*Bioinformatics*, 32(12), 2016, 1814–1822
doi: 10.1093/bioinformatics/btw034
Advance Access Publication Date: 11 February 2016
Original Paper

Gene expression

# Selecting a classification function for class prediction with gene expression data

**Victor L. Jong[1,2,*], Putri W. Novianti[1,3], Kit C.B. Roes[1] and Marinus J.C. Eijkemans[1]**

[1]Biostatistics & Research Support, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, 3508 GA, Utrecht, The Netherlands, [2]Viroscience Lab, Erasmus Medical Center Rotterdam, Rotterdam, CE 3015, The Netherlands and [3]Epidemiology & Biostatistics Department, Vrije University Medical Center Amsterdam, HV Amsterdam 1081, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on 12 October 2015; revised on 18 December 2015; accepted on 15 January 2016

## Abstract

**Motivation:** Class predicting with gene expression is widely used to generate diagnostic and/or prognostic models. The literature reveals that classification functions perform differently across gene expression datasets. The question, which classification function should be used for a given dataset remains to be answered. In this study, a predictive model for choosing an optimal function for class prediction on a given dataset was devised.
**Results:** To achieve this, gene expression data were simulated for different values of gene-pairs correlations, sample size, genes' variances, deferentially expressed genes and fold changes. For each simulated dataset, ten classifiers were built and evaluated using ten classification functions. The resulting accuracies from 1152 different simulation scenarios by ten classification functions were then modeled using a linear mixed effects regression on the studied data characteristics, yielding a model that predicts the accuracy of the functions on a given data. An application of our model on eight real-life datasets showed positive correlations (0.33–0.82) between the predicted and expected accuracies.
**Conclusion:** The here presented predictive model might serve as a guide to choose an optimal classification function among the 10 studied functions, for any given gene expression data.
**Availability and implementation:** The R source code for the analysis and an R-package 'SPreFuGED' are available at *Bioinformatics* online.
**Contact:** v.l.jong@umcutecht.nl
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Microarray gene expression profiling has become a widely used tool to identify particular disease subpopulations and to perform diagnostic and prognostic predictions (Huang *et al.*, 2010; van 't Veer *et al.*, 2002). In clinical practice, they are used in diagnostic and prognostic analyses while in preclinical studies (toxicogenomics), they involve predicting the toxicity of compounds in animal models with the goal of speeding up the evaluation of toxicity for new drug candidates (Shi *et al.*, 2010). Though class prediction analysis is a common practice, the question that remains to be addressed is, given the wide availability of classification functions nowadays, which classification function do we use for a particular dataset? Classification functions have been shown to perform differently across gene expression datasets (Lee *et al.*, 2005). Moreover, the MAQC-II initiative has pointed out that classification function is one of the variables that explains the variability between gene expression class prediction performance (Shi *et al.*, 2010).

While substantial amount of information is known about the characteristics of classification functions and class prediction building procedures, little is known about which data characteristics have impact on the performance of a class prediction model. For instance, diagonal linear discriminant analysis (DLDA) assumes no covariances and hence no correlations between variables and might fail if the data is highly correlated. On the other hand, linear discriminant analysis (LDA) assumes a common covariance matrix for the classes and thus to some extent, accounts for correlations (Hastie *et al.*, 2003). In addition, pernalized regressions like ridge, lasso, elastic net are capable to handle correlated variables. Support Vector Machine (SVM), though commonly understood as a method of finding the maximum-margin hyperplane, may also be seen as a regularization function estimation problem, corresponding to a hinge loss function with a quadratic penalty as that of ridge regression (Hastie *et al.*, 2003; Ye *et al.*, 2011). And it has been shown by (Yang *et al.*, 2006) that if a group of non-distinct variables are selected as input variable set, its training time lengthened and the errors become bigger. On the other hand, tree-based methods are by nature designed to capture interactions between variables while neural networks might capture other complex structures within a given dataset.

Given the above observations, it is obvious that the performance of these functions depends on the characteristics of the data in question. Despite this, the literature on how to choose a classification function for a given dataset is sparse. A common practice is comparing several classification functions and selecting the one with the minimum error rate but this has been pointed by Bernau *et al.* (2013), Ding *et al.* (2014), Tibshirani and Tibshirani (2009) and Varma and Simon (2006) to lead to selection bias. As such, some experimenters adhere to one or a few classification functions irrespective of the dataset, disease or medical question being addressed. While others choose a classification function for their datasets by affinity or familiarity without taking into account the characteristics of such data.

A simulation study by Kim and Simon (2011) shows that correlation is one of the data characteristics that affect the performance of most probabilistic classification functions. In addition, Jong et al. (2014) showed that correlation structures differ across gene expression data of different etiological diseases. The study by Novianti *et al.* (2015) shows that microarray gene expression data characteristics like $\log_2$ fold change of expression values, number of deferentially expressed genes and pairwise correlations between genes are associated to the accuracy of classification functions. However, this study was conducted in real-life gene expression datasets, where the magnitude and/or direction of association might have been confounded by unobserved data characteristics.

In this study, we aim to provide a guideline for making a choice of a classification function for a binary class prediction problem based on observed magnitudes and directions of the data characteristics, using accuracy as a measure of evaluation. We investigate the effect of sample size, proportion of deferentially expressed (DE) genes, genes' variances, log fold changes, pairwise correlations between DE and noisy genes on the accuracy of classification functions using extensive simulations.

The remainder of this article is organized as follows: methodology to simulate data, classification functions considered and the building and evaluation of class prediction models are presented in Section 2; Section 3 contains a predictive summary of the results of class prediction models for different simulated scenarios; Section 4 provides an application of our predictive model from the simulated results on real-life microarray gene expression datasets and Section 5 presents a discussion.

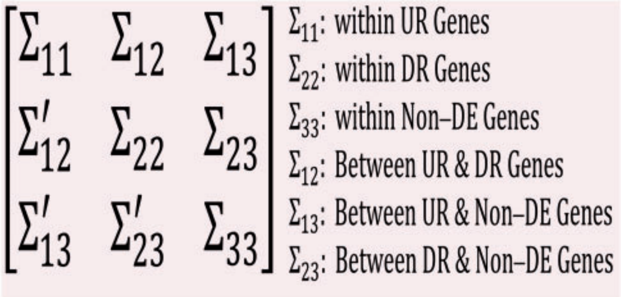
## 2 Methods

### 2.1 Simulated data (scenarios)

To simulated gene expression data, we hypothesized that sample size, proportion of DE genes, genes' variances, log fold changes, pairwise correlations between DE and noisy genes might be associated to the performance of classification functions. These six variables were to be systematically varied in our simulations.

From observed correlation structures in real-life gene expression datasets (Jong *et al.*, 2014), we generalized the structure as shown in Figure 1, containing three clusters referred to as up-regulated (UR), down-regulated (DR) and noisy genes. The absolute values of pairwise correlation for DE genes ($\rho$) were varied as 0.00, 0.25, 0.50 and 0.75 with UR cluster taking oppositely-signed correlation values for DR cluster. The pairwise correlations both within the noisy cluster and between the noisy and the DE clusters were per gene-pair randomly drawn from a normal distribution centered at zero with a standard deviation $\theta$ i.e. N(0, $\theta$) where $\theta = 0.00, 0.25, 0.50, 0.75$. The scenario $\rho = \theta = 0.00$ corresponds to complete independence. Resulting correlation values lying outside the interval $[-1, 1]$ were uniformly converted to the intervals $[-1, -0.15]$ and $[0.15, 1]$ for negative and positive values respectively. The variances of the genes $(\sigma^2 = \frac{1}{\lambda})$ were drawn from an exponential distribution i.e. exp($\lambda$) where $\lambda = 0.25,\ 0.50,\ 1.00,\ 1.50$. The distributional assumptions were made based on observation from real-life datasets as experienced by Jong *et al.* (2014) and Novianti *et al.* (2015). With the correlation values and the variances, the within covariance matrices $\Sigma_0$ and $\Sigma_1$ were constructed for the two classes. In addition, the proportion of DE genes ($\pi$) was also allowed to take up 1, 3 and 5% of the total number of genes, as values. This resulted to 192 different complex covariance matrices that were used to simulate the data for different values of other variables.

Finally, two different values of absolute $\log_2$ fold change ($\Delta$) and three different sample sizes ($n$) were considered (Table 1). For a fixed number of genes (p = 1000) and n samples, the samples' labels (0,1) were generated from a Bernoulli distribution with a probability 0.5 and the gene expression data of $p \times n$ dimension was generated from a multivariate normal distribution with mean vectors from a uniform distribution, U(6,10) of length $p$ and the covariance matrices corresponding to the above description, using Cholesky decomposition Golub and van Loan (1996) as a method to determine the root of the covariance matrix. The mean $\log_2$ expression values of DE genes were incremented or decremented with the corresponding $\log_2$ fold change value for samples in class 1. The choice of multivariate normal



$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma'_{12} & \Sigma_{22} & \Sigma_{23} \\ \Sigma'_{13} & \Sigma'_{23} & \Sigma_{33} \end{bmatrix}$$

$\Sigma_{11}$: within UR Genes
$\Sigma_{22}$: within DR Genes
$\Sigma_{33}$: within Non–DE Genes
$\Sigma_{12}$: Between UR & DR Genes
$\Sigma_{13}$: Between UR & Non–DE Genes
$\Sigma_{23}$: Between DR & Non–DE Genes

**Fig. 1.** Assumed correlation structure. Contains 3 clusters of up-regulated (UR), down-regulated (DR) and noisy (Non-DE) genes

distribution and mean vector corresponds to the practical assumption that gene expression data are normally distributed in $\log_2$ scale and based on observation that the $\log_2$ expression values often fall in the interval (0, 16). For each combination of the values of the data characteristics, the dataset was simulated as shown in Figure 2 (Algorithm 1), yielding 1152 different simulation scenarios, each of which was randomly replicated 1000 times.

## 2.2 Classification functions

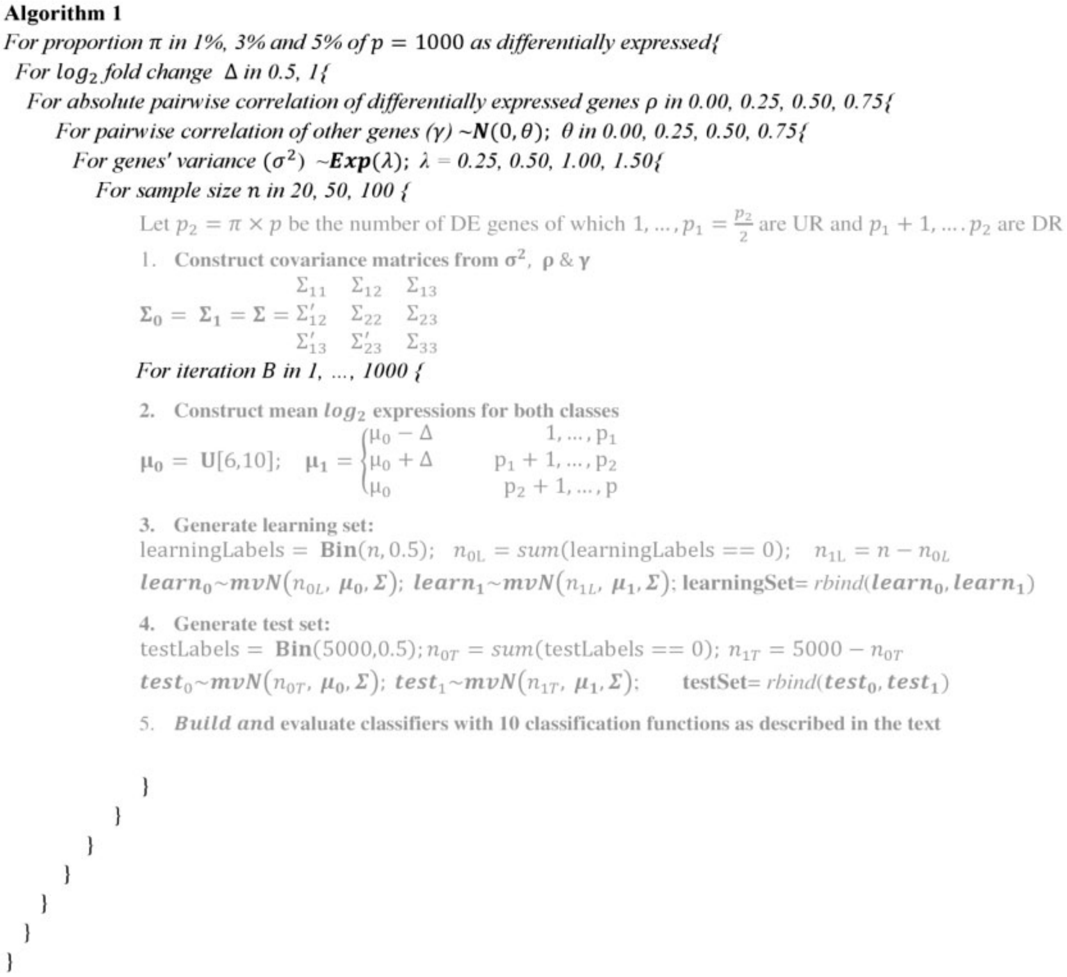
Ten elective choices of classification functions were chosen to represent the broad list in the literature that falls within the categories: discriminant analyses or Bayes classifiers, tree-based, regularization and shrinkage, nearest neighbors and neural networks methods. For discriminant analyses, linear discriminant analysis (LDA), quadratic discriminant analysis (McLachlan, 1992) and shrunken centroid discriminant analysis (SCDA) or prediction analysis of microarrays (PAM) (Tibshirani *et al.*, 2002) were selected. Random forest (RF) (Breiman, 2002) was chosen as tree-based method while support vector machines (SVM) (Schölkopf and Smola, 2002), $\ell_1$ penalized logistic regression or Lasso (PLR1) (Tibshirani, 1996), $\ell_2$ penalized logistic regression or Ridge (PLR2) (Zhu, 2004) as well as $\ell_1$ and $\ell_2$ penalized logistic regression or Elastic net (PLR12) (Zou and Trevo, 2005) were considered for regularized and shrinkage methods. Finally, k-nearest neighbors (KNN) and feed-forward neural network (NNET) (Ripley, 1996) were the lone choices for nearest neighbors and neural networks respectively.

**Table 1.** Simulated gene expression data characteristics

| Data characteristics | Values |
|---|---|
| Sample size ($n$) | 20, 50, 100 |
| Proportion of DE genes ($\pi$) | 1%, 3%, 5% |
| $\text{Log}_2$ fold change of DE genes ($\Delta$) | 0.5, 1 |
| Pairwise correlations of DE genes ($\rho$) | 0, 0.25, 0.5, 0.75 |
| Gene' variances ($\sigma^2 = 1/\lambda$) $\sim\text{Exp}(\lambda)$ | $\lambda = 0.25, 0.50, 1, 1.5$ |
| Pairwise correlations of noisy genes ($\gamma$) $\sim N(0,\theta)$ | $\theta = 0, 0.25, 0.5, 0.75$ |

50% of $\pi$ were each up- and down-regulated.

In machine learning, opinions are that super learners might provide good class predictions but model complexities of these learners are usually high. As such, super learners might not be useful in clinical practice where physicians often want simple class prediction models, that might yield a subset of genes (and possibly coefficients) for easy interpretation. This is because given a subset of genes, focus can be geared toward these genes rather than the entire genome for which experiments are often costly and time consuming. Thus, our choices of classification functions were driven by the choices often made and considered useful in clinical practice.



```
Algorithm 1
For proportion $\pi$ in 1%, 3% and 5% of $p = 1000$ as differentially expressed{
 For $log_2$ fold change $\Delta$ in 0.5, 1{
  For absolute pairwise correlation of differentially expressed genes $\rho$ in 0.00, 0.25, 0.50, 0.75{
   For pairwise correlation of other genes $(\gamma) \sim \mathbf{N}(0,\theta)$; $\theta$ in 0.00, 0.25, 0.50, 0.75{
    For genes' variance $(\sigma^2) \sim \mathbf{Exp}(\lambda)$; $\lambda = 0.25, 0.50, 1.00, 1.50${
     For sample size n in 20, 50, 100 {
      Let $p_2 = \pi \times p$ be the number of DE genes of which $1, \dots, p_1 = \frac{p_2}{2}$ are UR and $p_1 + 1, \dots. p_2$ are DR
      1. Construct covariance matrices from $\sigma^2$, $\rho$ & $\gamma$
      $\Sigma_0 = \Sigma_1 = \Sigma = \begin{matrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma'_{12} & \Sigma_{22} & \Sigma_{23} \\ \Sigma'_{13} & \Sigma'_{23} & \Sigma_{33} \end{matrix}$
      For iteration B in 1, ..., 1000 {
      2. Construct mean $log_2$ expressions for both classes
      $\mu_0 = \mathbf{U}[6,10]; \quad \mu_1 = \begin{cases} \mu_0 - \Delta & 1, \dots, p_1 \\ \mu_0 + \Delta & p_1 + 1, \dots, p_2 \\ \mu_0 & p_2 + 1, \dots, p \end{cases}$
      3. Generate learning set:
      learningLabels = $\mathbf{Bin}(n, 0.5)$;  $n_{0L} = sum$(learningLabels == 0);  $n_{1L} = n - n_{0L}$
      $\boldsymbol{learn}_0 \sim \boldsymbol{mvN}(n_{0L}, \mu_0, \Sigma)$; $\boldsymbol{learn}_1 \sim \boldsymbol{mvN}(n_{1L}, \mu_1, \Sigma)$; learningSet= rbind($\boldsymbol{learn}_0, \boldsymbol{learn}_1$)
      4. Generate test set:
      testLabels = $\mathbf{Bin}(5000,0.5)$; $n_{0T} = sum$(testLabels == 0); $n_{1T} = 5000 - n_{0T}$
      $\boldsymbol{test}_0 \sim \boldsymbol{mvN}(n_{0T}, \mu_0, \Sigma)$; $\boldsymbol{test}_1 \sim \boldsymbol{mvN}(n_{1T}, \mu_1, \Sigma)$;     testSet= rbind($\boldsymbol{test}_0, \boldsymbol{test}_1$)
      5. Build and evaluate classifiers with 10 classification functions as described in the text

      }
     }
    }
   }
  }
 }
}
```

**Fig. 2.** Algorithm to simulate data, build and validate class prediction models. For each value of the six variables, the covariance matrix was constructed in step 1, the learning and test data were simulated at steps 2–4 and class prediction models were built and validated in step 5. Steps 2–5 were then repeated 1000 times

### 2.3 Building and evaluating classifiers

To assess the dependency of the chosen classification functions on characteristics of the simulated gene expression data, we built on each simulated dataset, class prediction models with all the classification functions listed above. The simulated dataset was considered as a learning set and for classifiers that require pre-selection of genes because of their limitation to accommodate a number of parameters greater than the number of samples (i.e. LDA, QDA and NNET), the genes were ranked by their moderated t statistics (Smyth, 2004) using the learning set. The learning set was split into a $\frac{1}{3}$ inner- test set and $\frac{2}{3}$ learning set using 5-fold Monte-Carlo-cross-validation (MCCV) with stratification .

The parameter(s) of the classification functions were subsequently tuned using the inner-learning set and evaluated with the inner-test set. These tuning parameters were: number of genes (top *k*) for LDA and QDA; shrinkage intensity of class centroids for SCDA; with a fixed forest size of 500 trees, the number of variables randomly sampled as candidates at each split and minimum size of terminal nodes for RF; with a linear kernel, the cost of regularization for SVM; $\ell_1$ penalty for Lasso; $\ell_2$ penalty for Ridge; $\ell_1$ & $\ell_2$ penalties for Elastic net; number of nearest neighbors for KNN and finally, the number of genes (top *k*), number of units in a hidden layer and decay weights for NNET. With the optimal parameter(s) for each classification function, the class prediction models were built using the learning set. The resulting models were evaluated on a test set consisting of 5000 samples generated from the same model as the learning set (see Fig. 2). The error rates of the classification functions on this test set were recorded. The process was repeated 1000 times (sampling both learning and test sets) for each simulation scenario and the resulting error rates over the 1000 replications were used for further analyses.

### 2.4 Random effects linear regression

An average of the error rates of each and every classification function over 1000 replications for each simulated scenario was computed yielding 11 520 data points resulting from the 1152 different simulation scenarios by 10 classification functions. The error rates were then transformed to accuracies $(1-[\text{error rate}+0.001])$ and these accuracies were modeled using a linear random effects regression model with the classification function as the random effects clustering variable, by transforming the accuracies to an unbounded range using the logit function. For the $\ell^{\text{th}}$ standardized study factor, the random effects model is written as:

$$\log\left(\frac{\pi\ (\mathrm{x_{ij}})}{1-\ \pi\ (\mathrm{x_{ij}})}\right)=\mathrm{Y_{ij}}=\beta_0+\vartheta_{0\mathrm{j}}+(\beta_1+\vartheta_{1\mathrm{j}})\mathrm{X}^{\ell}_{\mathrm{ij}}+\epsilon_{\mathrm{ij}}$$

where $0<\pi(\mathrm{x_{ij}})<1$ is the average accuracy in scenario i for classification function j, $\vartheta_{\mathrm{j}}=(\vartheta_{0\mathrm{j}},\ \vartheta_{1\mathrm{j}})'\sim\mathrm{N}(0,\ \mathrm{D})$ are respectively the random intercepts and slopes of the classification functions while $\epsilon_{\mathrm{ij}}\sim\mathrm{N}(0,\sigma^2)$ are the independent and identically distributed residuals, also independent from the random effects $\vartheta_{\mathrm{j}}$. **D** is a $2\times2$ covariance matrix of the random effects. All the aforementioned study factors were evaluated by univariate and multivariate linear random effects regression models. Multivariate regression evaluation was done by a backward selection approach. In each step, two nested models, with and without a particular study factor, were compared by log-likelihood ratio test at 5% significance. Each factor $\ell$ was also evaluated by its explained-variation defined as:

$$\mathrm{Var}_\ell=\frac{\mathrm{MSE_{null}}\quad-\mathrm{MSE}_\ell}{\mathrm{MSE_{null}}}$$

where $\mathrm{MSE_{null}}$ and $\mathrm{MSE_l}$ are the mean square errors of the null (random intercept only) and the $\mathrm{l^{th}}$ standardized study factor models respectively. The explained variation of the selected multivariate model was also evaluated.

### 2.5 Software

All statistical analyses were performed in R software version 3.2.0, and Bioconductor (Gentleman *et al.*, 2009) using the following packages: *mvtnorm* (Genz and Bretz, 2009) for simulating data, *limma* (Ritchie *et al.*, 2015) for ranking genes via linear models, *CMA* (Slawski *et al.*, 2008) for predictive classification modeling, *lattice* (Sarkar, 2008) for visualization and *lme4* (Bates *et al.*, 2015) for linear random effects modeling. Additionally, we have developed an R package called '***SPreFuGED***': **S**electing a **Pre**dictive **Fu**nction for **G**ene **E**xpression **D**ata, that allows researchers to determine an optimal function for a given dataset.

## 3 Results

Figure 3 shows the average error rates over the 1000 random replicates (*y*-axis) of the functions (*x*-axis) for different combinations of variances, pairwise correlations of noisy (non-DE) genes and DE genes for a fixed sample size ($n=100$), proportion of DE genes ($\pi=5\%$) and $\log_2$ fold change ($\Delta=1$). From this figure, one sees that, the error rates for all functions increase with increasing variances (from top- to bottom-row), pairwise correlation values of non-DE genes (from left- to right-column) and pairwise correlation values of DE genes (different colored lines). On the other hand, other scenarios for different values of sample size, proportion of DE genes and $\log_2$ fold change (Supplementary Fig. S1A–C) indicate a negative association of sample size, proportion of DE gene and $\log_2$ fold change to the error rates. The non-constant variability of the error rates between classification functions across scenarios indicates a scenario-specific optimality for each and every classification function.

The average accuracies $(1-[\text{average error rates}+0.001])$ of the simulations were summarized to a data matrix as shown on Table 2. For each of the predictive variables, a linear random effects regression model was fitted as described in the method section. The individually explained variances of the study factors are depicted on Figure 4. This figure shows that sample size, pairwise correlations of non-DE genes and the proportion of DE genes are the leading factors respectively accounting for approximately 17, 14 and 13% of the null variance. While genes' variances and fold change respectively account for 8 and 7% of the null variance, pairwise correlations between DE genes accounts for simply 1%. As observed graphically, the univariate models (results not shown) confirmed a positive association of sample size, proportion of DE genes and fold change, and a negative association of pairwise correlations of non-DE, DE and the genes' variances to the accuracies.

For the multivariate linear random effects regression model, we started with a complex model of random intercepts and slopes and three ways interactions of the predictive factors. Starting with pairwise correlation between DE genes because of its low individually

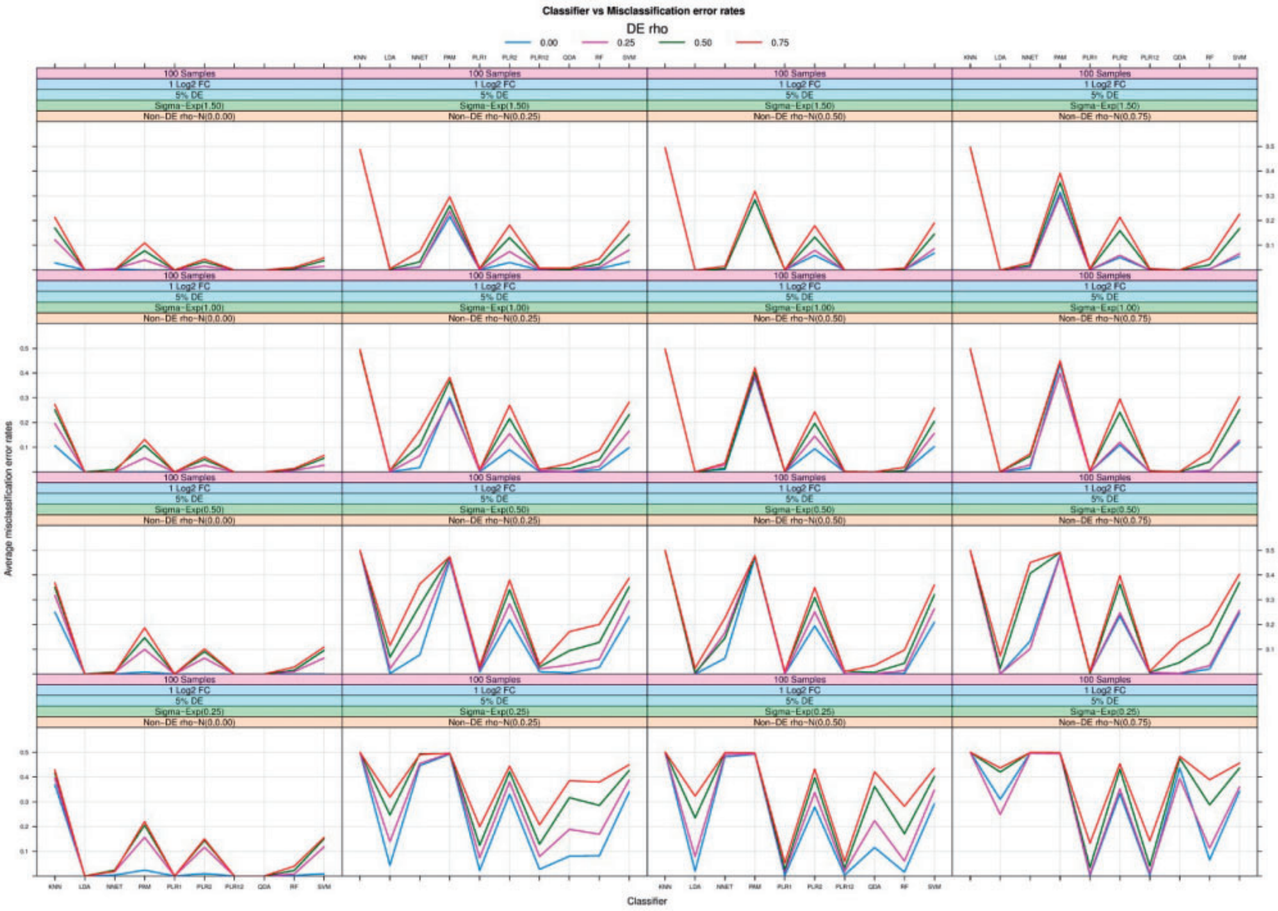


**Fig. 3.** Average misclassification error rates of the ten classification functions for sample size of 100, $\log_2$ fold change of 1 and 5% DE genes. Top-row to bottom row indicate increase in variance ($1/\lambda$) while from left-column to right-column indicate increase in the pairwise correlation of non-DE genes and the different colored lines from (blue–red) indicate increase in the pairwise correlations of DE genes (Color version of this figure is available at *Bioinformatics* online.)

**Table 2.** Structure of the performance data generated from evaluating the classification functions on the simulated data

| ID | Classifier | SampSize | propDE | Variance | deCorr | otherCorr | log2FC | Acc |
|---|---|---|---|---|---|---|---|---|
| 1 | SVM | 100 | 5 | 0.667 | 0.00 | 0.00 | 1 | 0.999 |
| 2 | SVM | 100 | 5 | 0.667 | 0.25 | 0.00 | 1 | 0.984 |
| 3 | SVM | 100 | 5 | 0.667 | 0.50 | 0.00 | 1 | 0.961 |
| 4 | SVM | 100 | 5 | 0.667 | 0.75 | 0.00 | 1 | 0.950 |
| 5 | KNN | 100 | 5 | 0.667 | 0.00 | 0.25 | 1 | 0.970 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 11515 | LDA | 20 | 1 | 4 | 0.50 | 0.75 | 0.5 | 0.498 |
| 11516 | LDA | 20 | 1 | 4 | 0.75 | 0.75 | 0.5 | 0.499 |
| 11517 | QDA | 20 | 1 | 4 | 0.00 | 0.75 | 0.5 | 0.500 |
| 11518 | QDA | 20 | 1 | 4 | 0.25 | 0.75 | 0.5 | 0.499 |
| 11519 | QDA | 20 | 1 | 4 | 0.50 | 0.75 | 0.5 | 0.499 |
| 11520 | QDA | 20 | 1 | 4 | 0.75 | 0.75 | 0.5 | 0.498 |

explained variance, we eliminated variables using the log-likelihood ratio test. We ended up with the model presented on Table 3 consisting of the fixed effects two ways interactions of all the six predictive factors, random intercepts and slopes. This model explains approximately 70% of the null variance as illustrated on Figure 4. The left panel of Table 3 presents the estimates of fixed effects, the standard errors and the t statistics while the top-right panel presents the net effect of a standard deviation (SD) unit increase of a given factor conditional on common values of other factors. Finally, the bottom-right panel presents the performances of the classification functions at different values of the predictive factors.

From the top-right panel of this table, one notices that a 1 SD unit increase in sample size, corresponding to $n = 89.67$ will lead to an increase in the Log odds (accuracy), with the highest increase observed when other variables are at their highest values. A similar effect is observed for a 1 SD unit increase in the proportion of DE genes. Though a 1 SD unit increase in fold change leads to an crease in the Log odds, as sample size and proportion of DE genes, its effect is highest when the other variables are at their lowest values. While on the average a 1 SD unit increase in the genes' variances, pairwise correlations of non-DE and DE genes will lead to a decrease in the accuracy, these effects become very severe when other variables are at their highest values. For very low values of other variables, a 1 SD unit increase of pairwise correlations between DE genes could even lead to an increase (a positive effect) on the accuracy as was previously observed and illustrated diagrammatically by Novianti *et al.* (2015). A similar effect is observed for a 1 SD unit increase in the genes' variances at very low values of other variables. These varying effects, indicate the complex interactions between the study factors and hence illustrate why classification functions will perform differently on different datasets.

Lastly, the bottom-right panel of the table shows that all classification functions will perform reasonably well if the predictive factors are at their average values (0 SD) with PLR1, PLR12, LDA and QDA having outstanding performances. For extremely small values

**Table 3.** Fixed effects estimates (left panel) and their conditional on other factors net effects (right panel)

| Parameter | Fixed effects<br>Estimate | <br>Std. Error | <br>t value |
|---|---|---|---|
| Intercept | 1.026 | 0.220 | 4.663 |
| StdSampSize ($\tilde{\mathbf{n}}$) | 0.573 | 0.123 | 4.660 |
| StdPropDE ($\tilde{\pi}$) | 0.508 | 0.108 | 4.695 |
| StdVariance ($\tilde{\sigma}^2$) | −0.400 | 0.081 | −4.915 |
| StdDECorr ($\tilde{\rho}$) | −0.158 | 0.027 | −5.878 |
| StdOtherCorr ($\tilde{\theta}$) | −0.560 | 0.085 | −6.557 |
| StdLog2FC ($\tilde{\Delta}$) | 0.378 | 0.068 | 5.565 |
| StdSampSize*StdLog2FC | 0.109 | 0.009 | 12.551 |
| StdPropDE*StdLog2FC | 0.145 | 0.009 | 16.718 |
| StdVariance*StdLog2FC | −0.109 | 0.009 | −12.588 |
| StdDECorr*StdLog2FC | −0.053 | 0.009 | −6.130 |
| StdSampSize*StdOtherCorr | −0.091 | 0.009 | −10.542 |
| StdPropDE*StdOtherCorr | −0.138 | 0.009 | −15.927 |
| StdVariance*StdOtherCorr | 0.102 | 0.009 | 11.819 |
| StdDECorr*StdOtherCorr | 0.019 | 0.009 | 2.182 |
| StdSampSize*StdDECorr | −0.067 | 0.009 | −7.706 |
| StdPropDE*StdDECorr | −0.119 | 0.009 | −13.745 |
| StdVariance*StdDECorr | 0.048 | 0.009 | 5.519 |
| StdSampSize*StdVariance | −0.161 | 0.009 | −18.571 |
| StdPropDE*StdVariance | −0.172 | 0.009 | −19.815 |
| StdSampSize*StdPropDE | 0.249 | 0.009 | 28.729 |

Conditional net effects of study factors

1 SD unit increase

| Other variables | | $\tilde{n}$ | $\tilde{\pi}$ | $\tilde{\sigma}^2$ | $\tilde{\rho}$ | $\tilde{\theta}$ | $\tilde{\Delta}$ |
|---|---|---|---|---|---|---|---|
| 2 SD | Log odds | 0.650 | 0.715 | −0.983 | −0.502 | −1.078 | 0.259 |
| 1 SD | | 0.612 | 0.473 | −0.691 | −0.330 | −0.819 | 0.318 |
| 0 SD | | 0.573 | 0.508 | −0.400 | −0.158 | −0.560 | 0.378 |
| −1 SD | | 0.534 | 0.543 | −0.108 | 0.015 | −0.300 | 0.438 |
| −2 SD | | 0.496 | 0.578 | 0.184 | 0.187 | −0.041 | 0.498 |
| 1 SD corresponds to | | $n = 89.67$ | $\pi = 4.63$ | $\sigma^2 = 3.22$ | $\rho = 0.65$ | $\theta = 0.65$ | $\Delta = 1.00$ |

| Other variables | | Classification functions<br>KNN | <br>LDA | <br>NNET | <br>PAM | <br>PLR1 | <br>PLR12 | <br>PLR2 | <br>QDA | <br>RF | <br>SVM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 SD | Log odds | −1.797 | 1.237 | −0.511 | −1.540 | 2.117 | 2.037 | −0.977 | 0.957 | 0.265 | −1.018 |
| 1 SD | | −0.445 | 1.835 | 0.550 | −0.232 | 2.451 | 2.396 | 0.178 | 1.632 | 1.085 | 0.150 |
| 0 SD | | 0.089 | 1.617 | 0.794 | 0.258 | 1.968 | 1.938 | 0.517 | 1.491 | 1.088 | 0.500 |
| −1 SD | | −0.193 | 0.580 | 0.221 | −0.070 | 0.667 | 0.662 | 0.038 | 0.532 | 0.273 | 0.033 |
| −2 SD | | −1.294 | −1.273 | −1.170 | −1.214 | −1.451 | −1.430 | −1.258 | −1.245 | −1.359 | −1.251 |

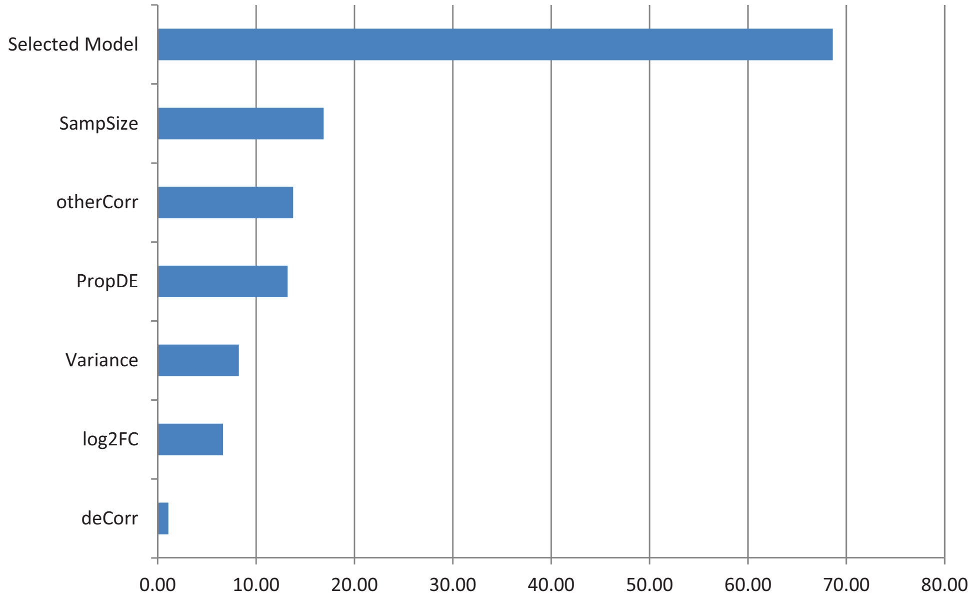


**Fig. 4.** Proportion of the null variance explained by each and every studied factor. The selected model refers to the predictive model presented on Table 3

(−2 SD) of the studied factors, all functions fail. An indication that the positively associated factors (sample size, proportion of DE genes and fold change) have a high combined net effect than the negatively associated factors (pairwise correlations between non-DE and DE genes and genes' variances). Additionally, for extremely large values (2 SD) of all predictive variables, classification functions like PLR1 and PLR12 clearly demonstrate their abilities to handle correlated variables and higher variances. That notwithstanding, the optimality of a function is scenario specific as illustrated on Supplementary Table S1 where both PLR1 and PLR12 fail when other variables are fixed at -2SD and otherCorr or DECorr is varied from -1SD to 2SD. It must be noted however that the combination of all other variables simultaneously being at −2, or at +2, is highly unlikely.

## 4 Application

To evaluate the predictive ability of the here presented random effects regression model on real-life data, eight Affymetrix gene

**Table 4**. Characteristics of the eight datasets used for evaluating the predictive model

| No. | Study | ID + | Affymetrix Platform | Probesets | SampSize | propDE | Variance | deCorr | otherCorr | log2FC |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CF* | E-GEOD-10406 | HG U133 Plus 2.0 | 54 675 | 15 (09, 06) | 0.267 | 0.143 | 1.205 | 0.414 | 0.408 |
| 2 | CS | E-MEXP-2236 | HG U133 Plus 2.0 | 5422 | 20 (10, 10) | 1.881 | 0.654 | 1.200 | 0.368 | 0.418 |
| 3 | Dia2 | E-CBIL-30 | HG U133A | 1749 | 26 (18, 08) | 2.859 | 0.444 | 0.604 | 0.611 | 0.434 |
| 4 | HIV2 | E-GEOD-14278 | HG U133 Plus 2.0 | 11 286 | 18 (09, 09) | 1.435 | 0.523 | 1.157 | 0.616 | 0.393 |
| 5 | UC2* | E-GEOD-21231 | HG 1.0 ST | 32 321 | 40 (20, 20) | 0.402 | 0.139 | 0.631 | 0.257 | 0.268 |
| 6 | UC3 | E-GEOD-36807 | HG U133 Plus 2.0 | 6541 | 28 (15, 13) | 6.849 | 0.735 | 1.381 | 0.715 | 0.503 |
| 7 | UC5 | E-MTAB-331 | HG 1.0 ST/HG 1.1 ST | 1402 | 59 (30, 29) | 1.427 | 0.461 | 1.390 | 0.951 | 0.286 |
| 8 | UC7* | E-GEOD-6731 | HG U95AV2 | 12 625 | 28 (11, 19) | 0.135 | 0.116 | 1.280 | 0.474 | 0.311 |

*No filtering was performed.; +: ArrayExpress accessing ID.

The sixth to the eleventh columns correspond to the variable under study. (.,.) represent the sample sizes for each class.

expression datasets of the 25 non-cancerous datasets described in one of our previous studies (Novianti *et al.*, 2015) were used. These datasets were selected to include a variety of Array platforms, both class-balance and class-imbalance, number of DE probesets, as well as various sample sizes. Three of these datasets were preprocessed without filtering while the other five were preprocessed and filtered as described by Novianti *et al.* (2015). We quantified the data characteristics studied and presented on Table 4 as follows: (i) sampSize, by counting the samples in the study, (ii) propDE, by ranking the probesets using limma (Ritchie *et al.*, 2015) and computing the proportion of DE probesets based on a $\log_2$ fold change cutoff of 1 if the number of DE is $\geq$10 or 0.5 otherwise, (iii) variance, was determined as the mean of the variances of all the probesets, (iv) log2FC, computed as the mean $\log_2$ fold changes of the DE probesets, (v) deCorr as the mean of the elements of the upper- (lower-) triangular of the correlation matrix of the DE probesets and (vi) otherCorr, was computed as the standard deviation (SD) of the elements of the upper- (lower-) triangular of the correlation matrix of non-DE probesets. This matrix was computed from all non-DE probesets if they were less than 20 000 or a sample of 20 000 from these non-DE probesets otherwise.

These data characteristics were standardized using the mean and SD of the respective variables from the simulated data. And our model was used to predict the accuracies for all classification functions in each dataset (Supplementary Fig. S2). We then built and evaluated classifiers using the classification functions by splitting the data into $\frac{2}{3}$ learning set and $\frac{1}{3}$ test set with stratification and a 3-fold inner cross-validation on the learning set for parameters optimization. This step was repeated a hundred times, each time predicting the accuracies of classification functions on the learning set using the random effects model and also recording the expected (observed) accuracies on the test set. These predicted and observed accuracies over the 100 repetitions are respectively presented on Supplementary Figure S3A and B. To compare the predicted to observed accuracies, and considering that we are interested in the ordering of performance (i.e. determining an optimal function for a given data), we used the ranked base Spearman correlation between the average predicted accuracies and the average observed accuracies.

The results of this comparison for each dataset are presented on Figure 5. The positive correlation values on this figure indicate agreement between our predicted and observed accuracies. Though these correlations are not very high in some datasets, our model more or less determined an optimal classification function for all the datasets except for UC7 where Ridge regression and SVM emerged first instead of fourth as predicted (i.e. 87.5% sensitivity). Nevertheless, the model was able to rule out on which classification(s) will perform poor on a given dataset, with approximately 100% certainty. As expected, the performance of the functions deteriorate on CF (small sample size and low proportion of DE probesets), Dia2 (high class-imbalance and small fold changes), UC2 (low proportion of DE probesets and small fold changes), UC3 (large variances and high correlations) and UC7 (low proportion of DE probesets). From Figure 5 and Supplementary Figure S3A and B, one sees that except on the UC3 data, our model's accuracies are less than or equal to observed accuracies. The model performs well on dataset with large sample sizes and balanced classes (UC2, UC3 and UC5). It attained its lowest performance on Dia2 where there is high class-imbalance and hence few samples of the small class in the learning set and on HIV2 and CF datasets with small sample sizes.

## 5 Discussion

We hypothesized that the performance of classification functions on gene expression data depends on sample size, proportion of DE genes, genes' variances, $\log_2$ fold changes between DE genes and magnitude of the pairwise correlation within DE genes and non-DE genes, and showed their association to the accuracies of ten often used and clinically relevant classification functions using simulations. Additionally, we built a predictive model to determine an optimal classification function among the studied functions using the simulation results. An application of the predictive model on eight non-cancerous real-life gene expression datasets predicted optimal function(s) for seven out of the eight and was able to rule out function(s) that will perform poor on almost all the datasets. This model may serve as a guide for choosing a classification function for a given gene expression data.

Classification functions have been shown to perform differently across gene expression datasets (Lee et al., 2005) and data characteristics have been shown to differ across datasets and are associated to the performance of classification functions (Jong *et al.*, 2014; Novianti *et al.*, 2015). While sufficient knowledge is available on the properties of most classification functions and procedures to build class prediction models using gene expression data have been outlined by Wessels *et al.* (2005), little is known about data characteristics that accounts for the variability in the performance of classification functions and how to use these characteristics to choose an optimal classification function for a specific dataset. As such, most researchers adhere to specific classification function(s) or randomly choose a classification for their class prediction models irrespective of the disease or data under study. A common practice is to evaluate several classification functions and select the one with smallest

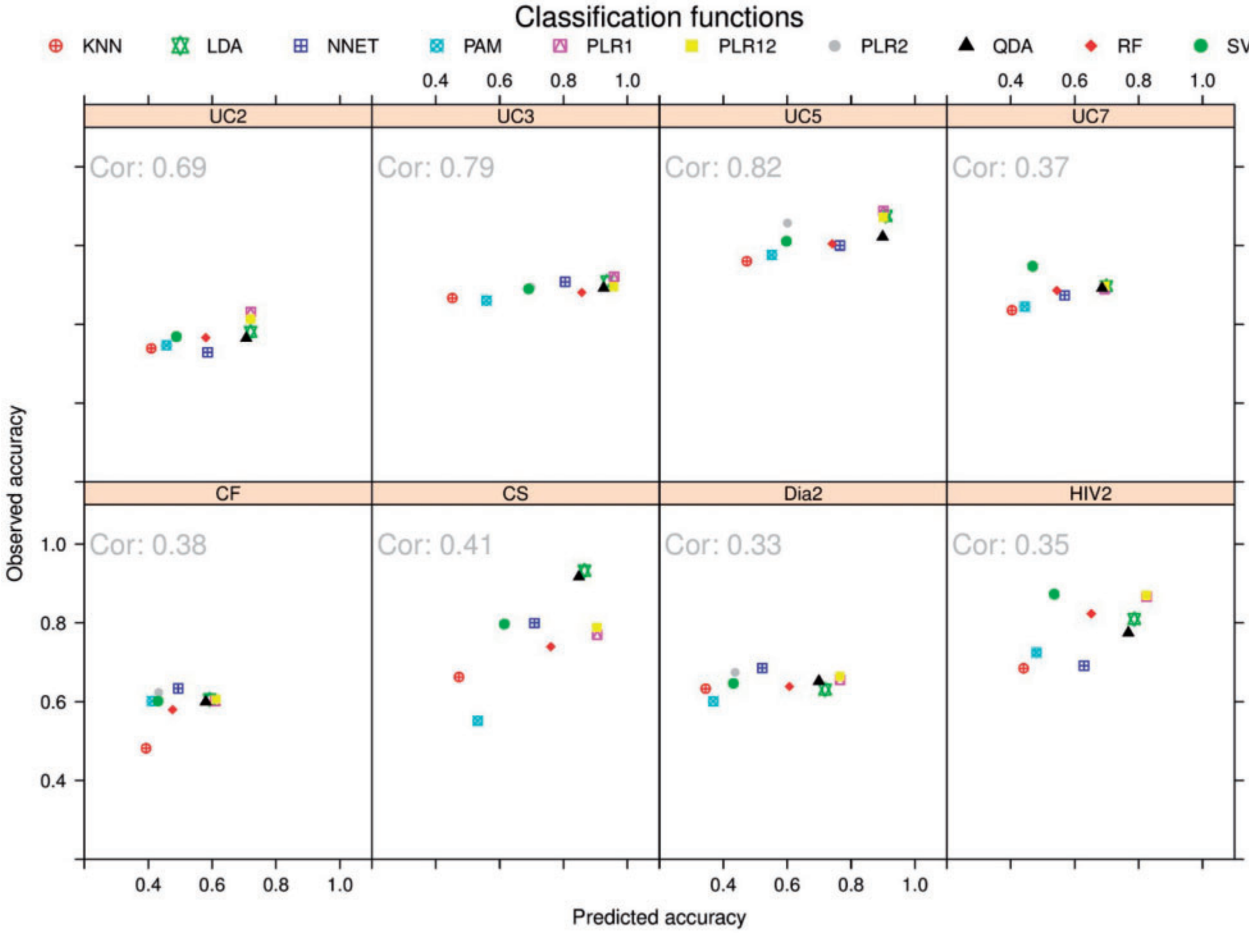


**Fig. 5.** Predicted versus Expected (Observed) accuracies. Cor represents Spearman correlations between the predicted and observed accuracies (Color version of this figure is available at *Bioinformatics* online.)

misclassification error but this leads to selection bias (Ding *et al.*, 2014).

In this study, we outlined data characteristics together with clinically relevant and often used classification functions and investigated their effects on classification performance using simulation studies. Based on these simulation studies, we provided a guide for choosing an optimal classification function for a specific dataset using the data's characteristics and the studied classification functions through a linear random effects predictive model. As a meta-model one would expect it to explain close to 100% of the variance in the simulated data but our predictive model accounts for approximately 70% of the variability in the simulated data. The remaining 30% unexplained variance may be associated to sampling variability stemming from the several (192) random covariance matrices used to generate both learning and test sets as well as the different learning and test sets generated at each iteration.

Although we used different classification functions and evaluated these functions using accuracy, our simulation results confirm the findings of Kim and Simon (2011) that classifiers tend to have poor performance on highly correlated data. Our results also agree with those of Novianti *et al.* (2015) that correlations, the absolute $\log_2$ fold changes and the number of DE probesets are associated to the accuracy of a class prediction model. In addition, these results specify clearly the directions of the association and point out the effects of other data characteristics like sample size, genes' variances that were not previously identified.

Most importantly, we have provided a predictive model that can serve as a guide to choose a classification function for a given dataset and its application on eight real-life datasets (both filtered and unfiltered) indicated a good predictive ability of the model. Although our model was reasonably good in its prediction on real-life data, we want to point out that it might have failed in some datasets because of the following reasons: (i) most of the eight non-cancerous datasets had small sample sizes and splitting these datasets to learning and test sets yielded even smaller sample sizes of the learning sets and hence might have led to poor estimates of the characteristics under study and (ii) the observed accuracies might not be the true accuracies because of the few Bootstrap samples. It could have been better if we had the means to perform several Bootstraps but due to the small sample sizes, the number of independent Bootstrap samples is limited. The fact that our predictions were most often slightly lower than the observed accuracies for almost all classification functions might indicate the general trend that the performance of a model usually decreases on an independent dataset. Hence, our model's predictions might reflect expected accuracies on independent datasets.

In the simulated data, we assumed exponential and normal distributions for the variances and pair-wise correlation of non-DE genes respectively. These distributional assumptions might be violated in some datasets. As such, it will be worth trying different distributions. Also, we used accuracy as a measure of evaluation by minimizing the loss function but in clinical applications, probabilities are more informative than simple yes or no predictions because they quantify the uncertainty of a prediction (Pepe, 2005). As such, it is worth evaluating these data characteristics on probabilistic classification functions where by the log-likelihood function is optimized, this might possibly provide a predictive model that will be most useful in clinical applications. Despite these limitations, our model was found to work well with data containing reasonably large and balanced sample sizes ($n \geq 30$). As such, our results apply to balanced class data. For data with class-imbalance some classification functions will have deteriorating performance, for which several solutions are proposed. However, this topic is outside the focus of the current study. In summary, our results serve as a guide to use data characteristics to choose an optimal classification function for a given dataset.

## Acknowledgements

The authors would like to thank the VIRGO consortium and its sponsors, the Netherlands Genomics Initiative and the Dutch Government for the financial support. Next to this, we thank Martin Marinus and the HPC-team at UMC Utrecht for the high performing computing facilities.

## Funding

This work has been supported by the VIRGO consortium, which is funded by the Netherlands Genomics Initiative and by the Dutch Government (FES0908). The funding agencies in no way influenced the outcome or conclusions of the study.

*Conflict of Interest*: none declared.

## References

Bates,D. *et al.* (2015). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-9

Bernau,C. *et al.* (2013) Correcting the optimal resampling-based error rate by estimating the error rate of wrapper algorithms. *Biometrics*, **69**, 693–702.

Breiman,L. (2002) Random forest. *Mach. Learn.*, **45**, 5–32.

Ding,Y. *et al.* (2014) Bias correction for selecting the minimal-error classifier from many machine learning models. *Bioinformatics*, **30**, 3152–3158.

Gentleman,R. *et al.* (2009) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Genz,A. and Bretz,F. (2009) *Computation of Multivariate Normal and T Probabilities*. Springer-Verlag, Heidelberg, Germany.

Golub,G. and Van Loan,C. (1996). *Matrix Computations*. 3rd edn. Johns Hopkins, Baltimore, USA.

Hastie,T. *et al.* (2003). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, NY, USA.

Huang,J. *et al.* (2010) Genomic indicators in the blood predict drug-induced liver injury. *Pharmacogenomics J.*, **10**, 267–277.

Jong,V. *et al.* (2014) Exploring homogeneity of correlation structures within and between gene expression datasets of different etiological disease categories. *Stat. Appl. Genet. Mol. Biol.*, **13**, 717–732.

Kim,K. and Simon,R. (2011) Probabilistic classifiers with high-dimensional data. *Biostatistics*, **12**, 399–412.

Lee,J. *et al.* (2005) An extensive comparison of recent classification tools applied to microarray data. *Comput. Statist. Data Anal.*, **48**, 869–885.

McLachlan,G. (1992) *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, NY, USA.

Novianti,P. *et al.* (2015) Factors affecting the accuracy of a class prediction model in gene expression data. *BMC Bioinf.*, **16**, 199.

Pepe,M. (2005) Evaluating technologies for classification and prediction in medicine. *Stat. Med.*, **24**, 3687–3696.

Ripley,B. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, MA, USA.

Ritchie,M. *et al.* (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

Sarkar,D. (2008) *Lattice: Multivariate Data Visualization with R*. Springer, NY, USA

Schölkopf,B. and Smola,A. (2002). *Learning with Kernels*. MIT Press, MA, USA.

Shi,L. *et al.* (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.*, **28**, 827–838.

Slawski,M. *et al.* (2008) CMA-a comprehensive Bioconductor package for supervised classification with high dimensional data. *BMC Bioinf.*, **9**, 439.

Smyth,G. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article3

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, **58**, 267–288.

Tibshirani,R. and Tibshirani,R. (2009) A bias correction for the minimum error rate in cross-validation. *Ann. Appl. Stat.*, **3**, 822–829.

Tibshirani,R. *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Nat. Acad. Sci. USA*, **99**, 6567–6572.

van 't Veer,L. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Varma,S. and Simon,R. (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinf.*, **7**, 91.

Wessels,L. *et al.* (2005) A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, **21**, 3755–3762.

Yang,K. *et al.* (2006) Correlation coefficient method for support vector machine input samples. *Mach. Learn. Cybern.*, 2857–2861.

Ye,G. *et al.* (2011) Efficient variable selection in support vector machines via the alternating direction method of multipliers. *Artif. Intell. Statist.*, **15**, 832–840.

Zhu,J. (2004) Classification of gene expression microarrays by penalized linear regression. *Biostatistics*, **5**, 427–443.

Zou,H. and Trevo,H. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, **67**, 301–320.