RESEARCH ARTICLE

RNA-Seq Count Data Modelling by Grey Relational Analysis and Nonparametric Gaussian Process

Thanh Nguyen[1]*, **Asim Bhatti**[1], **Samuel Yang**[2], **Saeid Nahavandi**[1]

1 Institute for Intelligent Systems Research and Innovation, Deakin University, Victoria, Australia,
2 Department of Emergency Medicine, Stanford University, California, United States of America

* thanh.nguyen@deakin.edu.au

Abstract

This paper introduces an approach to classification of RNA-seq read counts using grey relational analysis (GRA) and Bayesian Gaussian process (GP) models. Read counts are transformed to microarray-like data to facilitate normal-based statistical methods. GRA is designed to select differentially expressed genes by integrating outcomes of five individual feature selection methods including two-sample t-test, entropy test, Bhattacharyya distance, Wilcoxon test and receiver operating characteristic curve. GRA performs as an aggregate filter method through combining advantages of the individual methods to produce significant feature subsets that are then fed into a nonparametric GP model for classification. The proposed approach is verified by using two benchmark real datasets and the five-fold cross-validation method. Experimental results show the performance dominance of the GRA-based feature selection method as well as GP classifier against their competing methods. Moreover, the results demonstrate that GRA-GP considerably dominates the sparse Poisson linear discriminant analysis classifiers, which were introduced specifically for read counts, on different number of features. The proposed approach therefore can be implemented effectively in real practice for read count data analysis, which is useful in many applications including understanding disease pathogenesis, diagnosis and treatment monitoring at the molecular level.

OPEN ACCESS

Citation: Nguyen T, Bhatti A, Yang S, Nahavandi S (2016) RNA-Seq Count Data Modelling by Grey Relational Analysis and Nonparametric Gaussian Process. PLoS ONE 11(10): e0164766. doi:10.1371/journal.pone.0164766

Editor: Xiaofeng Wang, Cleveland Clinic Lerner Research Institute, UNITED STATES

Received: May 15, 2016

Accepted: September 30, 2016

Published: October 26, 2016

Data Availability Statement: The Mont-Pick dataset is available from the ReCount RNA-seq database: http://bowtie-bio.sourceforge.net/recount/. The cervical cancer dataset is available from Gene Expression Omnibus: http://www.ncbi.nlm.nih.gov/geo/ under accession number GSE20592.

Funding: The author(s) received no specific funding for this work.

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

Introduction

Discovery of genes that are differentially expressed is helpful in gaining insights into disease pathogenesis, and discovering biomarkers for diagnosing and predicting the clinical status of patients. Identifying gene biomarkers is often performed using DNA microarray, which measures gene expression of the entire human genome. DNA microarray technology however suffers from the cross-hybridization procedure that yields noisy gene expression profiles. RNA sequencing (RNA-seq) has been emerging as a favorite method against the microarray technology [1]. RNA-seq is a technique that is capable of generating RNA-seq count data based on the next generation sequencing (NGS) technologies. The count data are structured as a table, which reports the number of sequence fragments assigned to each gene for each sample. RNA-seq is increasingly preferable to DNA microarray because it produces low background noise count data that allow detecting transcripts at low expression levels [2, 3]. With the decreasing cost of sequencing, the use of RNA-seq for differential expression analysis has been increased rapidly. NGS is able to measure the expression levels of tens of thousands of transcripts simultaneously. Such information is useful for developing expression-based classification algorithms to determine the diagnostic category of disease, for example cancers [4, 5].

Fig 1 shows basic steps of a typical RNA-seq experiment. Specifically, an RNA-seq experiment normally requires a task of making a collection of cDNA fragments that are flanked by sequencing adapters. This library of cDNA fragments is then sequenced using a short-read sequencing platform. This step results in millions of short sequence reads that correspond to individual cDNA fragments.

As the RNA-seq technology provides count data, much interest has focused on statistical methods designed specifically for discrete counts, for example approaches using Poisson and negative binomial (NB) distributions. Witten et al. [6] introduced a Poisson linear discriminant analysis for modelling RNA-seq data. Alternatively, a specific nonlinear Poisson transformation was proposed in [7] and applied to the mRNA expression model to synthetically generate the RNA-seq data. Likewise, several over-dispersed Poisson models were introduced in [8–10]. A comparison of methods and software packages for detecting differential expression in RNA-seq studies was presented in [11, 12].

Due to the overdispersion issue, i.e. the variances are likely to exceed the means for a considerable number of genes [13], the Poisson distribution may not be suitable for modelling RNA-seq profiles when there are biological replicates. The NB distribution is therefore more general because it can mitigate this issue [14].

Robinson and Smyth [15] presented a quantile-adjusted conditional maximum likelihood estimator for the dispersion parameter of the NB distribution accompanying by the R package edgeR, which was detailed in [16]. Anders and Huber [17] proposed a method along with the DESeq package using the NB distribution with variance and mean linked by local regression. Hardcastle and Kelly [18] developed the algorithm baySeq that uses an empirical Bayes approach to discover patterns of differential expression by assuming a NB distribution for the data. Likewise, Wu et al. [19] introduced a shrinkage estimate of the dispersion parameters of the NB model for RNA-seq data. This estimator characterizes the variation in gene-specific dispersion and provides a better detection of differential expression genes compared with edgeR and DESeq. Love et al. [20] presented DESeq2, a successor to the DESeq method, to facilitate a more quantitative analysis of comparative RNA-seq count data using shrinkage estimators for dispersion and fold change.

Modelling sequencing data using count distributions is mathematically intractable and complicated because of the presence of extreme values, high skewness and the mean-variance dependency. Therefore, an alternative approach has emerged by using transformation procedures for the count RNA-seq data and applying normal-based microarray-like statistical methods. This reduces the disadvantages relating to the mathematical intractability of count distributions compared to the normal distribution and opens access to a wide range of known algorithms developed for microarray data. Several prevalent methods include logarithm transformation [3], variance-stabilizing transformation (VST) [17], TMM transformation [21], regularized logarithm [20], and variance modelling at the observation level "voom" method [22]. voom was verified and demonstrated that it performs as well or better than existing RNA-seq methods. This paper therefore promotes the use of voom method to process the RNA-seq data.

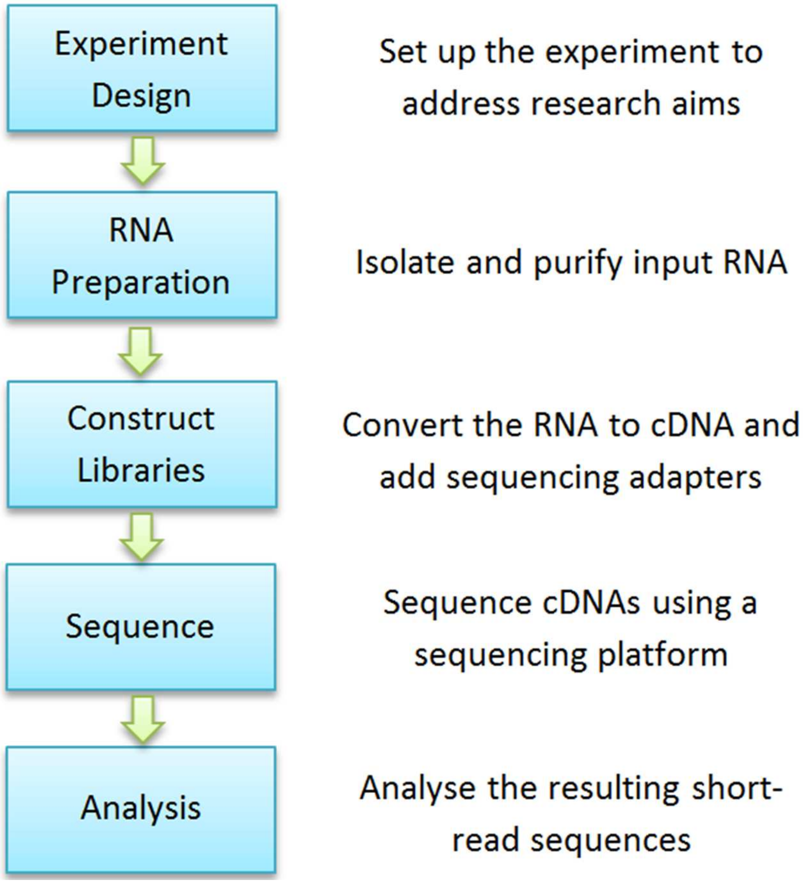
Experiment Design
Set up the experiment to address research aims
RNA Preparation
Isolate and purify input RNA
Construct Libraries
Convert the RNA to cDNA and add sequencing adapters
Sequence
Sequence cDNAs using a sequencing platform
Analysis
Analyse the resulting short-read sequences

Fig 1. Basic steps of a typical RNA-seq experiment.

doi:10.1371/journal.pone.0164766.g001

Using the voom transformation, we introduce an aggregate feature selection method based on the grey relational analysis (GRA) technique [23] to deal with transformed RNA-seq data. Compressed feature subsets obtained by the filter GRA method are fed into the Bayesian non-parametric Gaussian process (GP) models [24] for classification. Benchmark sequencing datasets are used to validate and show the significant dominance of the proposed approach against competing methods. We also perform rigorous statistical significance test to ensure the conclusions driven out of this study are valid and general. Next section presents in detail the proposed methodology and motivations of using GRA and GP methods.

Methods

The proposed methodology for analysis of RNA-seq read counts is graphically presented in Fig 2. One of the basic tasks in the analysis of RNA-seq count data is the detection of differentially expressed genes [25]. In this paper, the RNA-seq read counts are first transformed using the voom method [22]. The transform alleviates the typical skewness, dependency between mean and variance or extreme values of RNA-seq data. After the transformation, RNA-seq data can be treated as if it was microarray data. This means that any normal-based methods or gene set testing procedures can be applied to RNA-seq data. We then design the GRA-based aggregate feature selection method that combines outcomes of five individual methods including two-sample t-test, entropy test (known as Kullback-Liebler distance or divergence) [26],

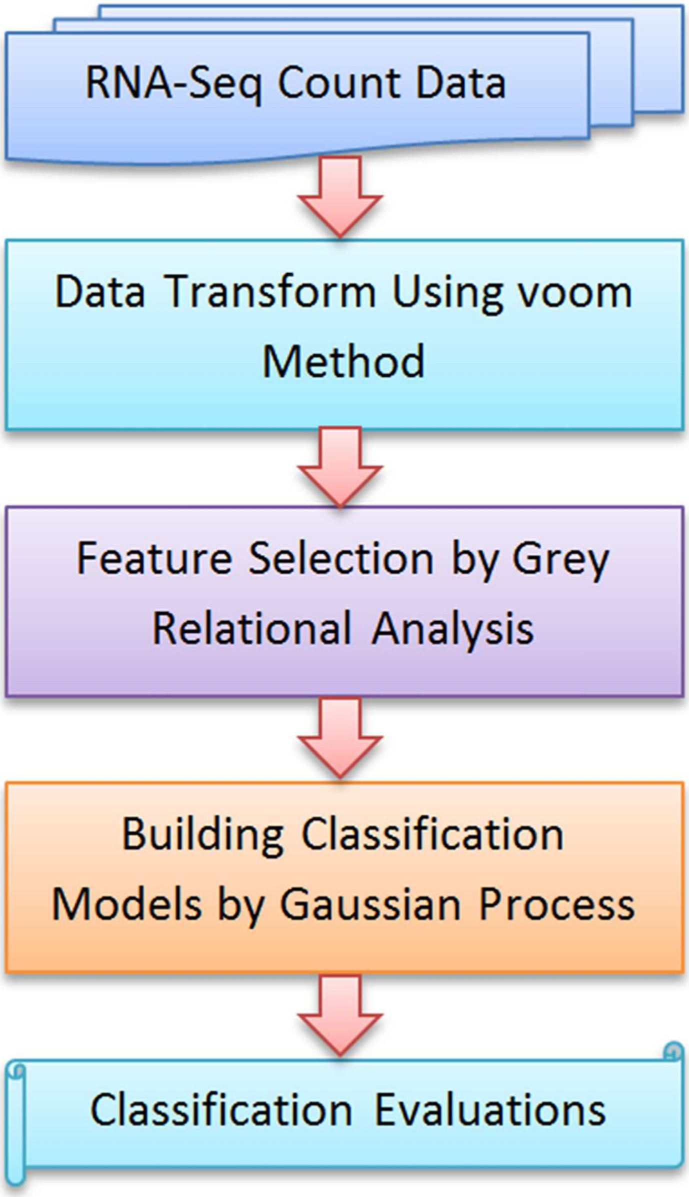
RNA-Seq Count Data
Data Transform Using voom Method
Feature Selection by Grey Relational Analysis
Building Classification Models by Gaussian Process
Classification Evaluations

Fig 2. Proposed methodology for analyzing RNA-seq count data.

doi:10.1371/journal.pone.0164766.g002

Bhattacharyya distance [27], Wilcoxon test [28] and receiver operating characteristic (ROC) curve [29] to select significant genes as biomarkers. GRA-based method performs as a filter approach based on the assumption that the features are independent. This assumption is often made for high-dimensional low-sample data as there are too few observations available to be able to effectively estimate the dependence structure among the features [6, 30–32].

Once discriminant feature subsets have been selected, they serve as inputs into the GP models for classification. GP is fast and computationally tractable based on analytic formulae. A GP is completely characterized by its mean and covariance functions but it is not limited by a parametric form. Being a nonparametric method, the number and nature of GP parameters are flexible and not fixed in advance but are determined from data. Therefore, uncertainty and complexity of RNA-seq data can be addressed effectively by GP models. Under the GP viewpoint, the models are transparent and hence amenable to interpretation compared to black-box methods such as neural networks [24]. Generalization capability of GP based on Bayesian formalism can yield high classification performance for RNA-seq data modelling. Details of the voom transform approach, GRA-based feature selection method and GP models are sequentially presented in the following subsections.

RNA-seq data transformation

Raw RNA-seq data are assembled in integer read counts. Specific characteristics of RNA-seq data that concern analysts are the presence of extreme values, high skewness, and the mean-variance dependency (i.e. heteroscedasticity). Logarithm transformation is a prevalent method to eliminate RNA-seq extreme values [3]. The variance-stabilizing transformation (VST) proposed in [17] is also often used to deal with skewed RNA-seq data. Alternatively, Love et al. [20] introduced regularized logarithm to transform RNA-seq data to render them homoscedastic. Law et al. [22] proposed voom method that converts the counts to log-counts per million with associated precision weights. After this, the normal-based methods can be applied to RNA-seq data as if it was microarray data. Details of the voom method are presented in the Supplementary Materials section.

GRA-based feature selection

GRA was introduced by Deng [33] and has been applied to solve multicriteria decision making (MCDM) problems in various fields [23, 34, 35]. GRA is part of grey system theory, which is capable of solving problems with complicated interrelationships between multiple factors and variables. We propose the use GRA as a filter feature selection approach that combines outcomes of individual methods including two-sample t-test, entropy test, Bhattacharyya distance, Wilcoxon test and ROC curve. Assume the MCDM problem has m alternatives and n criteria (attributes) where the ith alternative can be expressed as $Y_i = (y_{i1}, y_{i2}, \ldots, y_{ij}, \ldots, y_{in})$ where y_{ij} is the performance value of the criterion j of the alternative i. To formulate gene selection as an MCDM problem, we treat genes (features) as alternatives and individual methods as criteria. Therefore, there are m features corresponding to m alternatives. In this paper, n is equal to 5 as there are 5 individual methods corresponding to 5 criteria. Outcomes of individual methods are scores of every feature. For each individual method, we represent its scores as the performance values of corresponding features. The following presents steps of the GRA algorithm.

(1) Grey relational generating: This step is to translate the performance values of all alternatives into a comparability sequence. It normalizes data sequence for the experimental results within 0 and 1. If the larger target value of the original sequence is the better, then the normalization is performed by:

$$x_{ij} = \frac{y_{ij} - \min_i y_{ij}}{\max_i y_{ij} - \min_i y_{ij}} \tag{1}$$

Alternatively, if the smaller target value is the better then the original sequence is normalized by:

$$x_{ij} = \frac{\max_i y_{ij} - y_{ij}}{\max_i y_{ij} - \min_i y_{ij}} \tag{2}$$

where x_{ij} is the generating value of the grey relational analysis, $\min_i y_{ij}$ is the minimum value of y_{ij} among all alternatives $i = 1,2,\ldots,m$ and $\max_i y_{ij}$ is the maximum value of y_{ij}.

(2) Define the reference sequence: Once the grey relational generating procedure is complete, all performance values are scaled into [0, 1]. An alternative will be the best choice if all of its performance values are equal to or close to 1. Therefore, we define the reference sequence X_0 as $(x_{01},x_{02},\ldots,x_{0j},\ldots,x_{0n}) = (1,1,\ldots,1,\ldots,1)$ and then find the alternative whose comparability sequence is the closest to the reference sequence X_0.

(3) Calculate the grey relational coefficient: This coefficient is used to determine how close x_{ij} to x_{0j}. The larger the coefficient is the closer between x_{ij} and x_{0j}. This coefficient can be computed by:

$$\delta(x_{0j}, x_{ij}) = \frac{\Delta_{\min} + \alpha\Delta_{\max}}{\Delta_{ij} + \alpha\Delta_{\max}} \tag{3}$$

where $\Delta_{ij} = |x_{0j} - x_{ij}|$, $\Delta_{\min} = \min_{i,j}\Delta_{ij}$, $\Delta_{\max} = \max_{i,j}\Delta_{ij}$ and α is the distinguishing coefficient, $\alpha \in [0, 1]$. The distinguishing coefficient may expand or compress the range of the grey relational coefficient. In this paper, we set $\alpha = 0.5$ for all experiments.

(4) Calculate the grey relational grade between X_i and X_0 using:

$$\phi(X_0, X_i) = \sum_{j=1}^{n} w_j \delta(x_{0j}, x_{ij}) \tag{4}$$

where w_j is the weight of attribute j and $\sum_{j=1}^{n} w_j = 1$. The above equation is applied to all m alternatives $i = 1,2,\ldots,m$. The grey relational grade represents the degree of similarity between the comparability sequence and the reference sequence. Therefore, if a comparability sequence for an alternative achieves the greatest grey relational grade with the reference sequence, that alternative is the best choice.

For the purpose of feature selection for classification, we rank alternatives (features) based on their corresponding grey relational grades. Features have the top grey relational grades are selected to form a feature set.

The next subsections scrutinize background of individual feature selection filter methods whose outcomes are used for the proposed GRA approach. These methods are accomplished by ranking features via scoring metrics. They are statistic tests based on two sets of data samples in the binary classification problem. The sample means are denoted as μ_1 and μ_2, whereas σ_1 and σ_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes [36].

Two-sample t-test. The two-sample t-test is a parametric hypothesis test that is applied to compare whether the average difference between two independent sets of data samples is really significant. The test statistic is calculated by:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{5}$$

In the application of t-test for gene selection, the test is performed on each gene by separating the expression levels based on the class variable. The absolute value of t is used to evaluate the significance among genes. The higher the absolute value, the more important is the gene.

Entropy test. Relative entropy, also known as Kullback-Liebler distance or divergence is a test assuming classes are normally distributed. The entropy score for each gene is computed using the following expression:

$$e = \frac{1}{2}\left[\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} - 2\right) + \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)(\mu_1 - \mu_2)^2\right] \tag{6}$$

After the computation is complete for every gene, genes with the greatest entropy scores are selected to serve as inputs to the classification techniques.

Bhattacharyya distance. The Bhattacharyya distance can be calculated from the standard deviation and mean of each class as follows:

$$BD = \frac{1}{4}\ln\left[\frac{1}{4}\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} + 2\right)\right] + \frac{1}{4}\left[\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}\right] \tag{7}$$

Wilcoxon method. The Wilcoxon rank sum test [28] is a test for equality of population locations (medians). The null hypothesis is that two populations enclose identical distribution functions whereas the alternative hypothesis states that two distributions differ regarding the medians. The normality assumption regarding the differences between the two samples is not required. That is why this test is used instead of the two-sample t-test in many applications when the normality assumption is concerned. The steps of the Wilcoxon test are summarized below [29]:

1. Assemble all observations of the two populations and rank them in the ascending order.
2. The Wilcoxon statistic is calculated by the sum of all the ranks associated with the observations from the smaller group.
3. The hypothesis decision is made based on the p-value, which is found from the Wilcoxon rank sum distribution table.

In the applications of the Wilcoxon test for gene selection, the absolute values of the standardized Wilcoxon statistics are utilized to rank genes.

Receiver operating characteristic curve. Denote the distribution functions of X in the two populations as $F_1(x)$ and $F_2(x)$. The tail functions are specified respectively $T_i(x) = 1 - F_i(x)$, $i = 1, 2$. The ROC is given as follows:

$$ROC(t) = T_1\left[T_2^{-1}(t)\right], t \in (0, 1) \tag{8}$$

and the area under the curve (AUC) is computed by:

$$AUC = \int_0^1 ROC(t)dt \tag{9}$$

The larger the AUC, the less is the overlap of the classes. Genes with the greatest AUC therefore are chosen to form a gene set.

Gaussian process models

A nonparametric GP is a generalization of the Gaussian probability distribution based on a Bayesian methodology. GP is defined as a collection of random variables, any finite number of which have a joint Gaussian distribution. GP can be used for function approximation problems including both classification and regression. In the regression problems, likelihood function is often assumed to be Gaussian, which combines with a GP prior to yield a posterior GP over

Table 1. Summary of RNA-seq datasets.

Datasets	Features	Samples	Classes
Mont-Pick [42]	12,984 genes	129	CEU/YRI
Cervical cancer [44]	714 microRNAs	58	tumor/non-tumor

doi:10.1371/journal.pone.0164766.t001

functions. This exact Bayesian inference manipulation is analytically tractable. In classification problems, as the targets are discrete class labels, the Gaussian likelihood is therefore inappropriate. Therefore, an approximate inference is needed for classification problems. Several methods have been proposed that include Laplace's method, Expectation Propagation (EP), variational approximations and Markov chain Monte Carlo (MCMC) modelling, e.g. see [37, 38].

The GP method for binary classification in this study is implemented by customizing the Gaussian processes for machine learning toolbox developed by Rasmussen and Nickisch [39]. Details of GP and its parameter settings are presented in the Supplementary Materials section.

Experimental RNA-seq datasets

Two benchmark real datasets are utilized in this study for comparisons. We name the first dataset as “Mont-Pick” as it is obtained from a combination of two studies Montgomery et al. [40] and Pickrell et al. [41]. This data set is available through the ReCount RNA-seq database developed by Frazee et al. [42]. The data can be used to analyze differential expression between two ethnicities: the Montgomery group sequenced Utah people with ancestry from northern and western Europe (the HapMap CEU population) and the Pickrell group sequenced Yoruba residents in Ibadan, Nigeria (the HapMap YRI population). These two groups of ethnicities are treated as two separate classes in this study. There are 60 samples from the CEU group and 69 samples from the YRI group. A total of 52,580 genes are processed by which genes with zero counts in all samples are removed. The number of nonzero count genes of the Mont-Pick dataset is 12,984.

The second data set “cervical cancer” is available from Gene Expression Omnibus [43] under accession number GSE20592, which was utilized in [6, 44]. The data include 29 tumor and 29 non-tumor cervical tissue samples measured on 714 microRNAs, which are small RNAs with 18–30 nucleotides in length. The classification task is to distinguish between tumor and non-tumor samples. Details of the experimental datasets are presented in Table 1.

In the Mont-Pick dataset, the CEU and YRI samples were sequenced by different groups using potentially different facilities. Therefore, the batch effect would be a factor that affects the performance comparisons of RNA-seq data analysis approaches. To deal with this issue, we have included the design option that addresses the batch effect in the voom transformation method, which is implemented in the limma package [45]. Figs 3 and 4 show pseudocolor heat maps of the expression levels before and after voom transformation for the Mont-Pick and cervical cancer datasets respectively. The x-axis represents genes or genomic features of interest whilst the y-axis represents data samples of different groups (classes). Both datasets used in this study have two classes of samples and the horizontal red line in every heat map divides the samples into the two classes. In each heat map, the corresponding color bar representing expression levels is plotted adjacent to the color map. The white color represents mid points, warm colors represent high expression levels and cool colors indicate low expression or sparse regions.

Before voom transform, the white region locates at the bottom of color bars in both datasets (see Figs 3a and 4a). This shows that read counts follow a positively skewed distribution, which

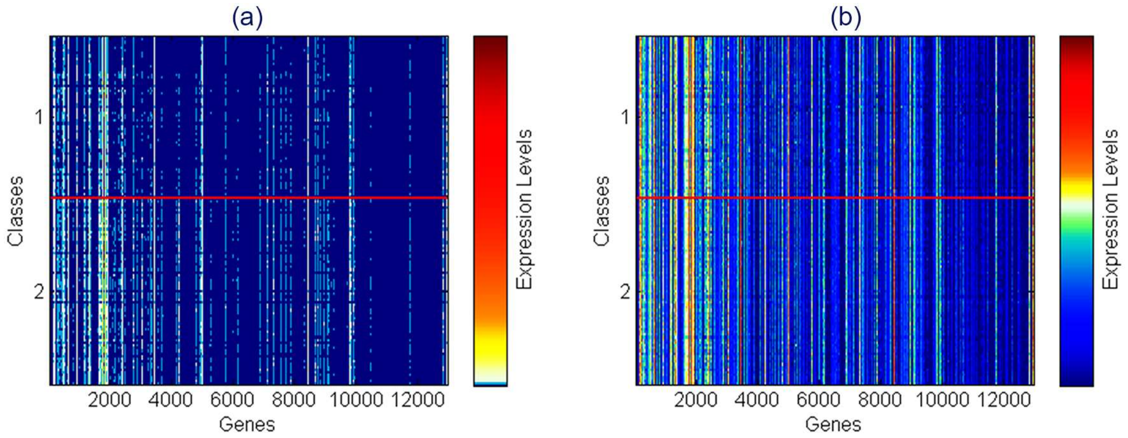

Fig 3. Heat maps showing expression levels of the Mont-Pick dataset (a) before and (b) after voom transform.

doi:10.1371/journal.pone.0164766.g003

would hinder the application of normal-based methods. Moreover, color maps of data before voom are almost blue with a very small proportion of warm color spots, which represent extreme values or outliers. The range of expression levels before voom transformation is extremely large, from 0 to 91,991 in the Mont-Pick dataset and from 0 to 476,438 in the cervical cancer dataset. The large blue areas in color maps represent sparse data, especially in Fig 4a. In contrast, the data after voom transformation is continuously distributed with the white region locates near the middle of color bars. In addition, the data after voom transform are less sparse than the original count reads as heat maps are more colorfully diversified with the combination of cool and warm colors. This demonstrates that the transformed data practically follow a normal distribution and can be processed by normal-based statistical methods.

Performance evaluation metrics

To highlight the advantages of GRA-based feature selection method, we implement a number of competing methods for comparisons including ReliefF [46], iterative search margin based algorithm (Simba) [47], signal-to-noise ratio (SNR) [48], and information gain (IG) [49].

The following methods are also applied for comparisons with the designed GP classifier: k-nearest neighbors (kNN) [50], multilayer perceptron (MLP) [51], support vector machine

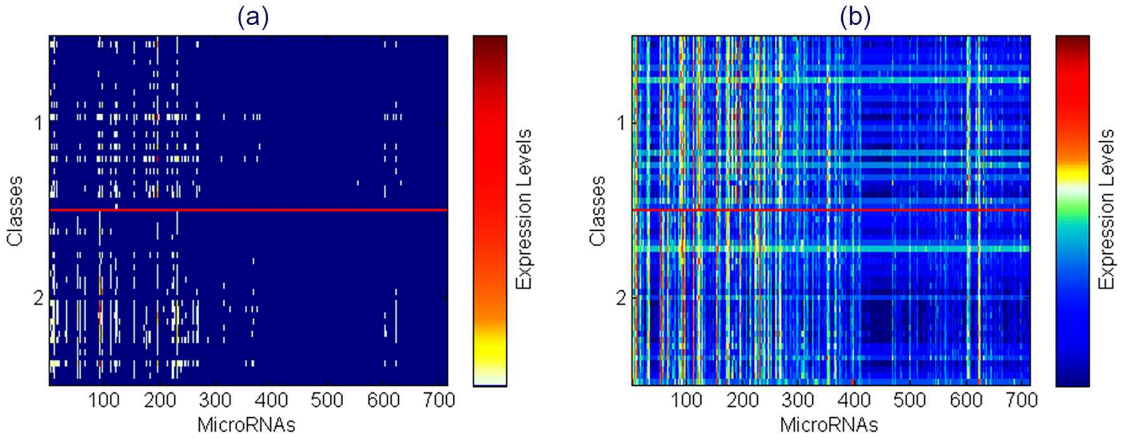

Fig 4. Heat maps of the expression levels in the cervical cancer dataset (a) before and (b) after voom transform.

doi:10.1371/journal.pone.0164766.g004

(SVM) [52] and ensemble learning AdaBoost [53]. Specifically, the number of nearest neighbors in kNN is equal to 5 and SVM kernel function is the Gaussian radial basis function with the scaling factor of 1. MLP is constructed with two hidden layers and each layer comprises five nodes. AdaBoost uses a collection of individual learners that are 100 decision trees.

Four different performance metrics including accuracy rate, F1 score statistics (F-measure), AUC and mutual information (MI) are used to evaluate performance of classification methods. F-measure considers both the "Precision" (denoted as *Pr*) and "Recall" (*Re*) of the classification procedure to compute the score expressed by:

$$F - measure = 2 \times \frac{\mathrm{Pr} \times \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}} \tag{10}$$

The MI between estimated and true label is calculated by:

$$MI(\widehat{C}, C) = \sum_{\widehat{c}=0}^{1} \sum_{c=0}^{1} p(\widehat{c}, c) \log \frac{p(\widehat{c}, c)}{p(\widehat{c})p(c)} \tag{11}$$

where $p(\widehat{c}, c)$ is the joint probability distribution function of estimated and true class labels $\widehat{C}$ and C, and $p(\widehat{c})$ and $p(c)$ are the marginal probability distribution functions of $\widehat{C}$ and C respectively.

The five-fold cross validation method is employed for experiments. Data samples are divided at random into five distinct subsets and four subsets are used for training classifiers whilst the last subset is for testing. For unbiased comparisons, each classifier is repeated 30 times and the average performance is reported along with the standard error.

To draw convincing conclusions in evaluating performance of feature selection methods and classifiers, we implement the Mann-Whitney U-test [54] for comparing two sets of results. The Mann-Whitney U-test is a nonparametric test of the null hypothesis that two populations have distributions with equal medians, against the alternative hypothesis that they do not. As the results over 30 trials may not be normally distributed, the use of Mann-Whitney U-test is more appropriate than that of normal-based methods [55].

Note that the test is performed to compare between the set of 30 outcomes generated by GRA method and that obtained by each of the competing feature selection methods using the same classifier. Similar procedure is performed to compare the GP classifier with its competing methods, i.e. kNN, MLP, SVM and AdaBoost using the same feature selection method.

Results and Discussions

After voom transformation, GRA-based gene selection is employed for RNA-seq data to select genes that are differentially expressed for classification. GRA-based method performs as a filter method that ranks genes by combining outcomes of individual methods t-test, entropy, Bhattacharyya distance, Wilcoxon, and ROC. It therefore obtains the quintessence of these individual methods and provides stable and most discriminant subsets of genes. Fig 5 shows 3D presentations of feature subsets obtained by GRA-based method using the Mont-Pick and cervical cancer datasets. Obviously, GRA method is able to provide a clear separation between samples of two classes in both datasets. This facilitates the great classification performance of classifiers that use GRA-based feature subsets.

Comparisons of GRA-based method with ReliefF, Simba, SNR, and IG

Feature subsets of top ten genes selected by GRA-based method serve as inputs into classifiers for demonstration although different number of genes can be used. For comparison, the same

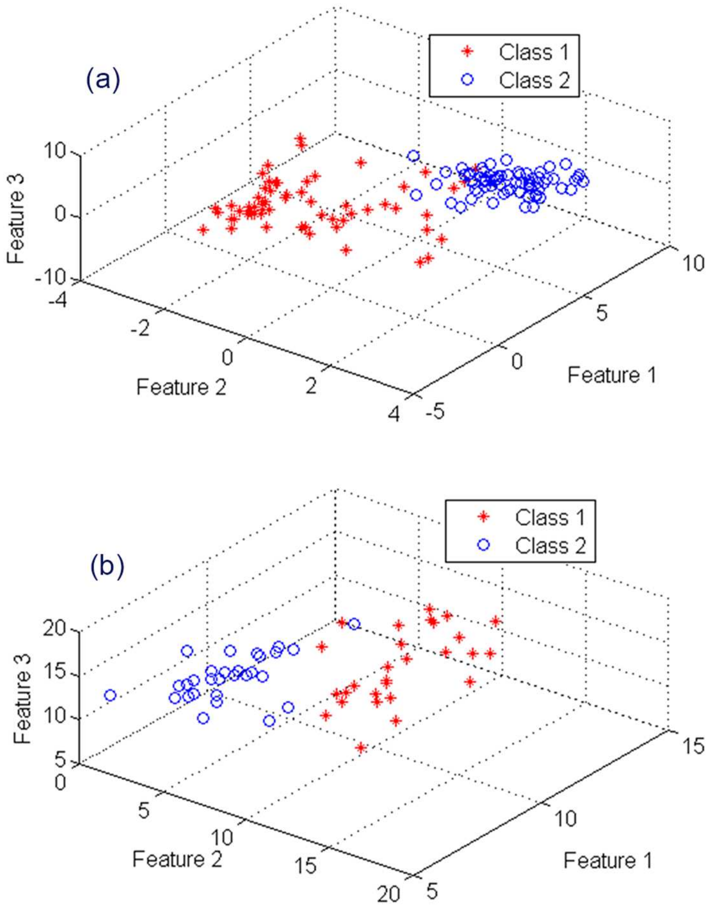

Fig 5. Distribution of data samples of the (a) Mont-Pick dataset and (b) cervical cancer dataset.

doi:10.1371/journal.pone.0164766.g005

number of genes is selected by other methods in order to form feature subsets. Tables 2 and 3 present classification results of different feature selection methods for the Mont-Pick and cervical cancer datasets respectively. The classification is performed by the GP method and results for the accuracy, F-measure, AUC and MI metrics are reported in percentage.

Each cell in these tables represents the mean and standard error of 30 classification outcomes. The value in brackets shows the *p*-value of the statistical Mann-Whitney U-test between each of the competing methods and the GRA method. For example, the value of 0.003 in the cell Accuracy-ReliefF in Table 2 is the *p*-value of the Mann-Whitney U-test between two sets of accuracy outcomes: one set is generated by using ReliefF and the other is obtained by GRA. The *p*-value smaller than 0.05 (the 5% significance level) indicates that the difference between two sets are statistically significant. In other words, the GRA method is significantly better than the ReliefF method. Values in italic in Tables 2 and 3 are *p*-values that are greater than 0.05.

Table 2. Results of feature selection methods using the Mont-Pick dataset (batch effect is addressed due to potentially different facilities).

Metrics	ReliefF	Simba	SNR	IG	GRA
Accuracy	93.78±0.80 (0.003)	84.81±1.26 (0.000)	96.05±0.49 (*0.154*)	90.80±1.22 (0.000)	96.77±0.71
F-measure	95.44±0.82 (0.022)	86.24±1.27 (0.000)	94.87±0.66 (0.003)	91.32±1.54 (0.001)	97.64±0.48
AUC	95.16±0.69 (0.005)	86.76±1.13 (0.000)	95.39±0.65 (0.031)	90.92±1.21 (0.000)	97.43±0.51
MI	75.89±3.15 (0.012)	42.99±3.60 (0.000)	78.88±2.46 (0.018)	72.68±4.73 (0.008)	88.56±2.71

doi:10.1371/journal.pone.0164766.t002

Table 3. Results of feature selection methods using the cervical cancer dataset.

Metrics	ReliefF	Simba	SNR	IG	GRA
Accuracy	88.33±1.54 (0.013)	87.35±2.02 (0.029)	88.23±1.65 (0.031)	90.05±1.80 (*0.199*)	93.43±1.28
F-measure	87.92±1.67 (0.023)	84.95±2.74 (0.042)	87.67±1.70 (0.020)	90.77±1.58 (*0.280*)	92.91±1.57
AUC	88.23±1.47 (0.006)	87.61±1.98 (0.022)	89.21±1.58 (0.043)	91.89±1.38 (*0.316*)	94.07±1.22
MI	58.48±4.60 (0.024)	57.07±6.08 (0.044)	57.83±4.81 (*0.085*)	66.16±5.10 (*0.361*)	73.96±4.85

doi:10.1371/journal.pone.0164766.t003

In the Mont-Pick dataset, GRA shows a great performance compared with its competing methods. Specifically, GRA's accuracy is of 96.77%, which is higher than ReliefF, Simba, SNR and IG of 93.78%, 84.81%, 96.05% and 90.80% respectively. Similar finding is seen in the cervical cancer dataset where GRA method outperforms other feature selection methods with regard to the accuracy metric. GRA obtains 93.43% of accuracy whilst those of ReliefF, Simba, SNR and IG are 88.33%, 87.35%, 88.23% and 90.05% respectively. Via other performance metrics, i.e. F-measure, AUC and MI, GRA feature selection method also demonstrates a considerable superiority to its competing methods. For example, in the Mont-Pick dataset, MI of GRA is of 88.56%, which is much greater than that of ReliefF at 75.89%, Simba at 42.99%, SNR at 78.88% and IG at 72.68%. Likewise, in the cervical cancer dataset, GRA's MI is of 73.96%, which is the highest performance among the examined feature selection methods.

The GRA method obtains not only greater performance but also more stable results compared with its competing methods. This is demonstrated via standard errors of the results. In the Mont-Pick dataset, GRA results' standard errors are mostly smaller than those of ReliefF, Simba, SNR and IG. In the cervical cancer dataset, GRA also obtains smaller standard errors than those of its competing methods except only one case MI-ReliefF.

Results of the Mann-Whitney U-test for evaluating feature selection methods demonstrate the statistical significance of GRA against its competing methods. In the Mont-Pick dataset, *p*-values of the Mann-Whitney U-test are smaller than 0.05 except only one case Accuracy-SNR. Therefore, the Mann-Whitney U-test rejects the null hypothesis that results of two methods (GRA and each of the competing feature selection methods) come from the same distribution at the 5% significance level. This means that GRA is significantly better than its competing methods in terms of all performance metrics. In the cervical cancer dataset, most *p*-values of the Mann-Whitney U-test are smaller than 0.05, except the comparisons of GRA with the IG method (for all performance metrics) and the SNR method (for MI metric).

Comparisons of GP classifier with kNN, MLP, SVM, and AdaBoost

We use the GRA-based feature selection method to obtain subsets of top ten features that are fed into every classifier for comparisons. Results of all classifiers are presented in Tables 4 and 5 for the Mont-Pick and cervical cancer datasets respectively.

Clearly, the GP classifier achieves greater performance compared with its competing methods in both datasets. The difference between GP with kNN, MLP, SVM and AdaBoost in the cervical cancer dataset is more considerable than that in the Mont-Pick dataset. The gap

Table 4. Comparisons of classifiers using the Mont-Pick dataset (batch effect is addressed due to potentially different facilities).

Metrics	kNN	MLP	SVM	AdaBoost	GP
Accuracy	95.16±0.57 (0.019)	95.15±0.70 (0.038)	95.01±0.89 (*0.133*)	94.50±0.72 (0.026)	96.77±0.71
F-measure	96.09±0.52 (0.019)	94.19±0.87 (0.001)	93.97±0.94 (0.002)	94.99±0.66 (0.009)	97.64±0.48
AUC	96.78±0.52 (*0.473*)	94.93±0.55 (0.000)	94.50±0.96 (0.012)	95.30±0.64 (0.014)	97.43±0.51
MI	76.46±2.56 (0.007)	80.06±2.53 (0.018)	81.60±3.07 (*0.212*)	76.62±3.24 (0.028)	88.56±2.71

doi:10.1371/journal.pone.0164766.t004

Table 5. Comparisons of classifiers using the cervical cancer dataset.

Metrics	kNN	MLP	SVM	AdaBoost	GP
Accuracy	88.11±1.53 (0.012)	85.33±2.00 (0.002)	87.42±2.21 (0.045)	88.38±1.32 (0.017)	93.43±1.28
F-measure	87.63±1.67 (0.016)	86.18±1.93 (0.011)	83.29±3.40 (0.019)	87.60±1.34 (0.005)	92.91±1.57
AUC	88.67±1.40 (0.007)	87.22±1.77 (0.006)	88.84±2.19 (*0.065*)	88.90±1.34 (0.010)	94.07±1.22
MI	57.49±4.53 (0.028)	51.53±5.24 (0.003)	60.29±5.43 (*0.126*)	57.47±4.08 (0.030)	73.96±4.85

doi:10.1371/journal.pone.0164766.t005

between GP and its competing methods is more than 5% in the cervical cancer dataset. More considerably, GP's MI is greater than those of kNN, MLP, SVM and AdaBoost by more than 13%. MI of kNN, MLP, SVM and AdaBoost are respectively of 57.49%, 51.53%, 60.29% and 57.47%, which are much lower than 73.96% of the GP classifier.

With regard to the Mann-Whitney U-test results, p-values are almost smaller than 0.05, except three cases, AUC-kNN, Accuracy-SVM and MI-SVM, in the Mont-Pick dataset and two cases, AUC-SVM and MI-SVM, in the cervical cancer dataset.

Comparisons of the proposed approach with sPLDA classifiers

In this subsection, we compare our approach, GP classifier using GRA-based feature subsets, with sparse Poisson linear discriminant analysis (sPLDA) classifiers, which were proposed by Witten [6]. The classification rule assigns the test observation x^* to the class for which the following expression is largest:

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

where c and c' are constants that are not dependent on the class label, whilst $\widehat{\pi}_k$ is the estimate of the prior probability that an observation belongs to class k. We set $\widehat{\pi}_1 = \ldots = \widehat{\pi}_K = 1/K$, corresponding to the prior that all classes are equally likely. Alternatively, $\widehat{f}_k$ is an estimate of the density of an observation in class k. A Poisson model for RNA sequencing data states that $X_j^*|y^* = k \sim \text{Poisson}(s^* g_i d_{kj})$ where $s^* = s_1, \ldots, s_n$ are the size factors for the training data, which can be estimated using the total count, median ratio and quantile as follows:

Total count: $\widehat{s}^* = \sum_{j=1}^{p} X_j^*/X..$ where $X..$ is the total number of counts of the training data.

Median ratio: $\widehat{s}^* = m^*/\sum_{i=1}^{n} m_i$ where $m^* = \text{median}_j \left\{ \frac{X_j^*}{(\Pi_{i=1}^{n} X_{ij})^{1/n}} \right\}$ and $m_i = \text{median}_j \left\{ \frac{X_{ij}}{(\Pi_{i'=1}^{n} X_{i'j})^{1/n}} \right\}$.

Quantile: $\widehat{s}^* = q^*/\sum_{i=1}^{n} q_i$ where q^* is the 75th percentile of counts for the test observation, and q_i is the 75th percentile of counts for the ith training observation.

The estimate of g_i is given by $\widehat{g}_j = X_{\cdot j}$ where $X_{\cdot j} = \sum_{i=1}^{n} X_{ij}$ and estimate of d_{kj} for sparse features is provided by:

$$\widehat{d}_{kj} = \begin{cases} \frac{a}{b} - \frac{\rho}{\sqrt{b}}, & \text{if } \sqrt{b}\left(\frac{a}{b} - 1\right) > \rho, \\ \frac{a}{b} + \frac{\rho}{\sqrt{b}}, & \text{if } \sqrt{b}\left(1 - \frac{a}{b}\right) > \rho, \\ 1, & \text{if } \sqrt{b}\left|1 - \frac{a}{b}\right| < \rho, \end{cases} \quad (13)$$

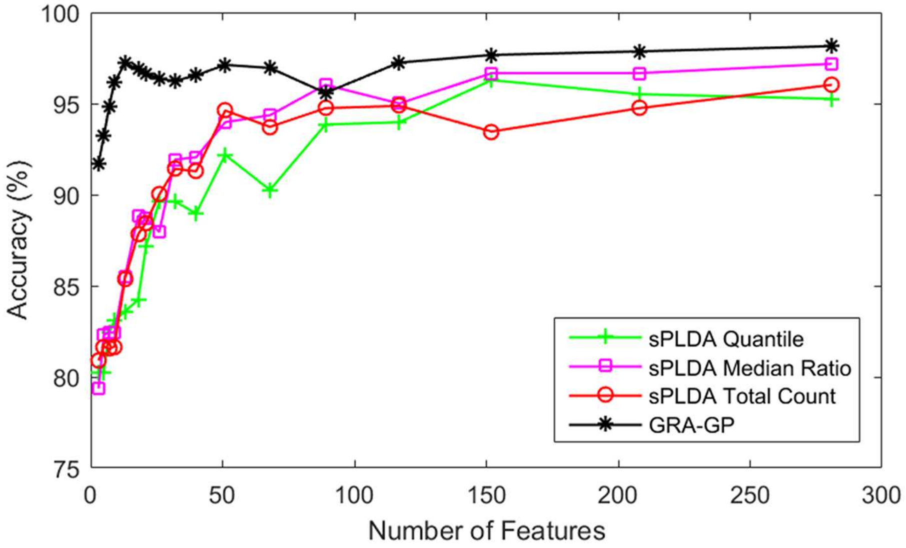

Fig 6. Comparisons of GRA-GP method with sPLDA classifiers using the Mont-Pick dataset.

doi:10.1371/journal.pone.0164766.g006

where $a = X_{C_{kj}} + \beta$, $b = \sum_{i \in C_k} \widehat{s}_i \widehat{g}_j + \beta$, and ρ is a nonnegative tuning parameter that is chosen by cross-validation. The number of features involved in classification is different when ρ is different. For unbiased comparisons, the number of features selected by our approach GRA-GP is the same with those determined by sPLDA. There are three approaches of sPLDA based on three corresponding methods of estimating the size factors for the training data, i.e. total count, median ratio and quantile. The comparisons are graphically depicted in Figs 6 and 7 by using the Mont-Pick and cervical cancer datasets respectively.

Results presented in these figures are obtained using the five-fold cross validation for each of the competing methods. We limit the number of features to approximately 300 and the performance is measured by accuracy in percentage. For each of the specified number of features, each classifier is repeated 30 times and the average result is reported. It is clear that GRA-GP method significantly dominates all three methods of sPLDA in both datasets based on different

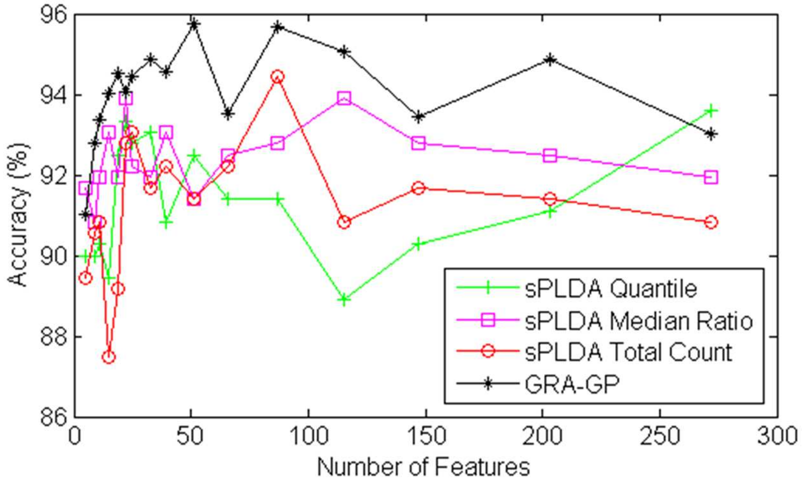

Fig 7. Comparisons of GRA-GP method with sPLDA classifiers using the cervical cancer dataset.

doi:10.1371/journal.pone.0164766.g007

number of features. In the Mont-Pick dataset, the gaps between GRA-GP method with its competing methods are very large when the number of features are smaller than 100. GRA-GP is still constantly superior to sPLDA classifiers when the number of feature increases. In the cervical cancer dataset, there are small gaps between GRA-GP and its competing methods when the number of features are smaller than 25. These gaps are larger when the number of features increases. sPLDA median ratio relatively ranks as the second best method after the GRA-GP. This highlights the effectiveness of our approach against the sPLDA classifiers.

Conclusions and Future Work

This paper proposes a new approach to RNA-seq count data classification using GRA-based feature selection method and the nonparametric GP models. RNA-seq data are assembled in integer read counts that present extreme values, high skewness, and heteroscedasticity. The voom transformation applied to RNA-seq data has turned them into microarray-like data by which a range of normal-based statistical methods can be utilized. On one hand, GRA systematically combines outcomes of individual methods, i.e. two-sample t-test, entropy test, Bhattacharyya distance, Wilcoxon test and ROC, and provides stable and robust feature subsets. By incorporating advantages and quintessence of the individual methods, GRA has shown a clear superiority to its competing methods that include ReliefF, Simba, signal to noise ratio, and information gain.

On the other hand, the nonparametric GP models based on the Bayesian inference methodology have addressed effectively the complexity of RNA-seq data. GP has demonstrated a considerable dominance in RNA-seq data classification against its competing methods including kNN, MLP, SVM and ensemble learning AdaBoost. Through analytic formulae, GP models are computationally tractable and easier to handle and interpret than their conventional counterparts such as neural networks. Via the characterization of mean and covariance functions, GP model fitting requires only the first- and second-order moments of the process to be specified. GP therefore has the generalization capability that has increased the classification performance. More considerably, the proposed GRA-GP approach has produced greater classification performance on different numbers of features against the sPLDA classifiers, which were proposed particularly for read counts modelling.

The use of benchmark real datasets along with the employment of various evaluation metrics, i.e. accuracy rate, F-measure, AUC and MI, ensure the findings of this research are well-justified. Application of the Mann-Whitney U-test has confirmed the statistical significance of the comparisons. This implies that the proposed approach can be implemented for many applications including finding potential markers of diseases, virus and bacteria type classification, and cancer prediction. Further work would be devoted to exploring different feature selection methods that may provide great performance specifically for count data classification. With the effectiveness in classifying RNA-seq data, GP models have demonstrated as a promising Bayesian approach in analysis of genomic data. Investigating Bayesian GP models to deal with challenges of other types of biological data is worth another future study.

Supporting Information

S1 File. voom transformation method. This file contains the description of the voom transformation method.
(PDF)

S2 File. Details of the Gaussian process method. This file contains the details of the Gaussian process method.
(PDF)

S3 File. Graphical comparisons of feature selection methods and classifiers. This file contains the graphical comparisons by box plots of feature selection methods and classifiers.
(PDF)

Author Contributions

Conceptualization: TN AB SY SN.

Data curation: TN AB SN.

Formal analysis: TN AB SY SN.

Methodology: TN AB SY SN.

Resources: TN AB.

Software: TN AB.

Validation: TN SY SN.

Writing – original draft: TN AB.

Writing – review & editing: TN AB SY SN.

References

1. Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication?. Bioinformatics. 2014 Feb 1; 30(3):301–4. doi: 10.1093/bioinformatics/btt688 PMID: 24319002
2. Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009 Jan 1; 10(1):57–63. doi: 10.1038/nrg2484 PMID: 19015660
3. Zwiener I, Frisch B, Binder H. Transforming RNA-seq data to improve the performance of prognostic gene signatures. PloS One. 2014 Jan 8; 9(1):e85150. doi: 10.1371/journal.pone.0085150 PMID: 24416353
4. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014 Aug 14; 512(7513):155–60. doi: 10.1038/nature13600 PMID: 25079324
5. Zhang W, Chang JW, Lin L, Minn K, Wu B, Chien J, et al. Network-based isoform quantification with RNA-seq data for cancer transcriptome analysis. PLoS Comput Biol. 2015 Dec 23; 11(12):e1004465. doi: 10.1371/journal.pcbi.1004465 PMID: 26699225
6. Witten DM. Classification and clustering of sequencing data using a Poisson model. The Annals of Applied Statistics. 2011 Dec 1:2493–518. doi: 10.1214/11-AOAS493
7. Ghaffari N, Yousefi MR, Johnson CD, Ivanov I, Dougherty ER. Modeling the next generation sequencing sample processing pipeline for the purposes of classification. BMC Bioinformatics. 2013 Oct 11; 14 (1):1. doi: 10.1186/1471-2105-14-307 PMID: 24118904
8. Srivastava S, Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Research. 2010 Sep 1; 38(17):e170-. doi: 10.1093/nar/gkq670 PMID: 20671027
9. Auer PL, Doerge RW. A two-stage Poisson model for testing RNA-seq data. Statistical Applications in Genetics and Molecular Biology. 2011 May; 10(1). doi: 10.2202/1544-6115.1627
10. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics (Oxford, England). 2012 Jul; 13(3):523–38. doi: 10.1093/biostatistics/kxr031 PMID: 22003245
11. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013 Mar 9; 14(1):1. doi: 10.1186/1471-2105-14-91 PMID: 23497356
12. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in Bioinformatics. 2015 Jan 1; 16(1):59–70. doi: 10.1093/bib/bbt086 PMID: 24300110
13. Si Y, Liu P. An optimal test with maximum average power while controlling FDR with application to RNA-seq data. Biometrics. 2013 Sep 1; 69(3):594–605. doi: 10.1111/biom.12036 PMID: 23889143

14. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Research. 2014 Jun 17; 42(11):e91-. doi: 10.1093/nar/gku310 PMID: 24753412

15. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics. 2008 Apr 1; 9(2):321–32. doi: 10.1093/biostatistics/kxm030 PMID: 17728317

16. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1; 26(1):139–40. doi: 10.1093/bioinformatics/btp616 PMID: 19910308

17. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010 Oct 27; 11(10):1. doi: 10.1186/gb-2010-11-10-r106 PMID: 20979621

18. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010 Aug 10; 11(1):422. doi: 10.1186/1471-2105-11-422 PMID: 20698981

19. Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics. 2013 Apr 1; 14(2):232–43. doi: 10.1093/biostatistics/kxs033 PMID: 23001152

20. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014 Dec 5; 15(12):1. doi: 10.1186/s13059-014-0550-8 PMID: 25516281

21. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010 Mar 2; 11(3):1. doi: 10.1186/gb-2010-11-3-r25 PMID: 20196867

22. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014 Feb 3; 15(2):1. doi: 10.1186/gb-2014-15-2-r29 PMID: 24485249

23. Kuo Y, Yang T, Huang GW. The use of grey relational analysis in solving multiple attribute decision-making problems. Computers & Industrial Engineering. 2008 Aug 31; 55(1):80–93. doi: 10.1016/j.cie.2007.12.002

24. Rasmussen C, Williams C. Gaussian Processes for Machine Learning. The MIT Press, Cambridge, MA; 2006.

25. Laiho A, Elo LL. A note on an exon-based strategy to identify differentially expressed genes in RNA-seq experiments. PloS One. 2014 Dec 26; 9(12):e115964. doi: 10.1371/journal.pone.0115964 PMID: 25541961

26. Theodoridis S, Koutroumbas K. Pattern Recognition. Academic Press, 4th edition, 2009.

27. Choi E, Lee C. Feature extraction based on the Bhattacharyya distance. Pattern Recognition. 2003 Aug 31; 36(8):1703–9. doi: 10.1016/S0031-3203(03)00035-9

28. Deng L, Pei J, Ma J, Lee DL. A rank sum test method for informative gene discovery. In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2004 Aug 22 (pp. 410–419). ACM.

29. Nguyen T, Nahavandi S, Creighton D, Khosravi A. Mass spectrometry cancer data classification using wavelets and genetic algorithm. FEBS Letters. 2015 Dec 21; 589(24):3879–86. doi: 10.1016/j.febslet.2015.11.019 PMID: 26611346

30. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statistical Science. 2003 Feb 1:104–17. doi: 10.1214/ss/1056397488

31. Bickel PJ, Levina E. Some theory for Fisher's linear discriminant function,'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli. 2004 Dec 1:989–1010. doi: 10.3150/bj/1106314847

32. Witten DM, Tibshirani R. Penalized classification using Fisher's linear discriminant. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011 Nov 1; 73(5):753–72. doi: 10.1111/j.1467-9868.2011.00783.x PMID: 22323898

33. Ju-Long D. Control problems of grey systems. Systems & Control Letters. 1982 Mar 1; 1(5):288–94. doi: 10.1016/S0167-6911(82)80025-X

34. Wu HH. A comparative study of using grey relational analysis in multiple attribute decision making problems. Quality Engineering. 2002 Dec 1; 15(2):209–17. doi: 10.1081/QEN-120015853

35. Lu JC, Yeh MF. Robot path planning based on modified grey relational analysis. Cybernetics & Systems. 2002 Mar 1; 33(2):129–59. doi: 10.1080/019697202753435908

36. Nguyen T, Nahavandi S. Modified AHP for gene selection and cancer classification using type-2 fuzzy logic. IEEE Transactions on Fuzzy Systems. 2016 Apr; 24(2):273–87. doi: 10.1109/TFUZZ.2015.2453153

37. Seeger M. Pac-bayesian generalisation error bounds for gaussian process classification. Journal of Machine Learning Research. 2002; 3(Oct):233–69.

38. Kuss M, Rasmussen CE. Assessing approximate inference for binary Gaussian process classification. Journal of Machine Learning Research. 2005; 6(Oct):1679–704.

39. Rasmussen CE, Nickisch H. Gaussian processes for machine learning (GPML) toolbox. Journal of Machine Learning Research. 2010; 11(Nov):3011–5.

40. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010 Apr 1; 464(7289):773–7. doi: 10.1038/nature08903 PMID: 20220756

41. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010 Apr 1; 464(7289):768–72. doi: 10.1038/nature08872 PMID: 20220758

42. Frazee AC, Langmead B, Leek JT. ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. BMC Bioinformatics. 2011 Nov 16; 12(1):449. doi: 10.1186/1471-2105-12-449 PMID: 22087737

43. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, et al. NCBI GEO: mining millions of expression profiles—database and tools. Nucleic Acids Research. 2005 Jan 1; 33(suppl 1):D562–6. doi: 10.1093/nar/gki022 PMID: 15608262

44. 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 May 11; 8(1):1. doi: 10.1186/1741-7007-8-58 PMID: 20459774

45. Smyth GK. limma: linear models for microarray data. [http://www.bioconductor.org/packages/release/bioc/html/limma.html]. 2016.

46. Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning. 2003 Oct 1; 53(1–2):23–69. doi: 10.1023/A:1025667309714

47. Gilad-Bachrach R, Navot A, Tishby N. Margin based feature selection-theory and algorithms. In Proceedings of the Twenty-First International Conference on Machine Learning 2004 Jul 4 (p. 43). ACM.

48. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999 Oct 15; 286(5439):531–7. doi: 10.1126/science.286.5439.531 PMID: 10521349

49. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, et al. Gene selection from microarray data for cancer classification—a machine learning approach. Computational Biology and Chemistry. 2005 Feb 28; 29(1):37–46. doi: 10.1016/j.compbiolchem.2004.11.001 PMID: 15680584

50. Mitchell T. Machine Learning. McGraw Hill, 1997.

51. Bishop C. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, 1995.

52. Kecman V. Learning and Soft Computing. MIT Press, Cambridge, MA, 2001.

53. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences. 1997; 55(1):119–39. doi: 10.1006/jcss.1997.1504

54. Gibbons JD, Chakraborti S. Nonparametric Statistical Inference, 5th Ed., Boca Raton, FL: Chapman & Hall/CRC Press, Taylor & Francis Group. 2011.

55. Nguyen T, Khosravi A, Creighton D, Nahavandi S. EEG signal classification for BCI applications by wavelets and interval type-2 fuzzy logic systems. Expert Systems with Applications. 2015 Jun 1; 42(9):4370–80. doi: 10.1016/j.eswa.2015.01.036