An Integrated RNA and DNA Molecular Signature for Colorectal Cancer Classification

Mohanad Mohammed (mohanadadam32@gmail.com)
University of Kwazulu-Natal https://orcid.org/0000-0001-5336-0286
Henry Mwambi
University of Kwazulu-Natal
Bernard Omolo
University Of South Carolina Upstate

Research article

Keywords: Colorectal cancer, FFPE, Microarray, RAS pathway signatures, RNASeq

Posted Date: January 8th, 2020

DOI: https://doi.org/10.21203/rs.2.20271/v1

RESEARCH

An Integrated RNA and DNA Molecular Signature for Colorectal Cancer Classification

Mohanad Mohammed[1*], Henry Mwambi[1] and Bernard Omolo[2,3]

*Correspondence:
mohanadadam32@gmail.com
[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, King Edward Avenue, Pietermaritzburg, South Africa
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Background: Colorectal cancer (CRC) is the third most common cancer among women and men in the USA and recent studies have shown an increasing incidence in less developed regions such as Sub-Saharan Africa (SSA). The *KRAS* gene is mutated in 40% of the CRC cases and hence the RAS pathway activation has become a major focus of drug targeting efforts. However, nearly 60% of patients with wild-type KRAS fail to respond to RAS-targeted therapies, for example the anti-epithelial growth factor receptor inhibitor (EGFRi) combination therapies. Thus, there is a need to develop more reliable molecular signatures to better predict mutation status. In this study, we develop a hybrid (DNA mutation and RNA expression) signature and assess its predictive properties for the mutation status and survival of CRC patients.

Methods: Publicly-available microarray and RNASeq data from 54 matched formalin-fixed paraffin embedded (FFPE) samples from the Affymetrix GeneChip and RNASeq platforms, were used to obtain differentially expressed genes between mutant and wild-type samples. For classification, the support-vector machines, artificial neural networks, random forests, k-nearest neighbor, naïve Bayes, negative binomial linear discriminant analysis, and the Poisson linear discriminant analysis algorithms were employed.

Results: Compared to the genelist from each of the individual platforms, the hybrid genelist had the highest accuracy, sensitivity, specificity and AUC for mutation status, across all the classifiers, and is prognostic for survival in patients with CRC.

Conclusions: This signature could be useful in clinical practice, especially for colorectal cancer diagnosis and therapy.

Keywords: Colorectal cancer; FFPE; Microarray; RAS pathway signatures; RNASeq

Background

Colorectal cancer (CRC) is one of the major emerging causes of mortality and morbidity around the world [1]. CRC is also the third leading cause of death among men and women [2, 3, 4, 5, 6]. There were about 1.80 million new cases and $862,000$ deaths in 2018, according to the World Health Organization (WHO) [7]. Furthermore, in 2019, CRC was reported to be the third most prevalent cancer among men and women and an estimated $101,420$ and $44,180$ new cases of colon and rectal cancer, respectively, and $51,020$ deaths in the USA alone [8, 2, 9].

Although the incidence rates of CRC are lower in developing countries than in developed countries, recent studies have shown an increase in the incidence rates in

Sub-Saharan Africa [10]. Many cancer types that are relatively curable in developed countries are detected only at advanced stages in developing countries, due to late or inaccurate diagnoses [11]. Cancer tumor classification based on morphological characteristics alone has been shown to have serious limitations in some studies [12]. Physicians aim to diagnose CRC as early as possible to design optimal treatment strategies that are patient-specific. Therefore, using genetic mutation and features of the tumor would most probably lead to better understanding and early detection of the disease and lead to finding suitable and targeted strategies [13].

Previously, most of the cancer classification research was based on clinical features of the tumors, which lacked accurate diagnostic ability, hence the need to develop new methods to better address this important problem [12, 14]. Recently, DNA microarray technology has greatly improved the classification of diseases into subtypes, particularly cancer. This technology allows the processing of thousands of genes simultaneously, hence providing critical information about a disease[15, 16]. Microarray gene expression data have been used widely for cancer detection, prediction and diagnosis [17]. In the last decade, next-generation sequencing (NGS) technology has emerged as an advancement in cancer and other disease research, based on RNA sequencing methodology. NGS technology permits the measurement of expression levels of tens of thousands of genes simultaneously. NGS platforms that are most common include Illumina, SOLiD, Ion Torrent semiconductor sequencing, and single-molecule real-time sequencing [18].

NGS technology has been the most attractive and dramatically been improved over the last few years. This technology is high-throughput and has become popular in the detection and analysis of differentially expressed genes[18, 19]. More recently, RNASeq data has been shown to be better than microarray data in terms of quality and accuracy in estimating transcript abundance. However, the two methodologies are different in design and implementation [20, 19, 21]. Although RNASeq experiments are expensive, in contrast, they have many advantages over microarrays. RNASeq allows detecting the variation of a single nucleotide, does not require genomic sequence knowledge, provides quantitative expression levels, provides isoform-level expression measurements, and offers a broader dynamic range than microarrays [20]. Moreover, RNASeq allows the detection of novel transcripts, low background signal, and increased specificity and sensitivity [22]. However, our view is that integrated use of data from both technologies may be the best approach, given the available information from both technologies.

Microarray and RNASeq technologies produce gene expression data in different forms. The structure of gene expression produced using microarrays is continuous data, while RNASeq provides a discrete type of data [23]. What is common between the two technologies is that both generate big datasets consisting of a few sample sizes, where each sample has a large number of genes.

Many statistical and machine learning methods have been used to analyze and extract information from massive amounts of gene expression data. These methods include the Poison linear discriminant analysis (PLDA), negative binomial linear discriminant analysis (NBLDA), support vector machines (SVM), artificial neural networks (ANN), linear discriminant analysis (LDA), and random forests (RF).

These methods have been used and examined in many studies based on RNASeq and microarray data. For example, Aziz]et al. [24] assessed the ANN performance

based on microarray data using six hybrid feature selection methods. Five gene expression datasets were used for evaluating these methods and for understanding how these methods can improve the performance of ANN. Statistical hypothesis tests were used to check the differences between these methods. They showed that the combination of independent component analysis (ICA) and genetic bee colony algorithm had superior performance. Salem et al. [25] proposed a new methodology for gene expression data analysis. They combined information gain (IG) and standard genetic algorithm (SGA) for feature selection and reduction, respectively. Their approach was tested on seven cancer datasets and then compared with the most recent approaches. Their results show that the proposed approach outperformed the most recent approaches. Jain et al. [26] presented a two-phase hybrid method for cancer classification using eleven microarray datasets for different cancer types. They combined correlation-based feature selection (CFS) and improved-binary particle swarm optimization (IBPSO). Naive Bayes with 10-fold cross-validation was used for assessment. Results indicated that their approach had better performance in terms of accuracy and the number of selected genes.

Anders and Huber [27] conducted differential expression analysis based on the negative binomial distribution, with variance and mean linked by local regression, for count data. They implemented their proposed method by the *DESeq* R package. Zararsiz et al. [23] presented a comprehensive simulation study on RNASeq classification using PLDA, NBLDA, single SVM, bagging SVM (bagSVM), classification and regression trees (CART), and RF. Their simulation results were applied and compared to two miRNA and two mRNA real experimental datasets. They found that the power-transformed PLDA, RF, and SVM were the best in classification performance.

Due to the small number of samples for gene expression data, combining independent datasets is novel in order to increase sample size and statistical power. Taminau et al. [28] worked on the integration of gene expression analysis using two approaches based on merging and meta-analysis. They used six gene expression datasets. Results showed that both meta-analysis and merging did well, but merging was able to detect more differentially expressed genes than meta-analysis.

Recently, combining two different gene expression data sources has been shown to improve classification accuracy as opposed to using only one source. Castillo and co-workers [20] introduced integration of multiple microarrays and RNASeq platforms. They first carried out a differential expression analysis, then applied the minimum-redundancy maximum-relevance (mRMR) feature selection approach for further reduction of the gene-list. The top 10 genes were selected and evaluated using four classification methods: k-Nearest neighbor (KNN), Naive Bayes (NB), RF, and SVM. Their results showed the highest accuracy and *f1*-score for the KNN. Here, we combined RNASeq and DNA expression data from colorectal cancer patients. We obtained a hybrid gene-list from the RNASeq and microarray datasets and assessed its classification performance based on the PLDA, NBLDA, SVM, RF, ANN, KNN, and NB algorithms.

The paper is structured as follows. Section 2 discusses the methods and the performance metrics used in the study. Section 3 shows the classification results of the microarray, RNASeq, hybrid gene lists, and the survival analysis. Discussion and conclusions are presented in sections 4 and 5 respectively.

Materials and Methods

Datasets

We used publicly available microarray and RNASeq data that is also reported in Omolo *et al.* [4]. The data consists of 54 matched formalin-fixed paraffin-embedded (FFPE) samples from colorectal cancer patients. The data is available in the gene expression omnibus (GEO) repository under the accession numbers GSE86562 and GSE86559 for RNASeq and microarray data, respectively. The microarray gene expression data consists of $60,607$ genes on 54 colorectal patients. KRAS mutation status was used as a class variable. As a first step, the Affymetrix microarray data was log2-transformed and quantile-normalized, and genes with more than 50% missing values were filtered out. Thereafter, class comparison was done using the two-sample t-test at the 0.005 significant level threshold, which yielded 165 differentially expressed genes.

The RNASeq dataset contained $57,905$ genes from the same colorectal cancer patients used to generate the microarray data. This data is in the form of counts i.e. discrete in nature. For this data, first, filtration was done to remove the genes with more than 50% of zeros across the samples, using the counts per million (CPM) method [29]. Genes whose CPM values are greater than 0.5 were retained. Thus, the dimension was reduced to $17,473$ genes. Differential expression analysis was performed using *DESeq2* package in *R*. This step reduced the genes to 282 genes using the 0.005 significance level threshold. The differential expression analysis tool in *DESeq2* uses a generalized linear model (GLM) of the following form:

$$\begin{aligned} g_{ij} &\sim NB(\mu_{ij}, \alpha_i) \\ \mu_{ij} &= s_j q_{ij} \\ log_2(q_{ij}) &= x_{j.}\beta_i, \end{aligned} \tag{1}$$

where g_{ij} is the counts for gene i in sample j. These counts are modeled using a negative binomial distribution with fitted mean μ_{ij} and a gene-specific dispersion parameter α_i. The fitted mean is decomposed into a sample-specific size factor s_j and a parameter q_{ij} proportional to the expected true concentration of fragments for sample j. The coefficients β_i represent the log_2-fold changes for gene i for each column of the model or design matrix X. Note that the model can be generalized to use sample- and gene-dependent normalization factors s_{ij}.

The dispersion parameter α_i defines the relationship between the variance of the observed count and its mean value. That is, how far we expect the observed count to be from the mean value, which depends both on the size factor s_j and the covariate-dependent part q_{ij} as defined above. Thus the variance function is given by

$$Var(g_{ij}) = E\left[(g_{ij} - \mu_{ij})^2\right] = \mu_{ij} + \alpha_i \mu_{ij}^2. \tag{2}$$

The steps performed by the *DESeq* function in *DESeq2* package are estimation of s_j, and α_i, and fitting negative binomial GLM for β_i and Wald statistics by *nbinomWaldTest*.

Counts per million can be computed as

$$CPM_i = \frac{g_i}{N} * 10^6, \tag{3}$$

where g_i denotes the counts observed from a gene of interest i and N is the number of sequenced fragments.

RNASeq and microarray data integration may help improve cancer classification accuracy. Several studies have addressed the classification problem using RNASeq, microarray, or a combination of both, based on heterogeneous samples ([20, 30, 31]). Our study aims to integrate homogeneous samples from the RNASeq and microarray platforms. In this regard, we obtained the differentially expressed genes from the two platforms based on the same set of samples. Thereafter, we used the database for annotation, visualization, and integrated discovery (DAVID) [32], and catalogue of somatic mutations in cancer (COSMIC) tools, to annotate the RNASeq transcripts list. The microarray genes symbol names were obtained from the dataset in [4]. Next, the intersection, complement of the intersection, and union between the two annotated lists were obtained.

Integration was done using the intersection, complement of the intersection, and the union of the two lists of genes. Due to the different nature of the two datasets, RNASeq was log_2 transformed and quantile- normalized in order to make both types of data consistent with each other. Subsequently, the integration was done based on binding the two gene-lists from the RNASeq and microarray datasets. To transform the RNASeq data, we let

$$\text{Transformed Data} = log_2(G + 1), \tag{4}$$

where G is the RNASeq counts data matrix, and $G + 1$ is the RNASeq counts data matrix with all zero counts changed to one.

Quantile normalization ensures that probe intensities of each array in a set of arrays have the same distribution. A quantile-quantile plot would help to confirm if two probe vectors have the same distribution (quantiles lie on the diagonal line) or not. This approach can be extended to n-dimensional data. Let $\boldsymbol{q}_k = (q_{k1}, \ldots, q_{kn})'$, $k = 1, \ldots, P$, be the vector of the k^{th} quantiles for all n arrays, and $\boldsymbol{d} = (\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})'$ be the unit diagonal. To transform from the quantiles so that they all lie along the diagonal, we projected $\boldsymbol{q}$ on to $\boldsymbol{d}$ as below [33]:

$$\text{proj}_{\boldsymbol{dq}_k} = \left(\frac{1}{n} \sum_{j=1}^{n} q_{1j}, \ldots, \frac{1}{n} \sum_{j=1}^{n} q_{Pj} \right). \tag{5}$$

Data Integration

Here, we used homogeneous data from matched-pair samples from microarray and RNASeq technologies. Using a set-theoretic approach of taking the intersection, complement of intersection, or union, we obtained four lists of genes from the two platforms at the 0.005 significance level. The intersection between the two lists was 23 genes, with 401 genes being the complement of the intersection.

Classification Methods

Several methods have been developed for classification and their performance evaluated in both microarray and RNASeq platforms. Below, we briefly describe seven classification methods and how to evaluate their performances based on the integration of the two platforms.

Poisson Linear Discriminant Analysis

The PLDA classifier was proposed by Witten [34]. Witten used the Poisson log linear model, and developed an analog of diagonal linear discriminant analysis for sequence data.

Let $\mathbf{G}$ denote a $n \times p$ matrix of read counts data, where n denotes the number of observations (samples), and p the number of genes. Let G_{ij} be the counts or reads for gene j in sample i; it is reasonable to assume that

$$G_{ij} \sim Poisson(\mu_{ij}), \tag{6}$$

where $\mu_{ij} = s_i g_j$. To avoid identifiability issues, one can require $\sum_{i=1}^{n} s_i = 1$, where s_i is the number of counts per sample i, and g_j is the number of counts per gene j.

Suppose that we have K different classes of samples. Then we can write

$$G_{ij}|y_i = k \sim Poisson(\mu_{ij} d_{kj}), \tag{7}$$

where y_i denotes the class of the ith sample ($y_i = 1, 2, 3, \cdots, K$) and d_{kj} denotes a measure of the level of the jth gene to be differentially expressed in class k.

Let $g_i = (g_{i1}, g_{i2}, \cdots, g_{ip})'$ indicate the entries of row i in the $\mathbf{G}$ matrix, which are the gene expression levels of sample i. Let, $G_{.j} = \sum_{i=1}^{n} G_{ij}$, $G_{i.} = \sum_{j=1}^{p} G_{ij}$, and $G_{..} = \sum_{i,j} G_{ij}$ denote the column, row and the overall totals respectively. The maximum likelihood estimate (MLE) for μ_{ij} assuming independence is $\hat{\mu}_{ij} = \frac{G_{i.} G_{.j}}{G_{..}}$, and $\sum_{i=1}^{n} \hat{s}_i = 1$ yields the estimates $\hat{s}_i = \frac{G_{i.}}{G_{..}}$ and $\hat{g}_j = G_{.j.}$. $\hat{s}_i$ is the estimate of the size factor for sample i. Maximum likelihood estimation provides the estimate of d_{kj} as $\hat{d}_{kj} = \frac{G_{c_k j}}{\sum_{i \in c_k} \hat{\mu}_{ij}}$, where c_k denotes the class of an observation.

If $\hat{d}_{kj} > 1$, then the jth gene is over-expressed relative to the baseline in the kth class, and if $\hat{d}_{kj} < 1$ then the jth gene is under-expressed relative to the baseline in the kth class. If $G_{c_k j} = 0$ (an event that is not unlikely if the true mean for jth gene is small), then the maximum likelihood estimate for d_{kj} equals zero.

Assume that we want to classify a new observation $g^* = (G_1^*, \cdots, G_p^*)$, and let y^* indicate the unknown class label. By Bayes rule,

$$P(y^* = k|g^*) \propto f_k(g^*)\pi_k, \tag{8}$$

where f_k is the density of a sample in class k and π_k is the prior probability that an observation belongs to class k. Then, if f_k is a normal density with a class specific

mean and common covariance, PLDA classifies a new sample to class k, which maximize equation (8). Consequently, the discriminant score of PLDA is

$$\log Pr(y^* = k|g^*) \approx \sum_{j=1}^{P} G_j^* \log d_{kj} - \sum_{j=1}^{P} s^* \lambda_j d_{kj} + \log \pi_k + C. \tag{9}$$

PLDA is implemented using the R package *MLSeq*.

Negative Binomial Linear Discriminant Analysis

Recently, Dong *et al.* [22] proposed NBLDA for RNASeq data analysis. NBLDA and Poisson linear discriminant analysis (PLDA) were considered the most suitable classifiers for RNASeq data, due to the discrete nature of data [22, 34].

Let G_{ij} denote the number of reads in sample i, and gene j, $i = 1, 2, 3, \cdots, n$ and $j = 1, 2, 3, \cdots, p$. Then G_{ij} is assumed to follow the negative binomial distribution

$$G_{ij} \sim NB(\mu_{ij}, \phi_j), \qquad \mu_{ij} = s_i \lambda_j, \tag{10}$$

where s_i is the size factor, used to scale gene counts for the ith sample due to different sequencing depth, λ_j is the total number of reads per gene, and $\phi_j \geq 0$ is the dispersion parameter. The mean and variance of the negative binomial distribution are given by:

$$\begin{aligned} E(G_{ij}) &= \mu_{ij} \\ V(G_{ij}) &= \mu_{ij} + \mu_{ij}^2 \phi_j. \end{aligned} \tag{11}$$

Suppose that we have M classes. Let C_M be an indicator variable such that $C_m \in \{1, 2, 3, \cdots, M\}$. Then, the model for RNASeq data is

$$(G_{ij}|y_i = m) \sim NB(\mu_{ij} d_{m,j}, \phi_j), \tag{12}$$

where $d_{m,j}$ denotes the differences among the M classes, and $y_i = m, m \in \{1, 2, 3, \cdots, M\}$ denotes the class of samples i. The assumption is that all the genes are independent.

Let $\boldsymbol{g^*} = (G_1^*, \cdots, G_p^*)$ be a new sample whose class is to be predicted, s^* is the size factor, and y_i^* the class label value. By Bayes' rule, we have

$$\Pr(y^* = m|\boldsymbol{g^*}) \propto f_m(\boldsymbol{g^*})\pi_m, \tag{13}$$

where f_m is the *pdf* of the sample in class m, and π_m is the prior probability that a sample comes from class m. The *pdf* of $G_{ij} = \mathrm{g}_{ij}$ in equation (12) is

$$\Pr(G_{i,j} = \mathrm{g}_{ij}|y_i = m) = \frac{\Gamma(\mathbf{g}_{ij} + \phi_j^{-1})}{\mathrm{g}_{ij}!\Gamma(\phi_j^{-1})}\left(\frac{s_i\lambda_j d_{mj}\phi_j}{1 + s_i\lambda_j d_{mj}\phi_j}\right)^{\mathrm{g}_{ij}}\left(\frac{1}{1 + s_i\lambda_j d_{mj}\phi_j}\right)^{\phi_j^{-1}}. \tag{14}$$

Thus, the discriminant score for NBLDA can be constructed from (13) and (14) as

$$\begin{aligned}\log\Pr(y^* = m|\boldsymbol{g}^*) = &\sum_{j=1}^{P} G_j^*[\log d_{mj} - \log(1 + s_i\lambda_j d_{mj}\phi_j)] \\ &- \sum_{j=1}^{P}\phi_j^{-1}\log(1 + s_i\lambda_j d_{mj}\phi_j) + \log\pi_m + C,\end{aligned} \tag{15}$$

where C is a constant independent of m. The class m, which maximizes the score in equation (15) will be assigned to the new sample $\boldsymbol{g}^*$. NBLDA is implemented using the *R* package *MLSeq*.

Support Vector Machines

The SVM method was first proposed by Boser, Guyon and Vapnik [35] at the Computational Learning Theory (COLT92) ACM Conference in 1992. The method is based on the idea of a hyperplane that lies furthermost from both classes. This plane is known as the *optimal (maximum) margin hyperplane*. The hyperplane is completely determined by a sub-set of the samples known as the *support vectors* [36]. SVM has the ability to handle problems where the data are not linearly separable by transforming the data using mapping kernel functions such as the radial basis function (RBF) kernel, polynomial function, and the linear function [37]. In addition, SVM has can handle high dimensional data, which is clearly an important advantage in dealing with genetic data from cancer studies. This strength makes SVM widely appealing and applicable to real-life data analysis problems such as handwritten character recognition, human face recognition, radar target identification, speech identification, and, quite recently, to gene expression data analysis [38, 39].

Suppose we have n samples and p genes. Further, assume samples belong to two distinct outcome classes represented by $+1$ or -1 and a feature vector $\mathbf{g}_i$ such that $(\mathbf{g}_i, y_i) \in G \times Y \quad i = 1, 2, \ldots n$, where $\mathbf{g}_i = (g_{i1}, g_{i2}, \ldots, g_{ip})'$ is the sample profile (vector) and $y_i \in \{+1, -1\}$ is the outcome class dichotomy. The goal is to classify the samples into one of the two classes by training the SVM which maps the input data (using a suitable kernel function) onto a high-dimensional space (feature space) $\{(\Phi(\mathbf{g}_i), y_i)\}_{i=1}^n$. This is achieved by constructing an optimal separating hyperplane that lies furthest from both classes.

The general form of a separating hyperplane in the space of the mapped data is defined by

$$\mathbf{w}^T\Phi(\mathbf{g}) + b = 0. \tag{16}$$

Here, $\mathbf{w} = (w_1, w_2, \ldots, w_n)'$, is the weight vector. We can rescale the $\mathbf{w}$ and b such that the following equation determines the point in each class that is nearest to the hyperplane defined by the equation

$$|\mathbf{w}^T\Phi(\mathbf{g}) + b| = 1. \tag{17}$$

Therefore, it should follow that for each sample i, $i \in \{1, 2, \ldots, n\}$,

$$\mathbf{w}^T\Phi(\mathbf{g}_i) + b = \begin{cases} \geq 1 & \text{if} \quad y_i = +1 \\ \leq -1 & \text{if} \quad y_i = -1. \end{cases} \tag{18}$$

After the rescaling, the distance from the nearest point in each class to the hyperplane becomes $\frac{1}{\|\mathbf{w}\|}$. Thus, the distance between the two classes is $\frac{2}{\|\mathbf{w}\|}$, which is called the *margin*. To maximize the margin, the solution of the following optimization problem is obtained:

$$\begin{aligned} &\min_{\mathbf{w},b} \quad \|\mathbf{w}\|^{\mathbf{2}} \\ &\text{subject to} \\ &y_i(\mathbf{w}^T\Phi(\mathbf{g}_i) + b) \geq 1, \quad i = 1, 2, \ldots, n. \end{aligned} \tag{19}$$

The square of the norm of $\mathbf{w}$ is considered in order to make the problem quadratic. Suppose $\mathbf{w}^*$ and b^* are the solutions to the optimization problem (19) above. Then this solution determines the hyperplane in the feature space where $(\mathbf{w}^*)^T\Phi(\mathbf{g}) + b^* = 0$. The points $\Phi(\mathbf{g}_i)$ that satisfy the qualities $y_i((\mathbf{w}^*)^T\Phi(\mathbf{g}_i) + b^*) = 1$ are called *support vectors* [36]. The SVM method is implemented using the *R* package *kernlab* [40].

Random Forests

Random forests were first introduced in 2001[41, 42]. They are an extension of classification and regression trees, and also an improvement over bagged trees by further modification using a random small tweak to de-correlate the trees. Growing random forests leads to an improvement in prediction accuracy compared to single or bagged trees [43].

We build a number of forests of decision trees on bootstrapped training samples from the original data. A tree is obtained by recursively splitting the genes such that at each node of the tree, a candidate gene for splitting is obtained from a random sample of size v. A typical choice for v is such that $v \approx \sqrt{p}$, where p is the number of candidate genes for splitting.

We then grow the trees to maximum depth. Therefore, the two-step randomization process helps to de-correlate the trees [44]. To determine the prediction for an unknown sample, an average over all the trees is taken for a regression problem and a majority vote for a classification problem [41, 45, 46].

Random Forest Algorithm for Regression or Classification [41]

1 For $b = 1$ to B (# random-forest trees):
   - Draw a bootstrap sample of size N from the training data.
   - Grow a random-forest tree, T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size, n_{min}, is reached.
   - Select v genes at random from the p genes.
   - Pick the best gene to split on among the v based on an impurity measure.
   - Using the selected gene, split the node into two daughter nodes.

2 To make a prediction for a new sample, x: Let $\hat{C}_b(x)$ be the class prediction of the b-th random-forest tree. Then

$$\hat{C}_{rf}^{B}(x) = \text{majority vote} \left\{ \hat{C}_b(x) \right\}_{b=1}^{B}$$

RF is implemented using the *R* package *randomForest* [47].

Artificial Neural Networks

Artificial neural networks (ANN) are multi-layered models that are constructed from three layers, each layer consisting of nodes called *neurons* [48]. The input layer contains nodes whose number is based on the input features. The output layer contains nodes equal to the number of classes, and finally, the hidden layer contains nodes determined by the level of tuning required. The inputs are weighted by multiplying each input by weight as a measure of its contribution. The layers are connected together via connection weights. These weights are determined through stages of model fitting. The hidden nodes receive the sum weighted from the input layer plus some bias. This summation is passed onto the transform function (activation function) to generate the results. These results are called *outputs* and interpreted as a class probability in our case.

There are many types of architecture of ANN. Neural networks are used widely in different fields such as prediction in time series models, economic modeling, and medical applications [37]. In addition, ANN can be applied to the classification problem using microarray gene expression data [48]. In this paper, we apply the method to both microarray and RNA Sequencing gene expression data.

Consider the simplest multi-layered network, with one hidden layer. Assume we have gene expression data where n denotes the number of genes. Then the input layer receives the n gene expression levels for a sample, each multiplied by the corresponding weight, $w_{ij}^{(1)} g_j$, as shown in Eq 20, below:

$$b_i = \sum_{j=0}^{n} w_{ij}^{(1)} g_j \quad i = 1, 2, \ldots, m, \tag{20}$$

where $\mathbf{g} = (g_0, g_1, g_2, \ldots, g_n)'$ is a vector of input features and $g_0 = 1$ is a constant input feature that with weight w_{i0}. The quantities, b_i, are called *activations*, and the parameters $w_{ij}^{(1)}$ are the weights. Note that alternatively b_i can be viewed as a summary of the n genes from sample i. The superscript "(1)" indicates that this is the first layer of the network. Each of the activations is then transformed by a nonlinear activation function f, typically a sigmoid, as in Eq 21 below:

$$z_i = f(b_i) = \frac{1}{1 + \exp(-b_i)}. \tag{21}$$

The quantities z_i are interpreted as the output of hidden units, so called because they do not have values specified by the problem (as is the case for input units) or target values used in the training (as is the case for output units).

In the second layer, the outputs of the hidden units are linearly combined to give the activations

$$a_k = \sum_{i=0}^{m} w_{ik}^{(2)} z_i \quad k = 1, 2, \ldots, K. \tag{22}$$

Again, $z_0 = 1$ corresponds to the bias. The transformations in the second layer of the neural network are parameterized by weights $w_{ik}^{(2)}$. The output units are transformed using an activation function. Again, a sigmoid function may be used as shown below:

$$y_k = f(a_k) = \frac{1}{1 + \exp(-a_k)}. \tag{23}$$

These equations may be combined to give the overall equation that describes the forward propagation through the network, and describes how an output vector is computed from an input vector, given the weight matrices as

$$y_k = f\left(\sum_{i=0}^{m} w_{ik}^{(2)} f\left(\sum_{j=0}^{n} w_{ij}^{(1)} g_i\right)\right). \tag{24}$$

ANN are implemented using the *R* package *nnet* [49].

Naive Bayes

The Naive Bayes classifier uses probability theory to find the most likely of the possible classes in a classification problem. The NB classifier relies on two assumptions, namely, that each attribute is conditionally independent from the other attributes given the class and that all the attributes have an influence on the class [50]. The popularity of this classifier is mainly due to its simplicity, yet exhibiting a surprisingly competitive predictive accuracy. The NB classifier has previously been applied in many fields, including microarray gene expression data [37, 48].

Consider an n x p gene expression data matrix, where n is the number of the samples and p is the number of the genes (features). Let $g_{kj}, j = 1, 2, \ldots, p$, denote the j-th gene on the k-th sample. Let C_i be the i-th class, $i = 1, 2, \ldots, L$. The Naive Bayes classifier uses the *maximum a posteriori* (MAP) classification rule to classify these samples. The probability of the k-th sample gene information vector, $\mathbf{G}_k = (g_{k1}, g_{k2}, \ldots, g_{kp})'$, is calculated and then the sample is assigned the class with largest probability from L conditional probabilities. Let $P(C_1|\mathbf{G}_k), P(C_2|\mathbf{G}_k), \ldots, P(C_L|\mathbf{G}_k)$ denote the set of L conditional probabilities. The NB classification depends on the Bayes rule, which states that a posterior probability

$$P(C_i|\mathbf{G}_k) = \frac{P(\mathbf{G}_k|C_i)P(C_i)}{P(\mathbf{G}_k)} \propto P(\mathbf{G}_k|C_i)P(C_i),\ k = 1, 2, \ldots, n, \tag{25}$$

where $P(\mathbf{G}_k)$ is considered a common normalizing factor for all the L probabilities.

The NB classification assumes that all input features are conditionally independent, that is,

$$\begin{aligned} P(g_{k1}, g_{k2}, \ldots, g_{kp}|C_i) &= P(g_{k1}|g_{k2}, \ldots, g_{kp}, C_i)P(g_{k2}, \ldots, g_{kp}|C_i) \\ &= P(g_{k1}|C_i)P(g_{k2}, \ldots, g_{kp}|C_i) \\ &= P(g_{k1}|C_i)P(g_{k2}|C_i)\ldots P(g_{kp}|C_i). \end{aligned} \tag{26}$$

Ultimately, NB classifies a new sample, $\mathbf{G}^*$, according to the model with MAP probability given the sample, as

$$\text{Class}(\mathbf{G}^*)_{MAP} = \text{argmax}(P(C_i|\mathbf{G}^*)). \tag{27}$$

NB is implemented using the *R* package *naivebayes*.

k-Nearest Neighbors

The k-nearest neighbor classifiers (KNN) are known to be the most useful instance-based learners. KNN is a non-parametric model [51]. If the classification is based on Euclidean distance in a feature space, then k determines the number of neighbors to be used. In the testing set, the new sample is assigned to the class that is most likely among the k neighbors. Then the number of neighbors can be tuned to choose the optimal fitted model parameters [37, 48].

The KNN uses the Euclidean distance measure to find the closest samples for the new sample. Suppose we have two samples, each one with n genes. Denote the two samples as $S_1 = (g_{11}, g_{12}, \ldots, g_{1n})'$ and $S_2 = (g_{21}, g_{22}, \ldots, g_{2n})'$. Then the Euclidean distance is calculated as the square root of the sum of the squared differences in their corresponding values. Using the Euclidean distance formula, the distance between two points, $\text{dist}(S_1, S_2)$, is given as

$$\text{dist}(S_1, S_2) = \sqrt{\sum_{j=1}^{n}(g_{1j} - g_{2j})^2}, \tag{28}$$

where a large dist(S_1, S_2) means the two samples belong to different classes and values near zero suggest that the samples are homogeneous. KNN is implemented using the *R* package *caret*.

Performance Metrics

Several performance measures exist in the literature that can be used to assess classification based on microarray and RNASeq gene expression data. The metrics include accuracy, sensitivity, specificity, kappa coefficient, AUC, balanced error rate (BER), Matthews correlation coefficient (MCC), and the F-measure (F1-score) [52, 53]. Let TP (number of positive samples correctly classified), TN (number negative samples classified correctly), FP (number of negative samples incorrectly classified as positive), and FN (number of positive samples incorrectly classified as negative) denote true positives, true negatives, false positives, and false negatives, respectively. The performance measures can be expressed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{(TP + FN)}$$

$$Specificity = \frac{TN}{(TN + FP)}$$

$$\text{Kappa coefficient} = \frac{\text{Accuracy} - \text{Random Accuracy (RA)}}{1 - \text{Random Accuracy (RA)}}$$

$$RA = \frac{(TN + FP) \times (TN + FN) + (FN + TP) \times (FP + TP)}{(TP + TN + FP + FN) \times (TP + TN + FP + FN)}$$

$$AUC = \frac{1}{2}\left(\frac{TP}{(TP + FN)} + \frac{TN}{(TN + FP)}\right)$$

$$BER = \frac{1}{2}\left(\frac{FP}{(TN + FP)} + \frac{FN}{(FN + TP)}\right)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$\text{F1-measure} = \frac{2TP}{2TP + FP + FN}$$

Results

The analysis of RNASeq data using the integrated list of genes was performed using *R* statistical software. Assessment of the methods was done using 10-fold cross-validation. Here, the 54 CRC samples were divided into 10 folds randomly, with each fold consisting of about 5 - 6 samples. Thereafter, 9 folds were used for model-building and one fold for the testing and validation. Thus, this process was self-iterated 10 times, and the average of the 10 iterations was used to obtain the model performance measures.

Table 1 below provides the number of genes obtained through the intersection, complement of intersection and union of the gene-lists from differential expression analysis (RNASeq: GSE86562, Microarray: GSE86559).

Table (1) Number of genes in the intersection, complement of intersection, and union of the two gene-lists.

Dataset	Total of DEGs	Intersection	Complement of intersection	Union
GSE86559	165	23	142	424
GSE86562	282		259	

Table 2 below shows the official gene symbols and the corresponding gene names.

Table (2) Official gene symbols and corresponding gene names for the 23 genes.

Ensemble Gene ID	Official Gene Symbol	Name
ENSG00000108511	HOXB6	homeobox B6(HOXB6)
ENSG00000169247	SH3TC2	SH3 domain and tetratricopeptide repeats 2(SH3TC2)
ENSG00000120068	HOXB8	homeobox B8(HOXB8)
ENSG00000025293	PHF20	PHD finger protein 20(PHF20)
ENSG00000136997	MYC	v-myc avian myelocytomatosis viral oncogene homolog(MYC)
ENSG00000143882	ATP6V1C2	ATPase H+ transporting V1 subunit C2(ATP6V1C2)
ENSG00000003096	KLHL13	kelch like family member 13(KLHL13)
ENSG00000131746	TNS4	tensin 4(TNS4)
ENSG00000196532	HIST1H3C	histone cluster 1 H3 family member c(HIST1H3C)
ENSG00000233101	HOXB-AS3	HOXB cluster antisense RNA 3(HOXB-AS3)
ENSG00000204104	TRAF3IP1	TRAF3 interacting protein 1(TRAF3IP1)
ENSG00000126003	PLAGL2	PLAG1 like zinc finger 2(PLAGL2)
ENSG00000120875	DUSP4	dual specificity phosphatase 4(DUSP4)
ENSG00000164070	HSPA4L	heat shock protein family A (Hsp70) member 4 like(HSPA4L)
ENSG00000111057	KRT18	keratin 18(KRT18)
ENSG00000260807	LMF1	lipase maturation factor 1(LMF1)
ENSG00000174136	RGMB	repulsive guidance molecule family member b(RGMB)
ENSG00000197818	SLC9A8	solute carrier family 9 member A8(SLC9A8)
ENSG00000187372	PCDHB13	protocadherin beta 13(PCDHB13)
ENSG00000140526	ABHD2	abhydrolase domain containing 2(ABHD2)
ENSG00000166068	SPRED1	sprouty related EVH1 domain containing 1(SPRED1)
ENSG00000182742	HOXB4	homeobox B4(HOXB4)
ENSG00000101193	GID8	GID complex subunit 8 homolog(GID8)

We performed an explanatory analysis of the RNASeq data. Figure 1 shows the most meaningful changes at the 0.005 significance level among the genes between the two conditions, based on the volcano plot [54]. The volcano plot shows the genes with smaller p-values (higher $-log_{10}$ values) in red.

Figure 2 illustrates the estimated dispersion of the RNASeq data using *DESeq2* package, with each gene having a gene-specific dispersion parameter. Good estimates of dispersion parameters lead to accurate detection of differentially expressed genes. Underestimating the dispersion parameters might lead to false positives (i.e., declaring genes to be differentially expressed when they are not truly differentially-expressed). On the other hand, overestimating the dispersion parameters might lead to false negatives [55].

Table 3 - 6 show the performance of the gene-lists in predicting mutation status, based on seven methods (algorithms), at the 0.005 significance level: the 282 gene-list (Table 3); the 23 gene-list (Table 4); the 424 gene-list (Table 5); and the 401 gene-list (Table 6).

It is apparent from Table 3, compared to Table 4 below, that NB, ANN, KNN, and PLDA were improved in the common 23 genes in terms of all performance measures, while RF and NBLDA had the same performance. SVM had a better result on the full list of 282 genes. Therefore, in general, four methods out of seven were improved on the 23 gene-list compared to the 282 genes-list. From Figures 4 and 5, we notice NBLDA works very well in both lists of genes.

Table (3) **Performance of the classification methods for the 282 gene-list, on the RNASeq dataset ($\alpha = 0.005$).**

Methods / Metric	SVM	NB	RF	ANN	KNN	NBLDA	PLDA
Accuracy	0.80	0.76	0.83	0.72	0.72	0.89	0.80
Sensitivity	0.89	0.59	0.78	0.78	0.67	0.81	0.81
Specificity	0.70	0.93	0.89	0.67	0.78	0.96	0.78
Kappa	0.59	0.52	0.67	0.44	0.44	0.78	0.59
AUC	0.80	0.76	0.83	0.72	0.72	0.89	0.80
BER	0.20	0.24	0.17	0.28	0.28	0.11	0.20
MCC	0.60	0.50	0.67	0.45	0.45	0.79	0.59
F1-measure	0.81	0.71	0.82	0.74	0.71	0.88	0.80

Table (4) **Performance of the classification methods for the 23 gene-list, on the RNASeq dataset ($\alpha = 0.005$).**

Methods / Metric	SVM	NB	RF	ANN	KNN	NBLDA	PLDA
Accuracy	0.78	0.80	0.83	0.80	0.76	0.89	0.87
Sensitivity	0.81	0.70	0.78	0.81	0.70	0.85	0.85
Specificity	0.74	0.89	0.89	0.78	0.81	0.93	0.89
Kappa	0.56	0.59	0.67	0.59	0.52	0.78	0.74
AUC	0.78	0.80	0.83	0.80	0.76	0.89	0.87
BER	0.22	0.20	0.17	0.20	0.24	0.11	0.13
MCC	0.56	0.60	0.67	0.59	0.52	0.78	0.74
F1-measure	0.79	0.78	0.82	0.80	0.75	0.88	0.87

Table (5) **Performance of the classification methods for the 424 gene-list, on the combined RNASeq and microarray datasets ($\alpha = 0.005$).**

Methods / Metric	SVM	NB	RF	ANN	KNN	NBLDA	PLDA
Accuracy	0.98	0.93	0.93	0.98	0.83	-	-
Sensitivity	1	0.89	0.89	1	0.74	-	-
Specificity	0.96	0.96	0.96	0.96	0.93	-	-
Kappa	0.96	0.85	0.85	0.96	0.67	-	-
AUC	0.98	0.93	0.93	0.98	0.83	-	-
BER	0.02	0.07	0.07	0.02	0.17	-	-
MCC	0.96	0.85	0.85	0.96	0.68	-	-
F1-measure	0.98	0.92	0.92	0.98	0.82	-	-

Table 5 presents the integration results using the union approach, and it is clear that SVM, NB, RF, ANN, and KNN methods were improved compared to the case of 282 differentially-expressed genes. These results were confirmed in Figures 4 and 6. Moreover, SVM and ANN had a higher accuracy than the other methods.

Table (6) **Performance of the methods for the 401 gene-list, on the RNASeq dataset ($\alpha = 0.005$).**

Methods / Metric	SVM	NB	RF	ANN	KNN	NBLDA	PLDA
Accuracy	0.98	0.93	0.93	0.96	0.83	-	-
Sensitivity	1	0.89	0.89	0.96	0.74	-	-
Specificity	0.96	0.96	0.96	0.96	0.93	-	-
Kappa	0.96	0.85	0.85	0.93	0.67	-	-
AUC	0.98	0.93	0.93	0.96	0.83	-	-
BER	0.02	0.07	0.07	0.04	0.17	-	-
MCC	0.96	0.85	0.85	0.93	0.68	-	-
F1-measure	0.98	0.92	0.92	0.96	0.82	-	-

As can be seen from the Table 6 above, the methods performed better for the gene-list of 401 genes, compared to the 282 gene-list. Furthermore, Figures 7 confirm these results.

We compared our gene-list of 23 genes with the 18-gene RAS signature (*DUSP4, DUSP6, ELF1, ETV4, ETV5, FXYD5, KANK1, LGALS3, LZTS1, MAP2K3, PHLDA1, PROS1, S100A6, SERPINB1, SLCO4A, SPRY2, TRIB2*, and *ZFP106*) (Dry *et al.* [56]) and found only one overlapping gene (*DUSP4*). It turned out that this was also the most predictive of the 7 genes (*DUSP4, DUSP6, ETV4, ETV5, PHLDA1, SERPINB1*, and *TRIB2*) that were discussed in Omolo *et al.* (2016) [4].

Additional analysis was performed to assess whether the 23 gene-list was predictive of overall survival (OS). Survival analysis was performed using the mutation status as a group variable and vital status (dead or alive) as the censoring variable. Overall, 20 deaths were recorded out of the 54 samples. The results showed that the median OS was 1692 months for the 54 samples. Kaplan-Meier curves were used to graphically compare survival probabilities (Fig 8) between the two mutation groups (RAS-mutant vs wild-type). The log-rank test was performed using the RAS mutation status as the group variable. There was no significant difference in OS between the two groups (log-rank = 1.8, p-value = 0.2). The Cox proportional

hazards (CPH) model was then applied to assess the significance of the 23 genes and RAS mutation status. The results show that 9 out of the 23 genes were significantly associated with OS, including *SPRED1, KLHL13, HOXB4, LMF1, HSPA4L* at the 0.05 level, and *ATP6V1C2, PLAGL2, MYC, SLC9A8* at the 0.1 level (LRT = 56.85, p-value = 0.0002) as can be seen in Tables 7).

Further analysis was performed on the top 9 genes using gradient boosted trees and Shapley additive explanations (SHAP) methods to identify the top-K genes ($1 < K < 9$) [57]. The SHAP approach determined the order of importance of our 9 genes. SHAP values were calculated to give the importance of a gene by comparing what a model predicts with and without the gene. A SHAP value of 0 means that the gene has no effect on the prediction, as shown in Fig 9. The vertical axis showed the gene names, arranged in the order of importance, from top to bottom while the adjacent value next to the gene name is the mean SHAP value. The horizontal axis showed the SHAP value, which indicated how much the change was in log-odds. From the log-odds, one can obtain the probability of success. The gradient color indicated the original value for that gene. Genes pushing the prediction higher were coloured blue, while those pushing the prediction lower were coloured yellow. Each point represented a row from the original dataset.

Table (7) Cox proportional hazards model for overall survival, using the 23 genes and RAS mutation status (class) as covariates.

Covariate	Coef	Hazard ratio (HR)	SE(Coef)	Z	Pr($>\|z\|$)
class	2.23E+00	9.34E+00	1.41E+00	1.589	0.112
ATP6V1C2	4.72E-03	1.01E+00	2.66E-03	1.779	0.0752 .
HOXB-AS3	7.96E-05	1.00E+00	1.30E-03	0.061	0.951
KRT18	-3.27E-05	1.00E+00	8.73E-05	-0.374	0.7084
RGMB	7.48E-04	1.00E+00	1.13E-03	0.662	0.5079
PLAGL2	-3.28E-03	9.97E-01	1.89E-03	-1.737	0.0824 .
DUSP4	2.12E-03	1.00E+00	2.24E-03	0.946	0.3441
SPRED1	1.50E-03	1.00E+00	7.61E-04	1.969	0.0489 *
SH3TC2	6.92E-04	1.00E+00	4.42E-04	1.564	0.1178
HOXB8	-1.04E-02	9.90E-01	7.21E-03	-1.444	0.1488
ABHD2	4.67E-04	1.00E+00	4.34E-04	1.078	0.2812
TNS4	-5.58E-04	9.99E-01	3.74E-04	-1.491	0.1358
HIST1H3C	-4.67E-03	9.95E-01	3.20E-03	-1.458	0.1449
KLHL13	8.28E-03	1.01E+00	3.24E-03	2.559	0.0105 *
MYC	1.21E-03	1.00E+00	7.16E-04	1.689	0.0911 .
HOXB4	9.10E-03	1.01E+00	3.60E-03	2.529	0.0114 *
HOXB6	-2.78E-04	1.00E+00	4.18E-03	-0.067	0.9469
PHF20	1.47E-03	1.00E+00	1.78E-03	0.825	0.4094
LMF1	3.19E-03	1.00E+00	1.43E-03	2.224	0.0262 *
SLC9A8	-4.98E-03	9.95E-01	2.56E-03	-1.946	0.0517 .
GID8	2.16E-03	1.00E+00	2.61E-03	0.828	0.4079
HSPA4L	-6.94E-03	9.93E-01	2.78E-03	-2.497	0.0125 *
PCDHB13	-3.90E-03	9.96E-01	4.09E-03	-0.953	0.3406
TRAF3IP1	-7.63E-03	9.92E-01	4.86E-03	-1.571	0.1163

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Discussion

The development of molecular signatures is a significant step towards understanding the molecular mechanisms of tumorigenesis, which could help with accurate prognosis and diagnosis and thus allow physicians to prescribe suitable patient-specific therapies.

Several studies have done cancer classification using either microarray or RNASeq data only, and few studies have shown integration of both types of data, based on

heterogeneous datasets. To the best of our knowledge, no cancer classification study has employed the integration of homogeneous datasets approach. In this study, we integrated homogeneous microarray and RNASeq datasets, and assessed whether such an approach could improve the classification accuracy using seven methods, namely, SVM with radial basis function kernel, NB, RF, ANN, KNN, NBLDA, and PLDA. We implemented the classification of the mutation status of CRC samples, using gene-lists obtained through the intersection, complement of intersection, and union of differentially-expressed genes from microarray and RNASeq datasets.

CRC is the third most common cancer and one of the leading causes of death around the world. The findings suggest that combining two homogeneous datasets from different technologies could lead to an increase in CRC classification accuracy. The study also corroborates the findings of Castillo *et al.* [20], which found that combining heterogeneous datasets from different platforms can improve the performance of a classifier, using multiple datasets. They used data from different technologies and platforms in order to obtain a larger sample size due to the lack of enough RNASeq samples. Our proposed approach is different from Castillo *et al.*(2019) in that we used homogeneous datasets and a balanced binary class problem. We used the 0.005 significance level to obtained the differentially expressed genes, which is restrictive enough to control the false positive rate.

A comparison of the performance of the classification methods for each gene-list revealed that SVM yielded the highest mean accuracy (0.885), followed by RF(0.880), ANN(0.865), NB(0.855), and KNN(0.785) across the four gene-lists. However, NBLDA performed better than PLDA as a classifier when the analysis was restricted to RNASeq (count) data. Castillo *et al.* [20] also showed that SVM performed second to KNN. Statnikov *et al.* [58] performed a comparison of 18 classification methods on 5 feature selection methods, using 8 datasets and showed that RF had the highest accuracy (0.954).

Survival analysis results showed that 9 out of the 23 genes were prognostic for overall survival for CRC patients. Upon subjecting the 9 genes to the Shapley additive explanations (SHAP) method to rank the genes in order of importance, the top-5 genes to emerge were *ATP6V1C2, MYC, LMF1, HSPA4L,* and *PLAGL2*.

Our findings were validated with other published molecular signatures from previous studies (e.g. Zumwalt *et al.* [59], He *et al.* [60], and Liu *et al.* [61]). Zumwalt *et al.* [59] showed that ***ATP6V1C2*** expression successfully distinguished between cancerous and non-cancerous samples in CRC. In addition, He *et al.* [60] reported that the expression of ***c-Myc***, which was one of the three related human genes encoded under ***MYC*** genes family, was observed in many human cancers and was elevated in up to 70 - 80 % in CRC. Liu *et al.* [61] identified 10 lncRNAs related with crucial outcomes in CRC and one of these 10 genes was ***LMF1***. Zhang *et al.* [62] obtained 34 genes using minimal redundancy maximal relevance (mRMR) and incremental feature selection (IFS) methods and found that the ***HSPA4L*** gene was the most highly expressed in CRC patients with chromosomal instability (CIN) mechanism. Zheng *et al.* [63] reported that the ***PLAGL2*** gene was vital in increasing the effect on glioblastoma and colorectal cancer. Su *et al.* [64] reported that ***PLAGL2*** served as an oncogenic function in multiple human malignancies, including colorectal cancer (CRC).

This study was limited by the number of homogeneous RNASeq and microarray datasets that were available. Only one matched-pair set of 54 CRC samples was analyzed. Future studies should extend the approach to more than one cancer type and multiple datasets. However, the number of samples in each dataset ($n = 54$) ensured that the training and validation sets were large enough for the magnitude and statistical significance of the classification accuracies.

Conclusions

In summary, data integration by taking the *intersection* of the individual gene-lists from the two data types, improved the classification accuracy of CRC. However, laboratory experiments should be conducted on this 23-gene signature to further assess its clinical significance in CRC research. NBLDA method was the best performer on the RNASeq data. Results suggest that the SVM method was the most suitable classifier for CRC across the two data types and had high accuracy before and after the integration. Future studies should determine the effectiveness of integration in cancer survival analysis and the application on unbalanced data (where the classes are of different sizes) as well as on data with multiple classes.

Abbreviations
CRC: Colorectal cancer
SSA: Sub-Saharan Africa
EGFRi: Anti-epithelial growth factor receptor inhibitor
FFPE: Formalin-fixed paraffin-embedded
AUC: Area under the ROC curve
WHO: World Health Organization
NGS: Next-generation sequencing
PLDA: Poison linear discriminant analysis
NBLDA: Negative binomial linear discriminant analysis
SVM: Support vector machines
ANN: Artificial neural networks
LDA: Linear discriminant analysis
RF: Random forests
NB: Naive Bayes
KNN: k-Nearest neighbors
IG: Information gain
SGA: Standard genetic algorithm
CFS: Correlation-based feature selection
IBPSO: Improved-binary particle swarm optimization
bagSVM: Bagging SVM
CART: Classification and regression trees
mRMR: Minimum-redundancy maximum-relevance
GEO: Gene expression omnibus
CPM: Counts per million
GLM: Generalized linear model
DAVID: Database for annotation, visualization, and integrated discovery
COSMIC: Catalogue of somatic mutations in cancer
MLE: Maximum likelihood estimate
RBF: Radial basis function
MAP: Maximum a posteriori
BER: Balanced error rate
MCC: Matthews correlation coefficient
TP: True Positive
TN: True Negative
FP: False Positive
FN: False Negative
RA: Random Accuracy
OS: overall survival
CPH: Cox proportional hazards
SHAP: Shapley additive exPlanations
IFS: Incremental feature selection
CIN: Chromosomal instability
HR: Hazard ratio

DECLARATIONS

Ethics approval and consent to participate
Not Applicable

Consent for publication
Not Applicable

Availability of data and materials
The datasets supporting the results of this article have been deposited to the public repository (GEO) under the series accession number GSE86566. The individual datasets are accessible under the accession number GSE86559 for the microarray data and GSE86562 for the RNASeq data. The mutation and overall survival data are available upon request from BO.

Competing interests
The authors declare that they have no competing interests.

Funding
MM was supported by the DELTAS Africa Initiative Grant No. 107754/Z/15/Z-DELTAS Africa SSACAB. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (Grant No. 107754/Z/15/Z) and the UK government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government.

Authors' contributions
BO conceived the study. MM performed all the analyses. MM and BO drafted the manuscript. MM, HM and BO proof-read, discussed and approved the final manuscript.

Acknowledgements
The authors wish to thank Prof. Bob Gagnon (GlaxoSmithKline (GSK)) for his comments on the manuscript.

Author details
[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, King Edward Avenue, Pietermaritzburg, South Africa. [2]Division of Mathematics & Computer Science, University of South Carolina-Upstate, 800 University Way, Spartanburg, USA. [3]School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, 1 Jan Smuts Avenue, Braamfontein, 2000 Johannesburg, South Africa.

References

1. Gandomani, H.S., Aghajani, M., Mohammadian-Hafshejani, A., Tarazoj, A.A., Pouyesh, V., Salehiniya, H., *et al.*: Colorectal cancer in the world: incidence, mortality and risk factors. Biomedical Research and Therapy **4**(10), 1656–1675 (2017)
2. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2019. CA: a cancer journal for clinicians **69**(1), 7–34 (2019)
3. cancer research fund, W.: Colorectal cancer statistics. https://www.wcrf.org/dietandcancer/cancer-trends/colorectal-cancer-statistics. [Online; accessed 02-Aug-2019] (2019)
4. Omolo, B., Yang, M., Lo, F.Y., Schell, M.J., Austin, S., Howard, K., Madan, A., Yeatman, T.J.: Adaptation of a ras pathway activation signature from ff to ffpe tissues in colorectal cancer. BMC medical genomics **9**(1), 65 (2016)
5. Mármol, I., Sánchez-de-Diego, C., Pradilla Dieste, A., Cerrada, E., Rodriguez Yoldi, M.: Colorectal carcinoma: a general overview and future perspectives in colorectal cancer. International journal of molecular sciences **18**(1), 197 (2017)
6. Granados-Romero, J.J., Valderrama-Treviño, A.I., Contreras-Flores, E.H., Barrera-Mera, B., Herrera Enríquez, M., Uriarte-Ruíz, K., Ceballos-Villalba, J., Estrada-Mata, A.G., Alvarado Rodríguez, C., Arauz-Peña, G.: Colorectal cancer: a review. Int J Res Med Sci **5**(11), 4667 (2017)
7. (WHO), W.H.O.: Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer Accessed 2019-08-14
8. DeSantis, C.E., Miller, K.D., Goding Sauer, A., Jemal, A., Siegel, R.L.: Cancer statistics for african americans, 2019. CA: a cancer journal for clinicians **69**(3), 211–233 (2019)
9. Society, A.C.: Cancer Facts and Figures
10. May, F.P., Anandasabapathy, S.: Colon cancer in africa: Primetime for screening? Gastrointestinal endoscopy **89**(6), 1238–1240 (2019)
11. Organization, W.H.: National Cancer Control Programmes: Policies and Managerial Guidelines. World Health Organization, ??? (2002)
12. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., *et al.*: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science **286**(5439), 531–537 (1999)
13. Wang, J., Tan, A.C., Tian, T.: Next Generation Microarray Bioinformatics: Methods and Protocols. Springer, ??? (2012)

14. Tan, A.C., Gilbert, D.: Ensemble machine learning on gene expression data for cancer classification (2003)
15. Vanitha, C.D.A., Devaraj, D., Venkatesulu, M.: Gene expression data classification using support vector machine and mutual information-based gene selection. procedia computer science **47**, 13–21 (2015)
16. Lusa, L., *et al.*: Class prediction for high-dimensional class-imbalanced data. BMC bioinformatics **11**(1), 523 (2010)
17. Rajeswari, P., Reena, G.S.: Human liver cancer classification using microarray gene expression data. International Journal of Computer Applications **34**(6), 0975–8887 (2011)
18. Datta, S., Nettleton, D.: Statistical Analysis of Next Generation Sequencing Data vol. 15. Springer, ??? (2014)
19. Rai, M.F., Tycksen, E.D., Sandell, L.J., Brophy, R.H.: Advantages of rna-seq compared to rna microarrays for transcriptome profiling of anterior cruciate ligament tears. Journal of Orthopaedic Research® **36**(1), 484–497 (2018)
20. Castillo, D., Galvez, J.M., Herrera, L.J., Rojas, F., Valenzuela, O., Caba, O., Prados, J., Rojas, I.: Leukemia multiclass assessment and classification from microarray and rna-seq technologies integration at gene expression level. PloS one **14**(2), 0212127 (2019)
21. Zhang, W., Yu, Y., Hertwig, F., Thierry-Mieg, J., Zhang, W., Thierry-Mieg, D., Wang, J., Furlanello, C., Devanarayan, V., Cheng, J., *et al.*: Comparison of rna-seq and microarray-based models for clinical endpoint prediction. Genome biology **16**(1), 133 (2015)
22. Dong, K., Zhao, H., Tong, T., Wan, X.: Nblda: negative binomial linear discriminant analysis for rna-seq data. BMC bioinformatics **17**(1), 369 (2016)
23. Zararsız, G., Goksuluk, D., Korkmaz, S., Eldem, V., Zararsiz, G.E., Duru, I.P., Ozturk, A.: A comprehensive simulation study on classification of rna-seq data. PloS one **12**(8), 0182507 (2017)
24. Aziz, R., Verma, C., Srivastava, N.: Artificial neural network classification of high dimensional data with novel optimization approach of dimension reduction. Annals of Data Science **5**(4), 615–635 (2018)
25. Salem, H., Attiya, G., El-Fishawy, N.: Classification of human cancer diseases by gene expression profiles. Applied Soft Computing **50**, 124–134 (2017)
26. Jain, I., Jain, V.K., Jain, R.: Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Applied Soft Computing **62**, 203–215 (2018)
27. Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome biology **11**(10), 106 (2010)
28. Taminau, J., Lazar, C., Meganck, S., Nowé, A.: Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis. ISRN bioinformatics **2014** (2014)
29. Chen, Y., McCarthy, D., Robinson, M., Smyth, G.: edger: differential expression analysis of digital gene expression data. Bioconductor User's Guide, 1–78 (2014)
30. Castillo, D., Gálvez, J.M., Herrera, L.J., San Román, B., Rojas, F., Rojas, I.: Integration of rna-seq data with heterogeneous microarray data for breast cancer profiling. BMC bioinformatics **18**(1), 506 (2017)
31. Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., Tegnér, J.: Data integration in the era of omics: current and future challenges. BioMed Central (2014)
32. Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C., *et al.*: David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic acids research **35**(suppl_2), 169–175 (2007)
33. Bolstad, B.M., Irizarry, R.A., Åstrand, M., Speed, T.P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics **19**(2), 185–193 (2003)
34. Witten, D.M., *et al.*: Classification and clustering of sequencing data using a poisson model. The Annals of Applied Statistics **5**(4), 2493–2518 (2011)
35. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152 (1992). ACM
36. Moguerza, J.M., Muñoz, A.: Support vector machines with applications. Statistical Science, 322–336 (2006)
37. Stephens, D., Diesing, M.: A comparison of supervised classification methods for the prediction of substrate type using multibeam acoustic and legacy grain-size data. PLOS One **9**(4), 93950 (2014)
38. Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C., Ares, M., Haussler, D.: Support vector machine classification of microarray gene expression data. University of California, Santa Cruz, Technical Report UCSC-CRL-99-09 (1999)
39. Chu, F., Wang, L.: Gene expression data analysis using support vector machines. In: Neural Networks, 2003. Proceedings of the International Joint Conference On, vol. 3, pp. 2268–2271 (2003). IEEE
40. Karatzoglou, A., Smola, A., Hornik, K., Karatzoglou, M.A.: Package 'kernlab' (2016)
41. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning vol. 1. Springer, ??? (2001)
42. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
43. Qi, Y.: Random forest for bioinformatics. In: Ensemble Machine Learning, pp. 307–323. Springer, ??? (2012)
44. Chen, X., Ishwaran, H.: Random forests for genomic data analysis. Genomics **99**(6), 323–329 (2012)
45. Pappu, V., Pardalos, P.M.: High-dimensional data classification. In: Clusters, Orders, and Trees: Methods and Applications, pp. 119–150. Springer, ??? (2014)
46. Do, T.-N., Lenca, P., Lallich, S., Pham, N.-K.: Classifying very-high-dimensional data with random forests of oblique decision trees. In: EGC (best of Volume), pp. 39–55 (2009). Springer
47. Breiman, L., Cutler, A., Liaw, A., Wiener, M.: Package'randomforest'. software available at URL: http://stat-www. berkeley. edu/users/breiman/RandomForests (2011)
48. Dwivedi, A.K.: Artificial neural network model for effective cancer classification using microarray gene expression data. Neural Computing and Applications, 1–10 (2016)
49. Ripley, B., Venables, W., Ripley, M.B.: Package 'nnet'. R Package Version, 7–3 (2016)
50. De Campos, L.M., Cano, A., Castellano, J.G., Moral, S.: Bayesian networks classifiers for gene-expression data. In: Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference On, pp. 1200–1206 (2011). IEEE

51. Yao, Z., Ruzzo, W.L.: A regression-based k nearest neighbor algorithm for gene function prediction from heterogeneous data. BMC Bioinformatics **7**(1), 11 (2006)
52. Tharwat, A.: Classification assessment methods. Applied Computing and Informatics (2018)
53. Mohammed, M., Mwambi, H., Omolo, B., Elbashir, M.K.: Using stacking ensemble for microarray-based cancer classification. In: 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), pp. 1–8 (2018). IEEE
54. Li, W.: Volcano plots in analyzing differential expressions with mrna microarrays. Journal of bioinformatics and computational biology **10**(06), 1231003 (2012)
55. Landau, W.M., Liu, P.: Dispersion estimation and its effect on test performance in rna-seq data analysis: a simulation-based comparison of methods. PloS one **8**(12), 81415 (2013)
56. Dry, J.R., Pavey, S., Pratilas, C.A., Harbron, C., Runswick, S., Hodgson, D., Chresta, C., McCormack, R., Byrne, N., Cockerill, M., *et al.*: Transcriptional pathway signatures predict mek addiction and response to selumetinib (azd6244). Cancer research **70**(6), 2264–2273 (2010)
57. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, pp. 4765–4774 (2017)
58. Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., Pei, Z., Blaser, M.J., Aliferis, C.F., Alekseyenko, A.V.: A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome **1**(1), 11 (2013)
59. Zumwalt, T.J., Shigeyasu, K., Weng, W., Okugawa, Y., Miyoshi, J., Goel, A.: The ATP6V1C2 (a vacuolar-ATPase gene) is a novel early prognosticator for colorectal cancer. AACR (2016)
60. He, W., Weng, X.-T., Wang, J.-L., Lin, Y.-K., Liu, T.-W., Zhou, Q.-Y., Hu, Y., Pan, Y., Chen, X.-L.: Association between c-myc and colorectal cancer prognosis: a meta-analysis. Frontiers in Physiology **9**, 1549 (2018)
61. Liu, H., Gu, X., Wang, G., Huang, Y., Ju, S., Huang, J., Wang, X.: Copy number variations primed lncrnas deregulation contribute to poor prognosis in colorectal cancer. Aging (Albany NY) **11**(16), 6089 (2019)
62. Zhang, T.-M., Huang, T., Wang, R.-F.: Cross talk of chromosome instability, cpg island methylator phenotype and mismatch repair in colorectal cancer. Oncology letters **16**(2), 1736–1746 (2018)
63. Zheng, H., Ying, H., Wiedemeyer, R., Yan, H., Quayle, S.N., Ivanova, E.V., Paik, J.-H., Zhang, H., Xiao, Y., Perry, S.R., *et al.*: Plagl2 regulates wnt signaling to impede differentiation in neural stem cells and gliomas. Cancer cell **17**(5), 497–509 (2010)
64. Su, C., Li, D., Li, N., Du, Y., Yang, C., Bai, Y., Lin, C., Li, X., Zhang, Y.: Studying the mechanism of plagl2 overexpression and its carcinogenic characteristics based on 3'-untranslated region in colorectal cancer. International journal of oncology **52**(5), 1479–1490 (2018)

Figure Legend

Figure (1) Volcano plot of the RNASeq dataset show the 282 differentially expressed genes in red points ($\alpha = 0.005$).

Figure (2) Dispersion for the RNASeq data.

Figure (3) Flow-chart of the analysis.

Figure (4) ROC curves based on the 282 gene-list for the RNASeq data ($\alpha = 0.005$).

Figure (5) ROC curves based on the 23 gene-list for the RNASeq data ($\alpha = 0.005$).

Figure (6) ROC curves based on the 424 gene-list for the RNASeq and microarray datasets ($\alpha = 0.005$).

Figure (7) ROC curves based on the 401 gene-list ($\alpha = 0.005$).

Figure (8) Kaplan-Meier curves for overall survival (in months).

Figure (9) Genes in ascending order of importance (Note: dots represent SHAP values of specific features).

Figures

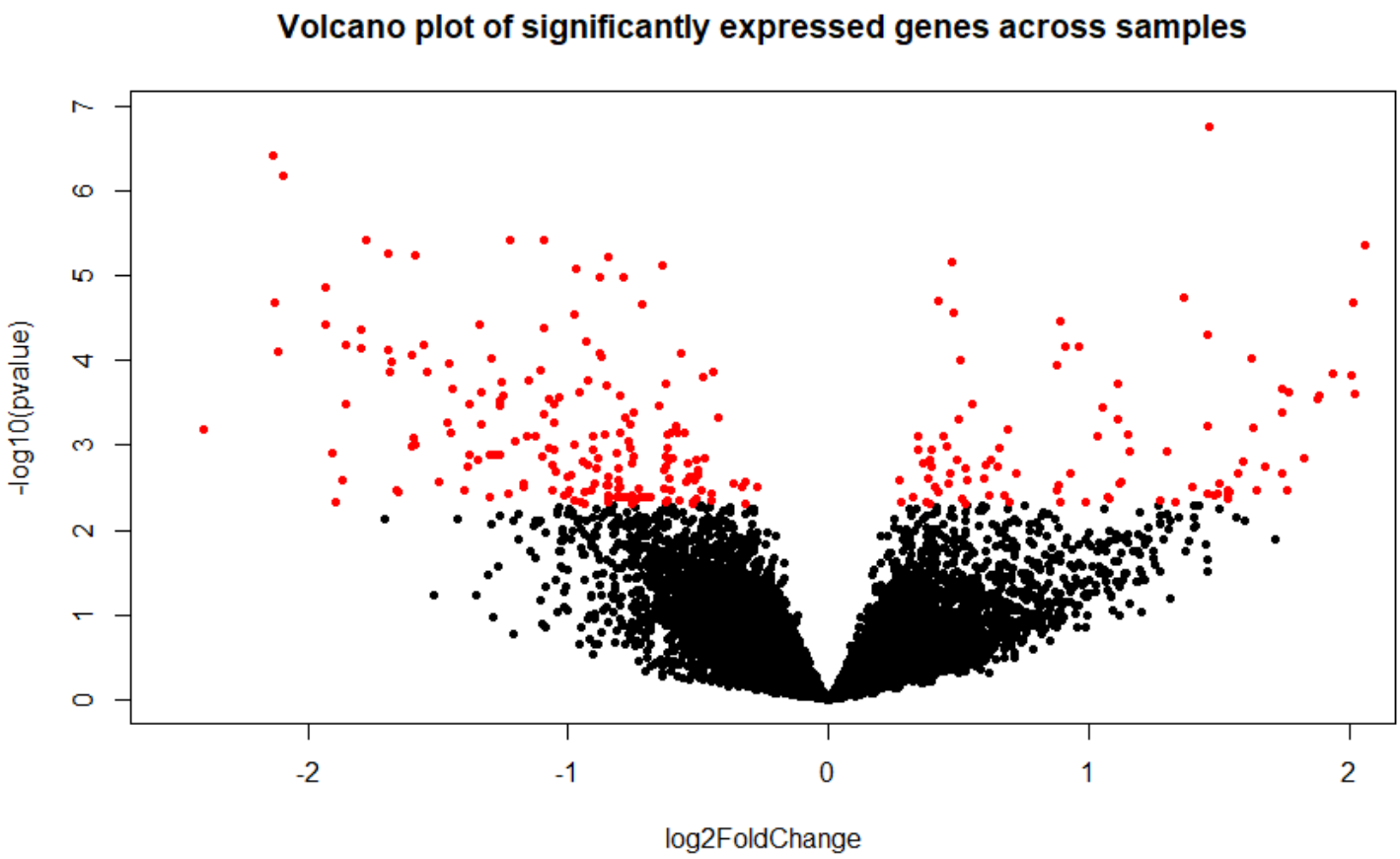

Figure 1

Volcano plot of the RNASeq dataset show the 282 differentially ex-pressed genes in red points (α = 0:005).

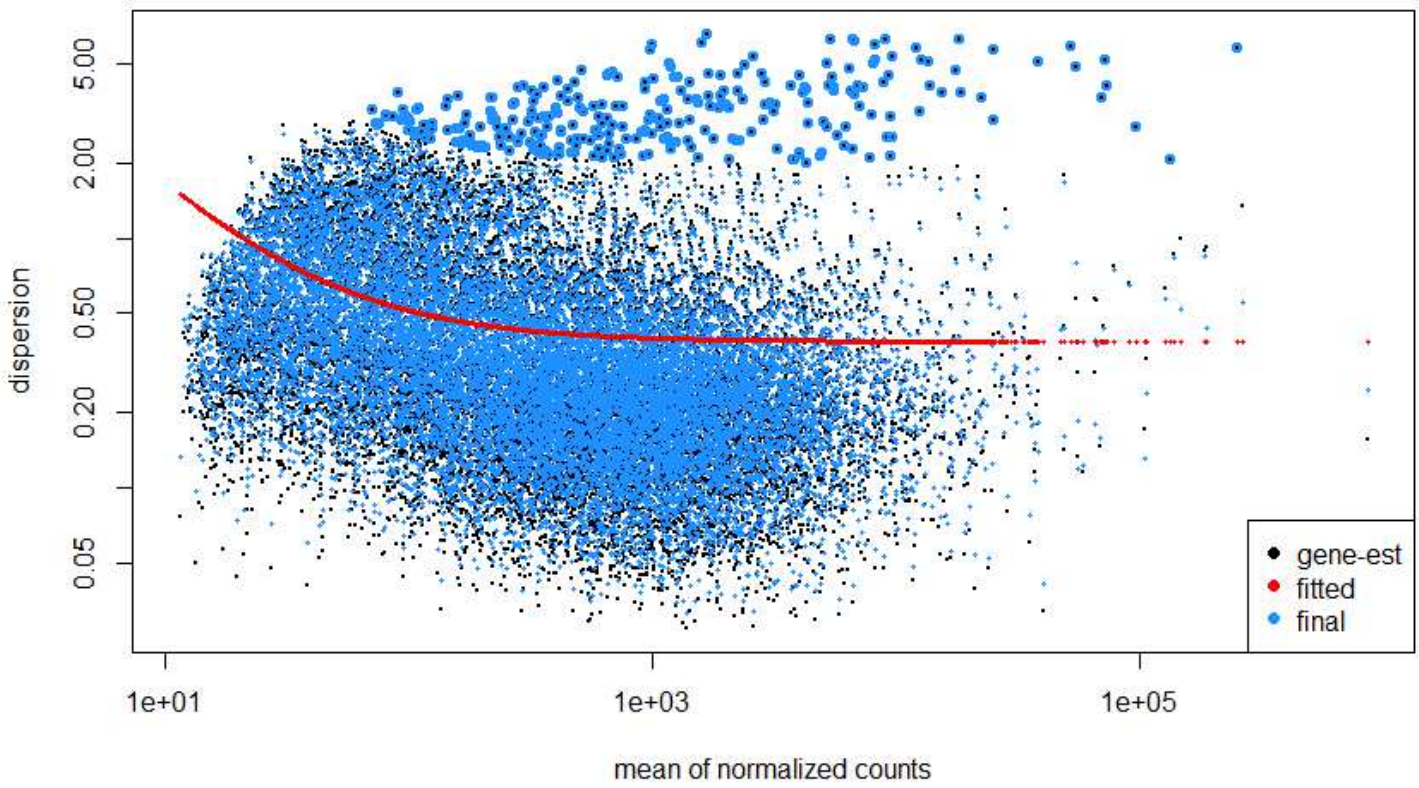

Figure 2

Dispersion for the RNASeq data.

Figure 3

Flow-chart of the analysis.

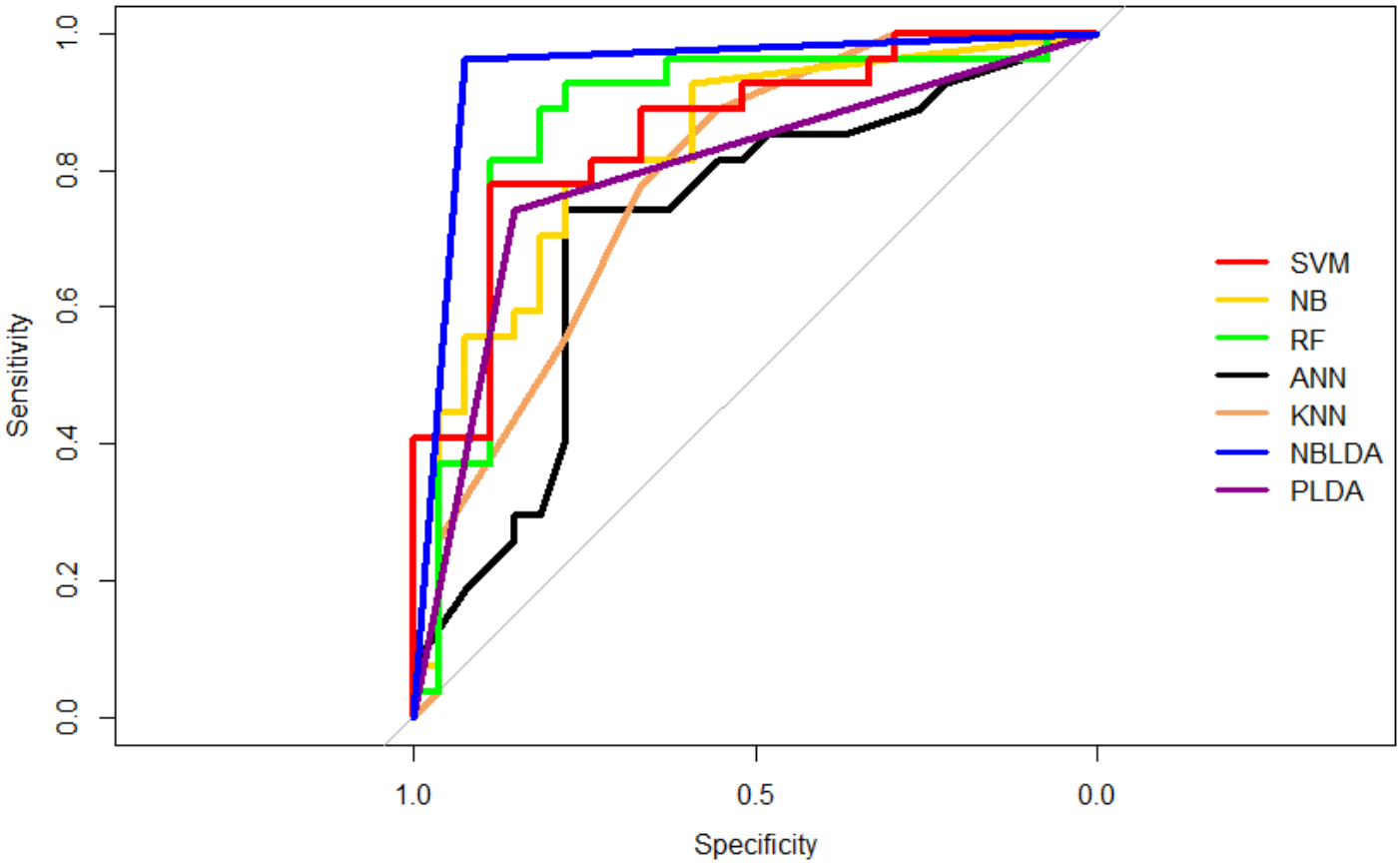

Figure 4

ROC curves based on the 282 gene-list for the RNASeq data (α =0:005).

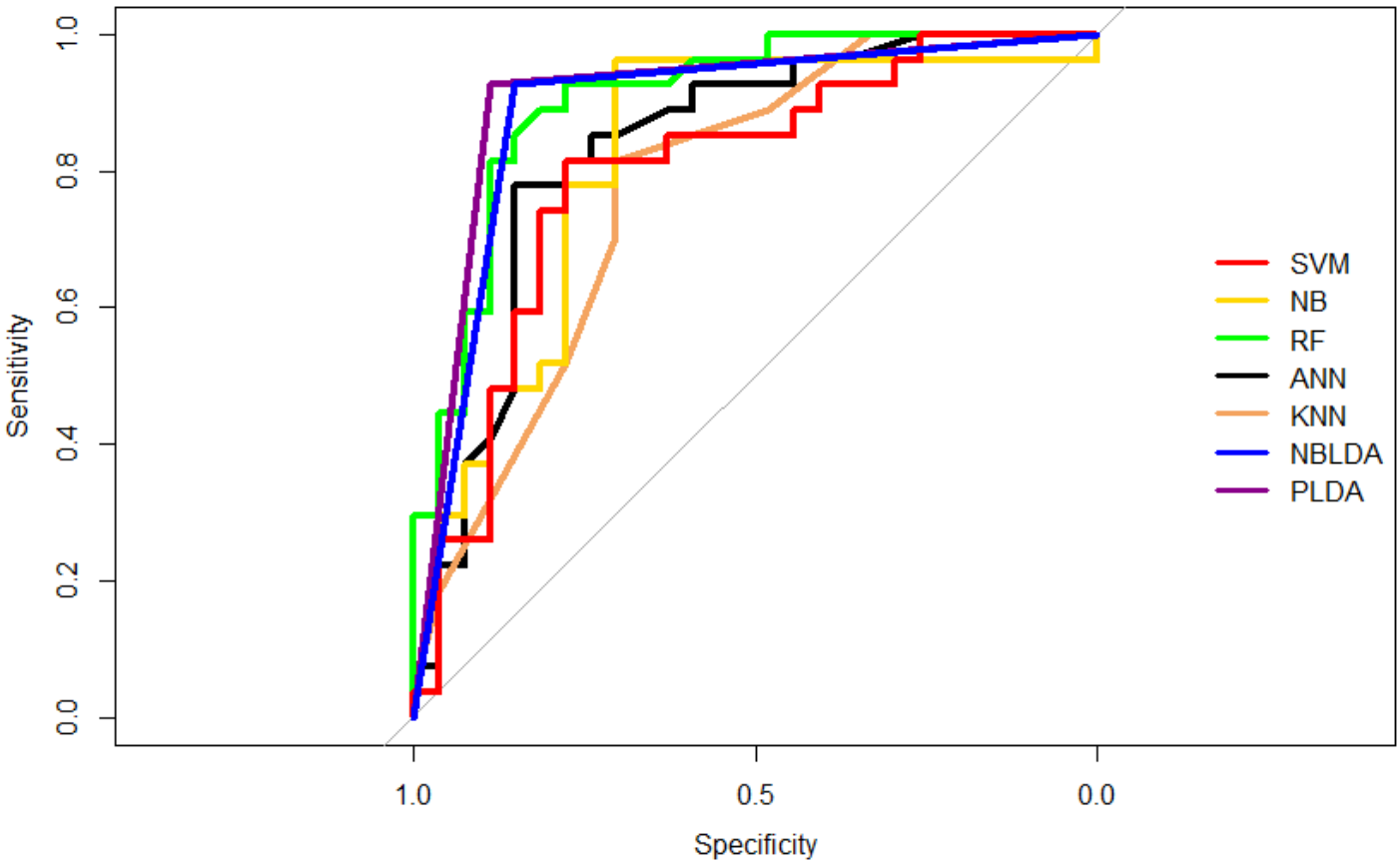

Figure 5

ROC curves based on the 23 gene-list for the RNASeq data (α = 0:005).

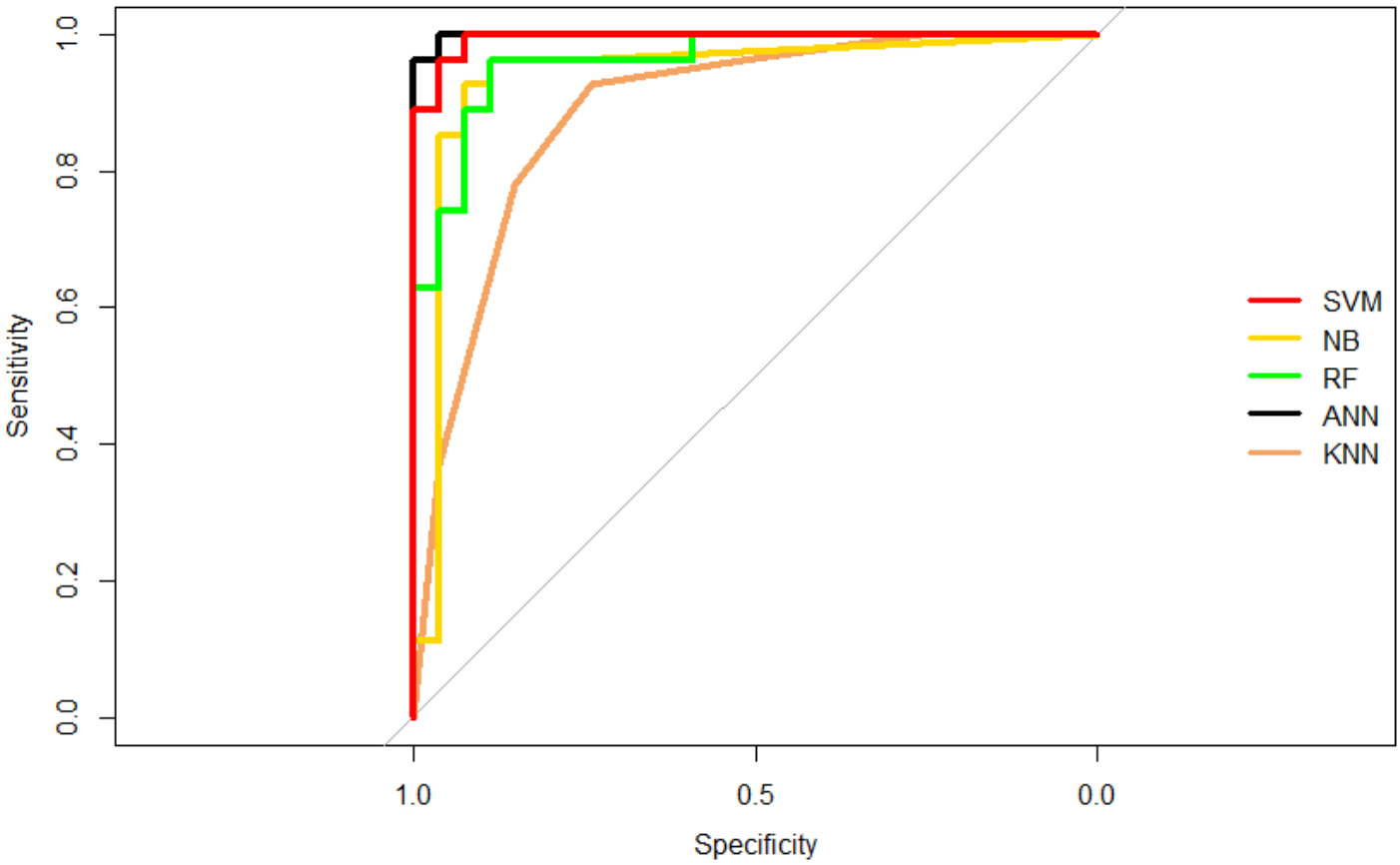

Figure 6

ROC curves based on the 424 gene-list for the RNASeq and microarray datasets (α = 0:005).

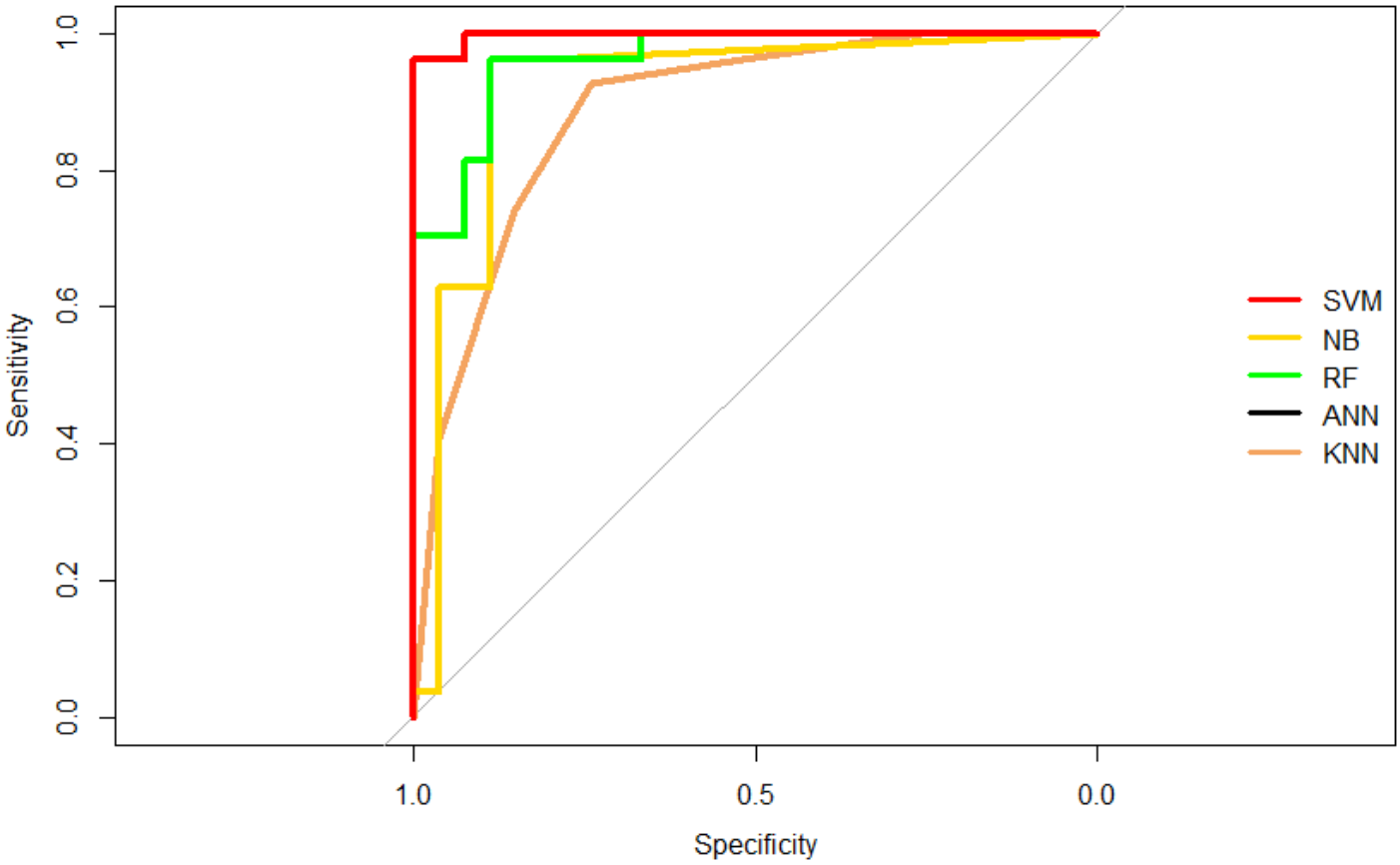

Figure 7

ROC curves based on the 401 gene-list (α = 0:005).

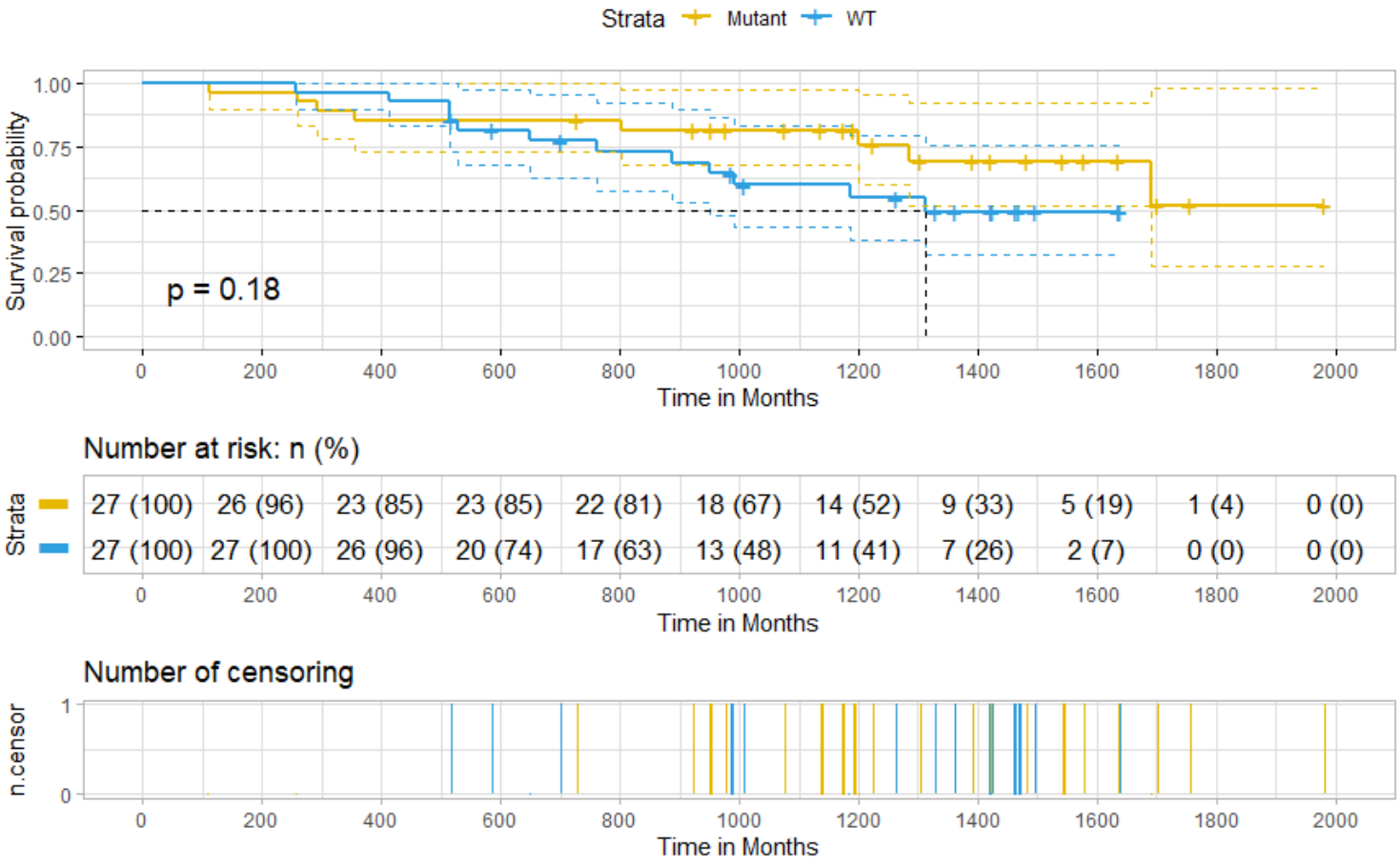

Figure 8

Kaplan-Meier curves for overall survival (in months).

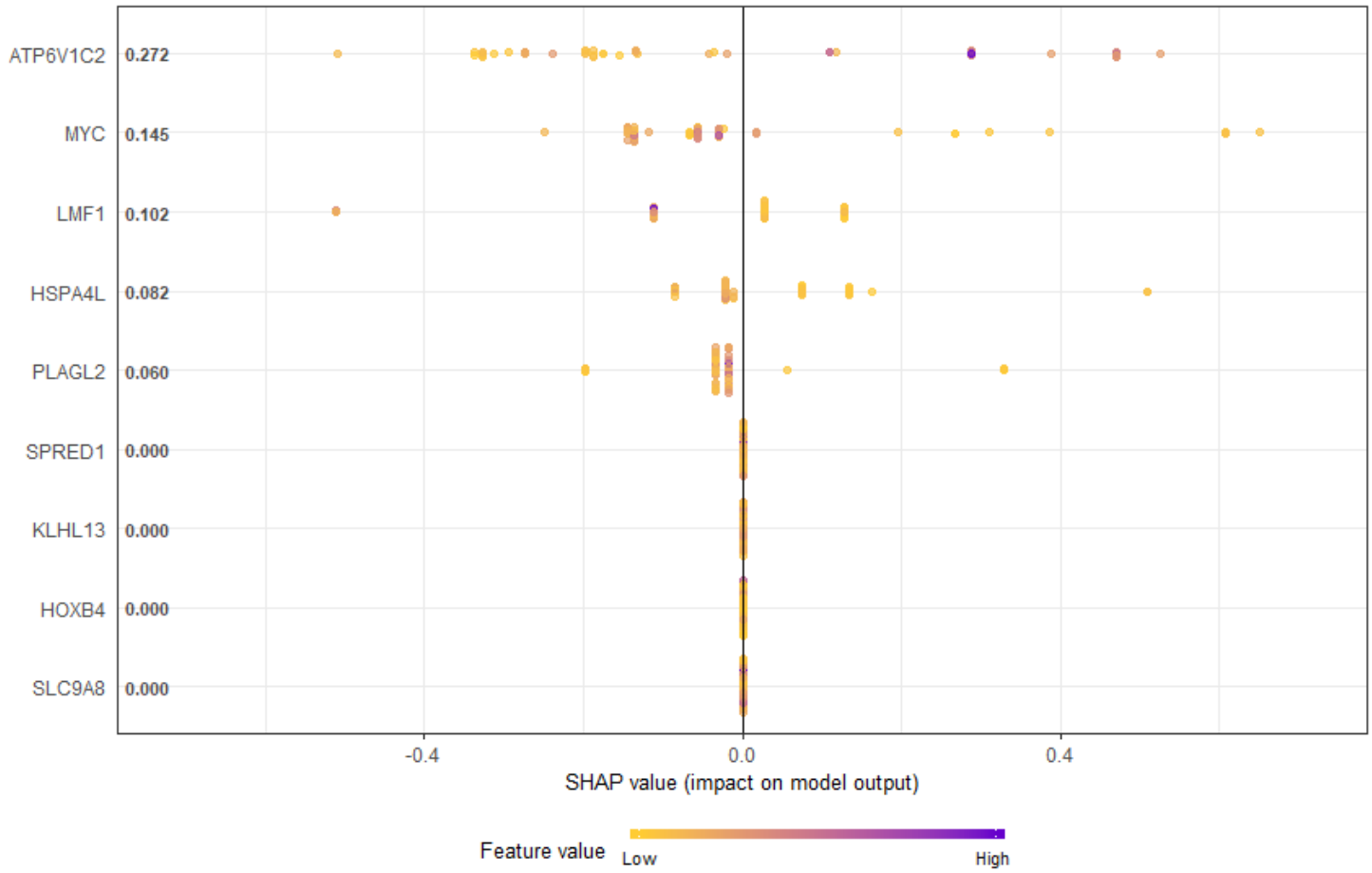

Figure 9

Genes in ascending order of importance (Note: dots represent SHAP values of specific features).