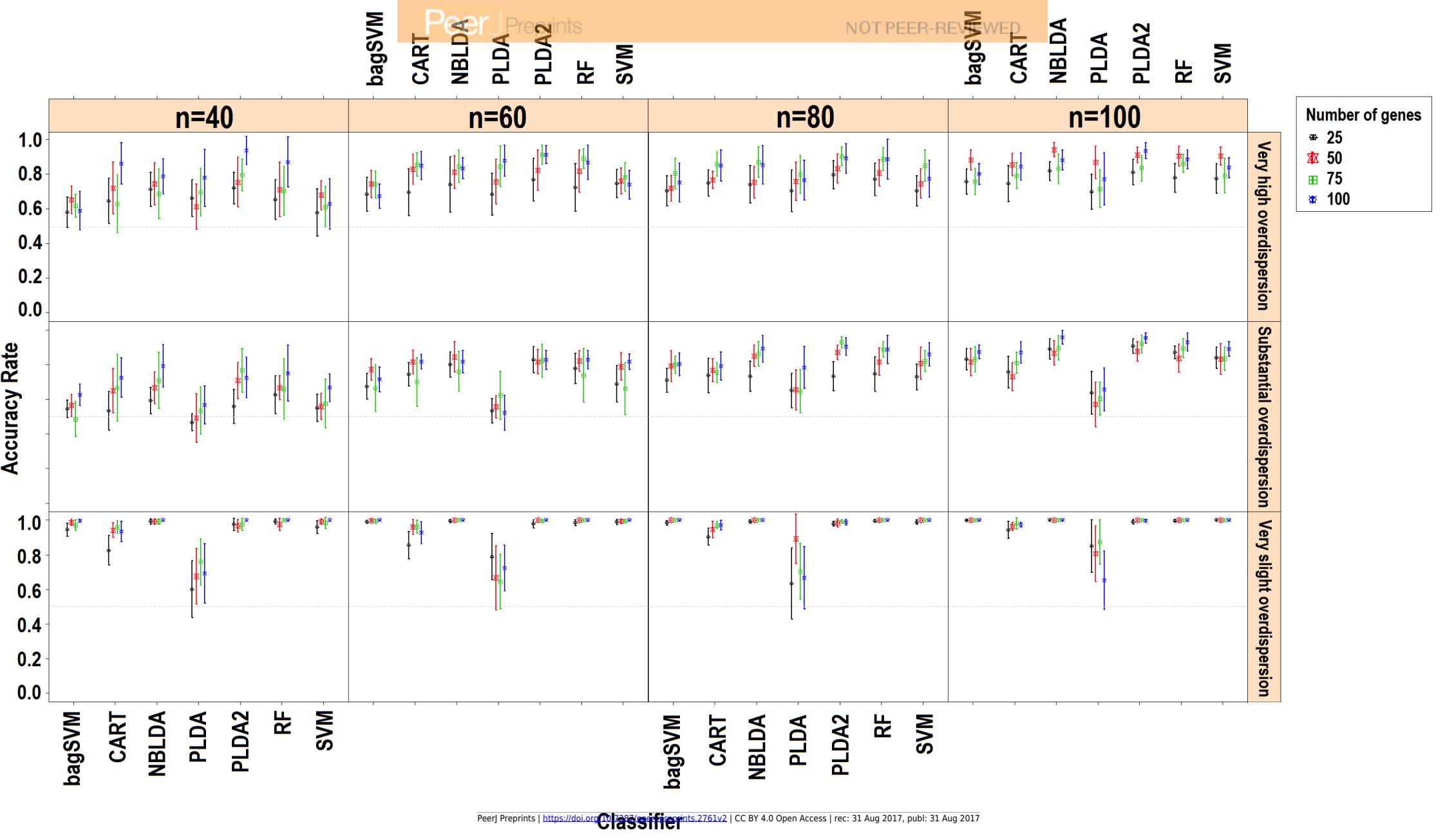


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2761v2 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017

Figure 4(on next page)

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 (φ=*0.01:* very sli

Fig 4 - 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 (φ=*0.01:* very slight, φ=*0.1:* substantial, φ=*1:* very high)



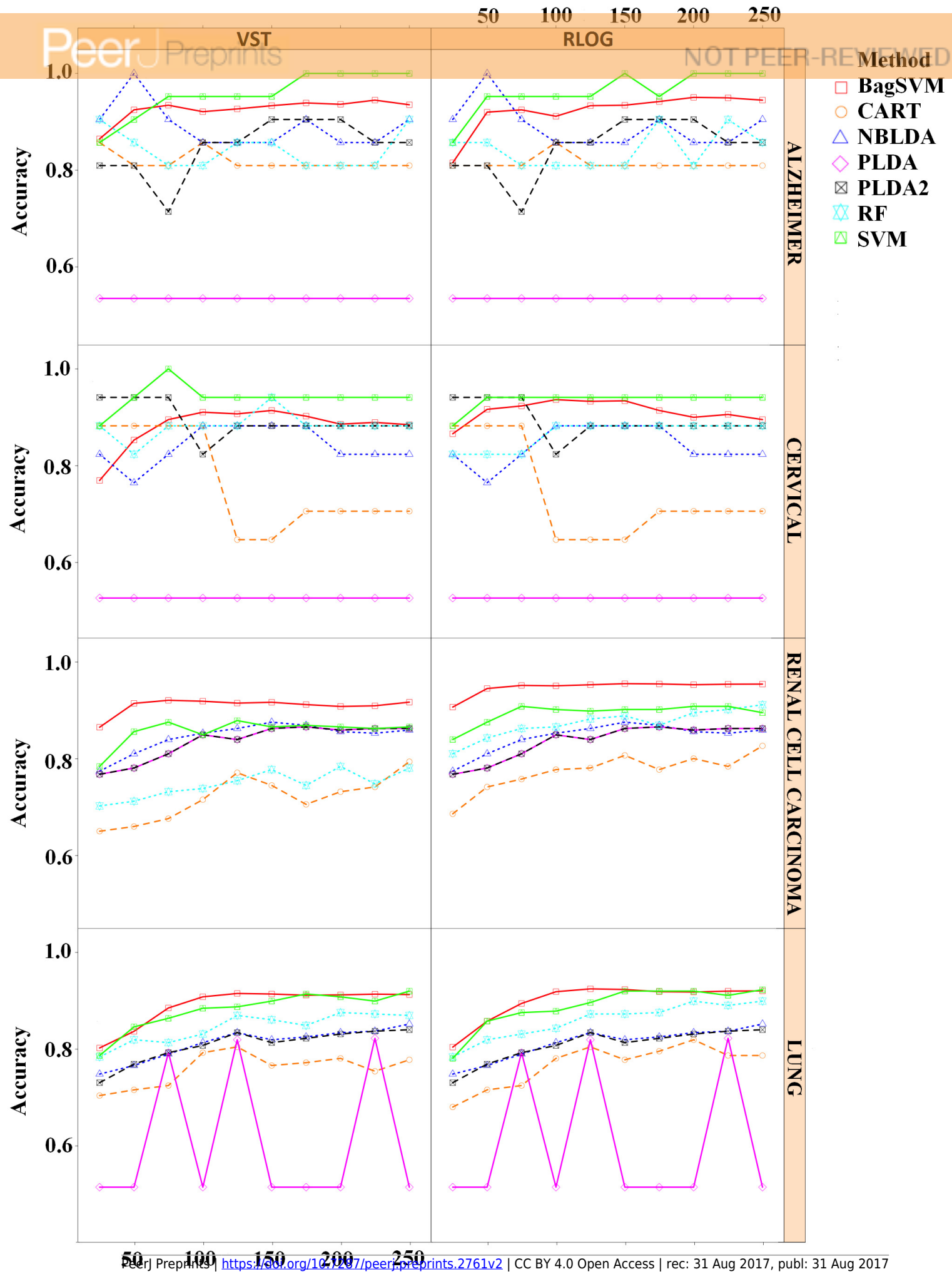
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2761v2 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017

Figure 5(on next page)

Results obtained from real datasets. Figure shows the performance results of classifiers for datasets with changing number of most significant number of genes

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2761v2 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2761v2 | CC BY 4.0 Open Access | rec: 31 Aug 2017, publ: 31 Aug 2017