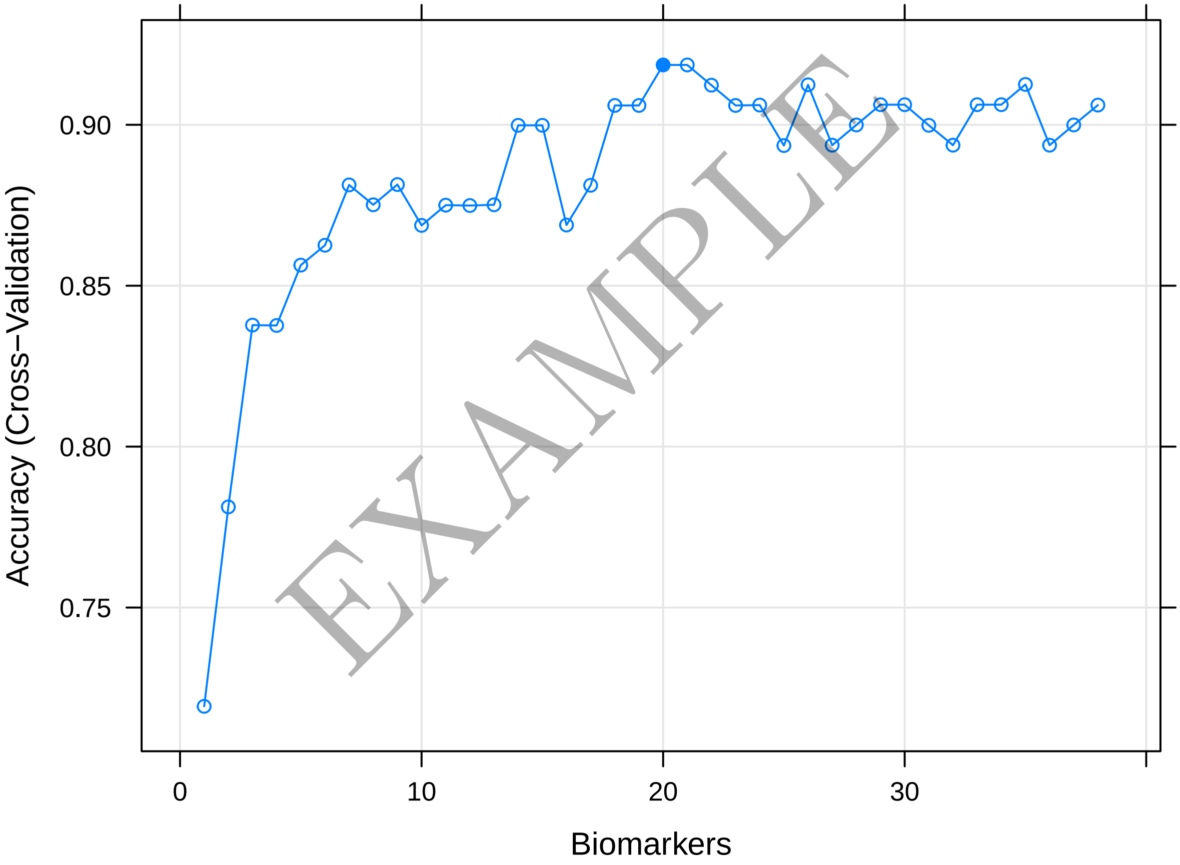
EXAMPLE REPORT
Biostatistics & Bioinfomatics Service
"Biomarker selection" Service



Bioinformatics Team, RayBiotech
December 06, 2018

Contents

1 Introduction

The "Biomarker selection" service aims to select a subset of biomarkers with modeling methods, which are able to distinguish samples of different groups efficiently. The "Biomarker selection" service requires a biomarker dataset from two or more groups of samples, and evaluates the performance of models of different biomarker combinations iteratively. The predictive/diagnositic performance of each biomarker/model is evaluated with receiver operating characteristic curve, accuracy and kappa value.

> *Need help understanding how the statistical analyses were performed in layman's terms? Please visit our* website.

2 Methods

2.1 Data filtration

Samples with missing data were identified and excluded from the analysis. Biomarkers showing no variation across all the subjects (i.e., zero-variance) were excluded from the analysis, too.

2.2 Data scaling

The raw biomarker values were scaled and centered during recursive feature selection and modeling to remove the effect of different scales in biomarker measurements.

2.3 Receiver operating characteristic (ROC)

The receiver operating characterisitic can describe the predictive performance of a continuous measurement at different cut-off values. It plots the curve of the ture positive rate (sensitivity) and false positive rate (1-specificity) at each data point of a continous measurement. In general, the area under curve (AUC) of a ROC reflects the diagnostic performance of a measurement, i.e., the larger AUC, the better.

ROCs can be used to evaluate not only single measurements, but also the performance of a predictive model developed with multiple measurements since the decision function of a model returns a single value for each input sample.

2.4 Recursive feature selection

The predictive model can use information from multiple measurements to predict response. A model may take into account all of the measurements; however, it is not uncommon that a subset of the measurements is used, achieving equivalent or better predictive performance. The recursive feature selection first develops a predictive model with all the measurements, and then removes measurements in a stepwise fashion to find the optimal combination of balancing model performance (e.g., AUC, accuracy) with complexity (e.g., number of included measurements).

There are many candidate models for feature selections. In this analysis we applied random forest because it is robust and non-parametric, requiring no assumptions on data distribution. During the selection procedure, the cross-validation with repeats technique was adopted to estimate the accuracy of models with different biomarker combinations.

2.5 Predictive modeling

Four predictive models were used: logistic regression (LR), linear discriminant analysis (LDA), support vector classification (SVC) and random forest (RF). First, the original dataset was split into training and testing datasets at a sample ratio of 3:1, respectively. Second, the model was developed with the training dataset using cross-validation for parameter tuning. Finally, the performance of the models was evaluated with the testing dataset.

The logistic regression model fits a sigmoidal function on the data:

$$Pr(y|X) = \frac{1}{e^{-(w^T X+b)} + 1}$$

The linear discriminant analysis starts from Bayes' rule:

$$Pr(y = k|X) = \frac{Pr(X|y = k)Pr(y = k)}{Pr(X)} = \frac{Pr(X|y = k)Pr(y = k)}{\sum_l Pr(X|y = l)Pr(y = l)}$$

and then estimates the conditional probability $Pr(X|y = k)$ from data with multivariate gaussian distribution density.

Support vector classification aims to find a hyperplane $y = w^T X + b$ such that the distance from a subset of datapoints (support vectors) to it is maximized. The task is to find these support vectors that define the hyperplane, which can be solved using a dedicated algorithm.

Random forest is an extension of a binary tree algorithm. It does not rely on a single best classification tree, but rather the majority vote across a group of trees (forest) generated by random resampling.

2.6 Software

All the analyses were conducted using R programming language V 3.5.1 (R Core Team 2017). ROC analysis was conducted with R package *pROC* (Robin et al. 2011). Modeling and recursive feature selection were conducted with R package *caret* (Kuhn et al. 2018).

3 Results

3.1 Data filtration

Samples with missing data: None.

Biomarkers with zero-variance: None.

All of the data were included into the analysis.

3.2 ROC analyses of individual biomarkers

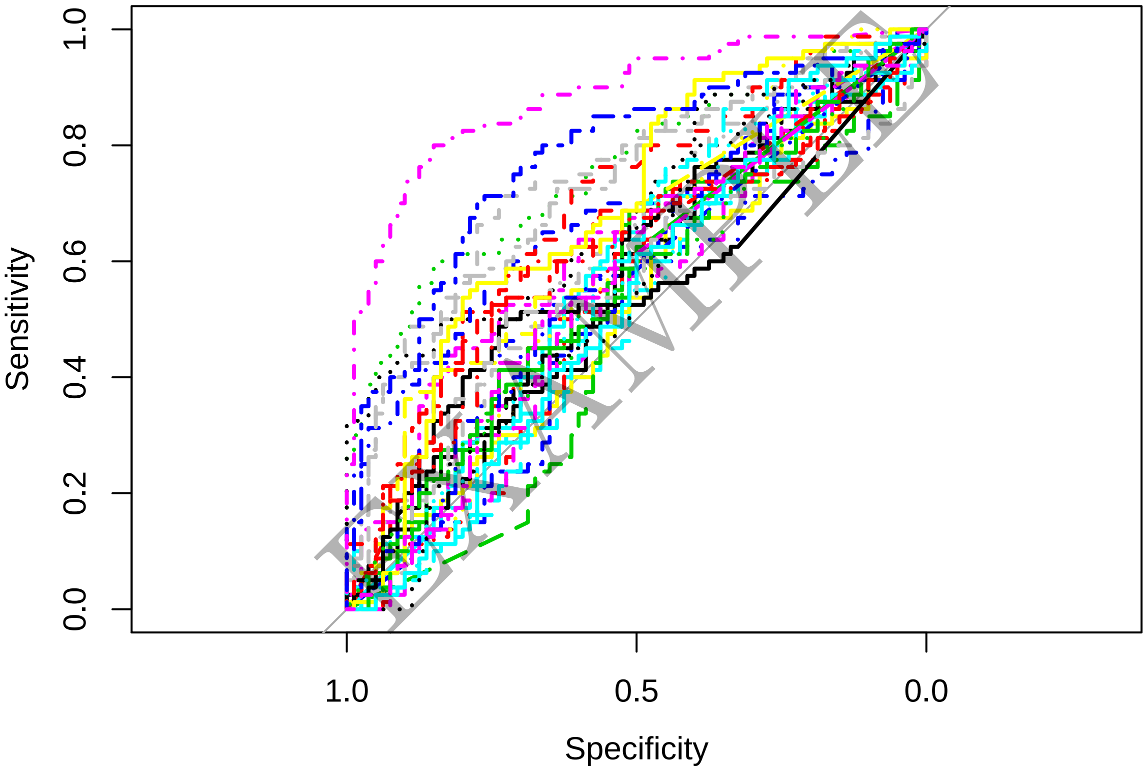


Figure 1: ROC curves of 38 biomarkers in 160 samples

Table 1: AUC of ROC curves of 38 biomarkers

Biomarker	AUC	Biomarker	AUC	Biomarker	AUC	Biomarker	AUC
MSPa	0.885	Leptin	0.621	IL-8	0.562	PDGF Rb	0.55
ApoA1	0.778	CEA	0.62	MIF	0.559	IL-2 Ra	0.541
BDNF	0.762	IL-6 sR	0.619	IL-1 R6	0.559	GROa	0.538
EGF	0.75	ICAM-1	0.615	VEGF	0.558	IGFBP-4	0.536
PDGF Ra	0.729	CA125	0.589	Prostasin	0.558	CA15-3	0.534
B2M	0.689	IL-6	0.581	transferrin	0.556	Adiponectin/ACRP30	0.533
PDGF-AA	0.686	AgRP	0.569	TIMP-4	0.553	CXCL16	0.517
EGF R	0.679	TIMP-2	0.567	HE4	0.552	IFNa	0.486
Mesothelin	0.651	MCSF	0.567	IGFBP-3	0.55		
OPN	0.632	AFP	0.565	Prolactin	0.55		

3.3 Recursive feature selection

We conducted recursive feature selection using the random forest model with 3-fold cross-validation of 10 repeats. The selection started with the full model of all 38 biomarkers in 160 samples, then decreased the number of biomarkers in the model at each iteration until only one biomarker remained in the model.

Table 2 lists the model performance, calculated from 10 3-fold cross-validations. The model with 20 biomarkers was selected (see Figure 2) based on parsimony principle, i.e., the model with less biomarkers being preferrable among those with smiliar performance.

Table 2: Accuracy of random forest models with different numbers of biomarkers during recursive feature selection

Variables	Accuracy	Kappa	AccuracySD	KappaSD
1	0.7193105	0.4401761	0.0782368	0.1578192
2	0.7812718	0.5636096	0.0284632	0.0563760
3	0.8377591	0.6761203	0.0457206	0.0914451
4	0.8376427	0.6757852	0.0197695	0.0399742
5	0.8563941	0.7133503	0.0199712	0.0404310
6	0.8625670	0.7255709	0.0092796	0.0189459
7	0.8813184	0.7630118	0.0094813	0.0192899
8	0.8751456	0.7505234	0.0276215	0.0555636
9	0.8814349	0.7630262	0.0381754	0.0765460
10	0.8687398	0.7379166	0.0014121	0.0025094
11	0.8750291	0.7503840	0.0102605	0.0209950
12	0.8749126	0.7502622	0.0121038	0.0238358
13	0.8751456	0.7505234	0.0276215	0.0555636
14	0.8998369	0.7999061	0.0294359	0.0588814
15	0.8998369	0.7999061	0.0294359	0.0588814
16	0.8688563	0.7380383	0.0174889	0.0354287
17	0.8812020	0.7627296	0.0115463	0.0232568
18	0.9060098	0.8123945	0.0332854	0.0662469
19	0.9060098	0.8122518	0.0382612	0.0764734
20	0.9185884	0.8373824	0.0292814	0.0584699
21	0.9185884	0.8373824	0.0292814	0.0584699
22	0.9122991	0.8248796	0.0293937	0.0586152
23	0.9060098	0.8123945	0.0332854	0.0662469
24	0.9061263	0.8125339	0.0195709	0.0390163
25	0.8935476	0.7875814	0.0295849	0.0586161
26	0.9124156	0.8251437	0.0117003	0.0231303
27	0.8936641	0.7878456	0.0223082	0.0440048
28	0.8999534	0.8001882	0.0114311	0.0228757
29	0.9062427	0.8126910	0.0188949	0.0378474
30	0.9062427	0.8129049	0.0188949	0.0375183
31	0.8998369	0.8001913	0.0294359	0.0582455
32	0.8936641	0.7877031	0.0119020	0.0234816
33	0.9062427	0.8129049	0.0188949	0.0375183
34	0.9062427	0.8126910	0.0188949	0.0378474
35	0.9125320	0.8253009	0.0104257	0.0209576
36	0.8936641	0.7877031	0.0119020	0.0234816
37	0.8999534	0.8001882	0.0114311	0.0228757
38	0.9061263	0.8126586	0.0195709	0.0387016

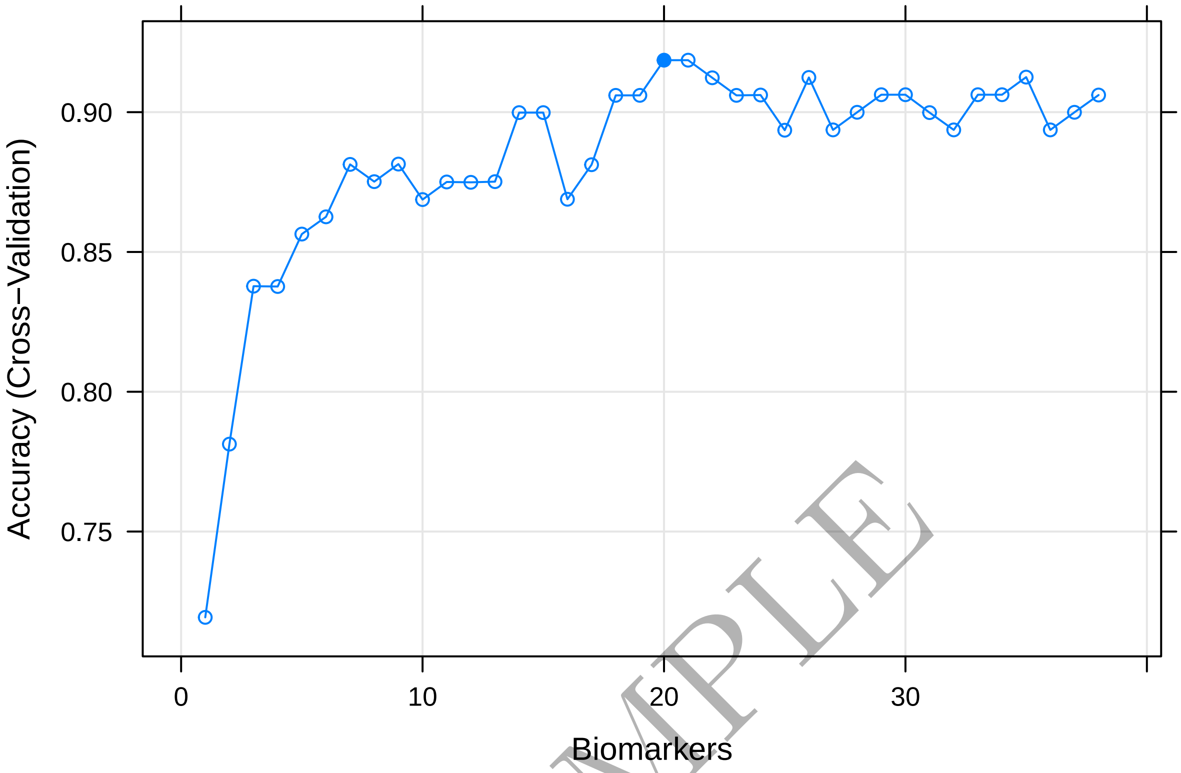


Figure 2: Recursive feature selection of 38 biomarker in 160 samples with random forest

Table 3 lists the biomarkers selected via recursive feature selection using the random forest model.

Table 3: 20 biomarkers selected via recursive feature selection with random forest model

	Control	Patient	MeanDecreaseAccuracy	MeanDecreaseGini
MSPa	0.1249990	0.0740078	0.0988050	18.243033
ApoA1	0.0401593	0.0246034	0.0322233	7.650248
BDNF	0.0339739	0.0112179	0.0224825	6.128297
CA125	0.0181102	0.0074944	0.0125277	3.892717
EGF R	0.0257428	0.0137007	0.0193219	4.559692
EGF	0.0356827	0.0208643	0.0279941	6.493262
PDGF Ra	0.0156242	0.0101955	0.0127225	3.427364
B2M	0.0078159	0.0096656	0.0088065	3.394235
PDGF-AA	0.0074144	0.0101417	0.0087509	2.875958
OPN	0.0088635	0.0054670	0.0072350	2.407283
Mesothelin	0.0073883	0.0018179	0.0045811	2.027645
PDGF Rb	0.0069362	0.0062613	0.0065727	2.106189
TIMP-4	0.0068601	0.0030756	0.0050138	2.580623
Leptin	0.0029104	0.0028819	0.0027284	1.964801
IL-6	0.0046661	0.0004480	0.0026577	1.700283
AgRP	0.0072556	0.0015203	0.0043236	2.006265
CEA	0.0033538	0.0018707	0.0024877	1.652084
ICAM-1	0.0054120	0.0081782	0.0068239	2.379354
HE4	0.0054693	0.0027919	0.0042011	1.834804
TIMP-2	0.0051403	0.0046450	0.0048746	2.204938

3.4 Modeling with biomarkers selected by recursive feature selection

Tables 4, 5 and Figure 3 show the performance of 4 models during cross-validation with 120 samples in the traing dataset.

Table 4: Accuracy of 4 models during cross-validation with 120 samples in the training data set

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
LR	0.8000000	0.8583333	0.9166667	0.9050000	0.9666667	1.0000000	0
LDA	0.7666667	0.8250000	0.8333333	0.8466667	0.8666667	0.9666667	0
RF	0.8666667	0.9000000	0.9500000	0.9366667	0.9666667	1.0000000	0
SVC	0.7666667	0.8000000	0.8333333	0.8533333	0.9083333	0.9666667	0

Table 5: Kappa values of 4 models during cross-validation with 120 samples in the training data set

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
LR	0.6000000	0.7166667	0.8333333	0.8100000	0.9333333	1.0000000	0
LDA	0.5333333	0.6500000	0.6666667	0.6933333	0.7333333	0.9333333	0
RF	0.7333333	0.8000000	0.9000000	0.8733333	0.9333333	1.0000000	0
SVC	0.5333333	0.6000000	0.6666667	0.7066667	0.8166667	0.9333333	0

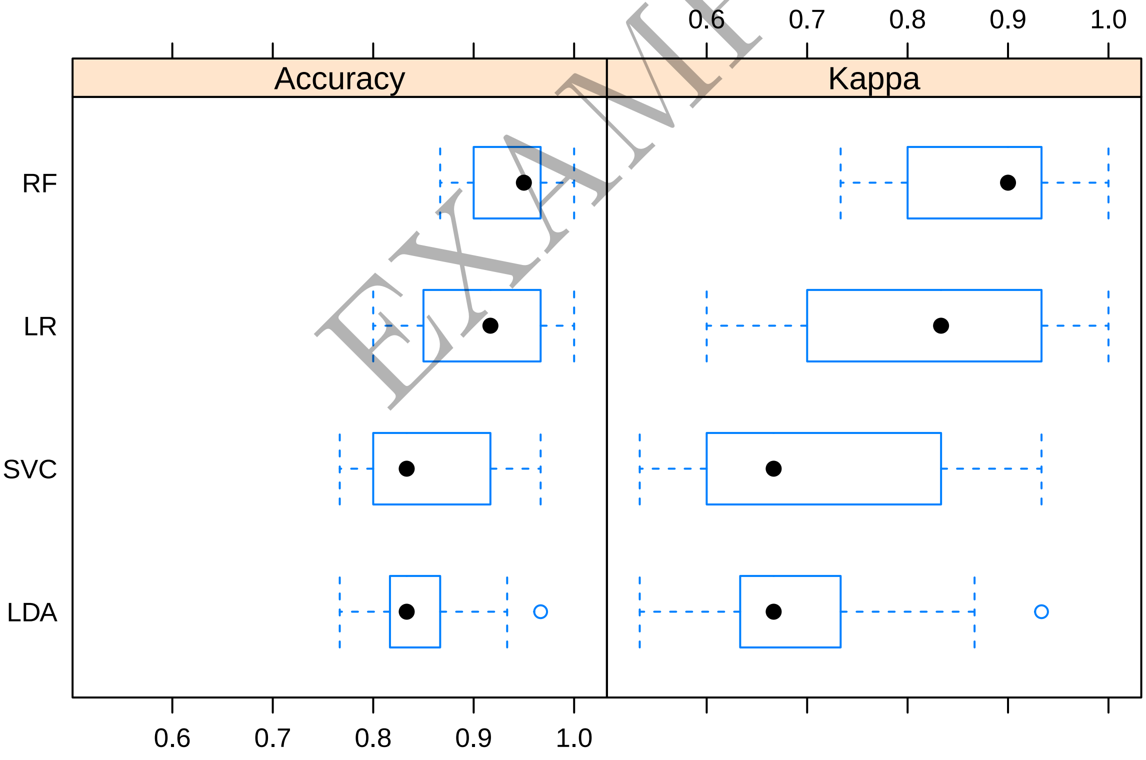


Figure 3: Performance of 4 models during cross-validation with 120 samples in the training data set

Tables 6 and 7 show coefficients in logisitic regression model and linear discriminant function, respectively.

Figures 4 - 7 show the performance of 4 models in 40 samples in the testing data set.

Table 6: Coefficients of logistic regression model

	Coefficient		Coefficient		Coefficient		Coefficient
(Intercept)	84.331443	‘EGF ‘	-29.777527	‘PDGF Rb ‘	22.524720	‘ICAM-1‘	-34.12761
‘MSPa ‘	222.102468	‘PDGF Ra‘	-19.916715	‘TIMP-4‘	15.191998	HE4	-29.59788
ApoA1	85.141653	B2M	-9.236591	‘Leptin ‘	-38.116400	‘TIMP-2‘	17.29349
BDNF	2.763249	‘PDGF-AA ‘	9.508341	‘IL-6‘	3.836784		
CA125	17.745830	OPN	119.260030	‘AgRP ‘	-70.473386		
‘EGF R‘	-39.087750	Mesothelin	-16.735473	CEA	49.384220		

Table 7: Coefficients of linear discriminants

	Coefficient		Coefficient		Coefficient		Coefficient
MSPa	0.7462224	PDGF Ra	-0.1288614	TIMP-4	0.1310751	HE4	0.0687977
ApoA1	0.3845782	B2M	-0.4123058	Leptin	-0.0949237	TIMP-2	0.1134803
BDNF	-0.1278794	PDGF-AA	-0.0804336	IL-6	0.0626574		
CA125	0.5032002	OPN	0.3461649	AgRP	-0.2670569		
EGF R	0.0805933	Mesothelin	-0.3544958	CEA	0.1057591		
EGF	-0.3323830	PDGF Rb	-0.0045838	ICAM-1	0.0088342		

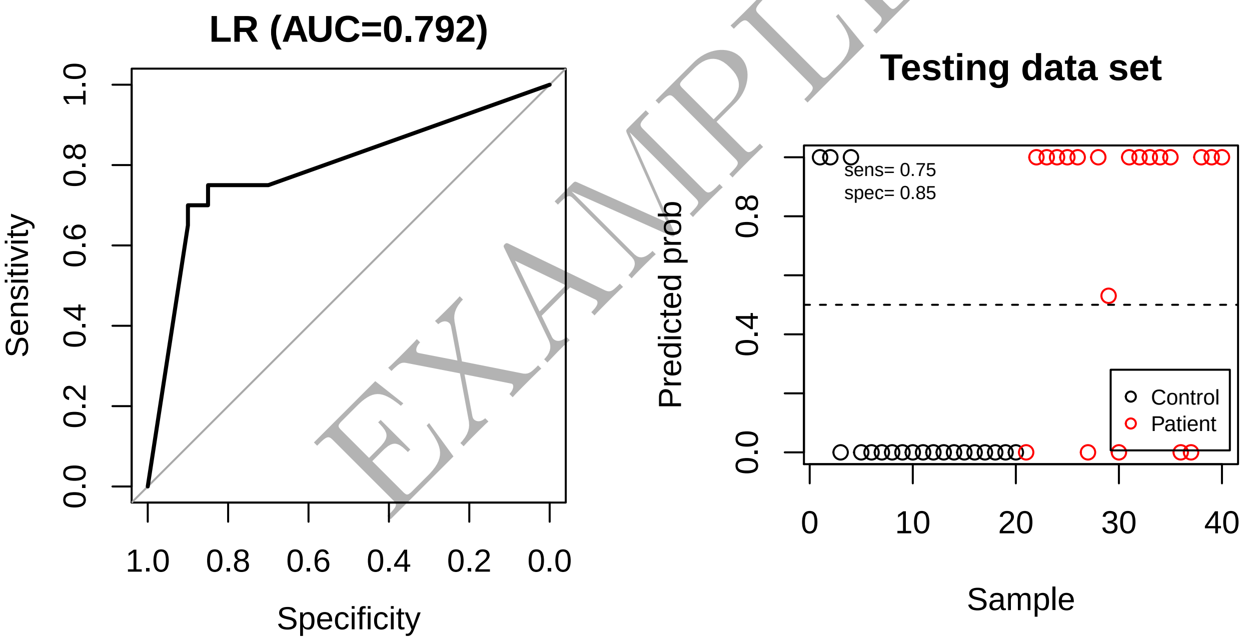


Figure 4: Performance of the logistic regression model of 20 biomarkers in 40 samples

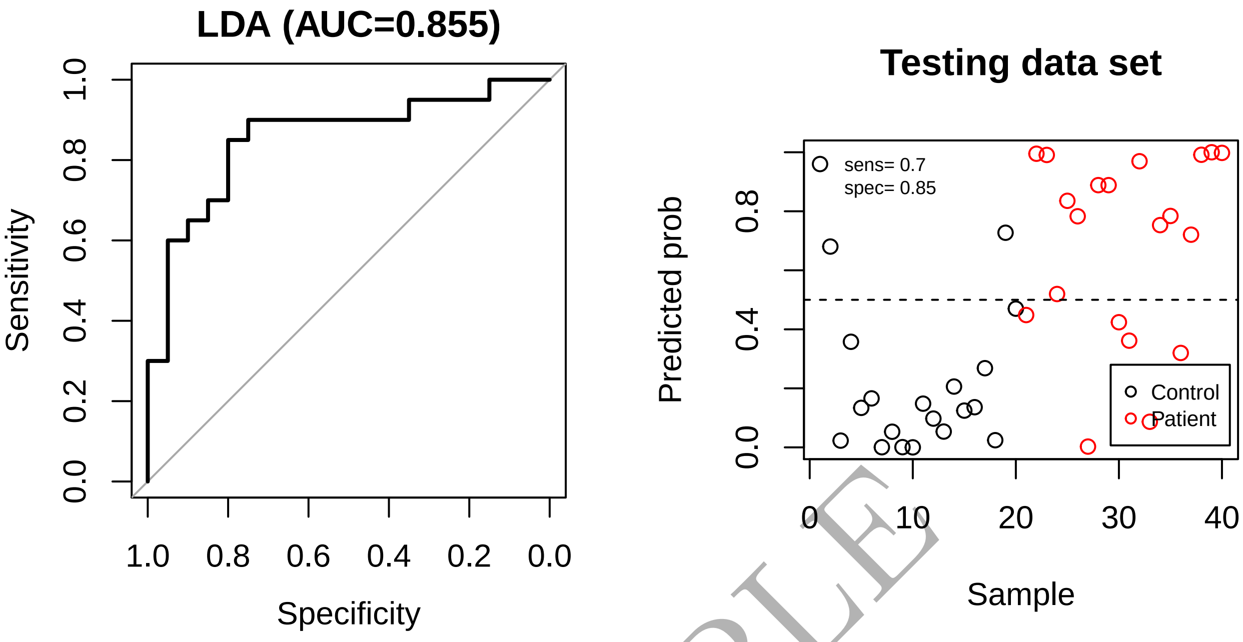


Figure 5: Performance of the linear discriminant analysis model of 20 biomarkers in 40 samples

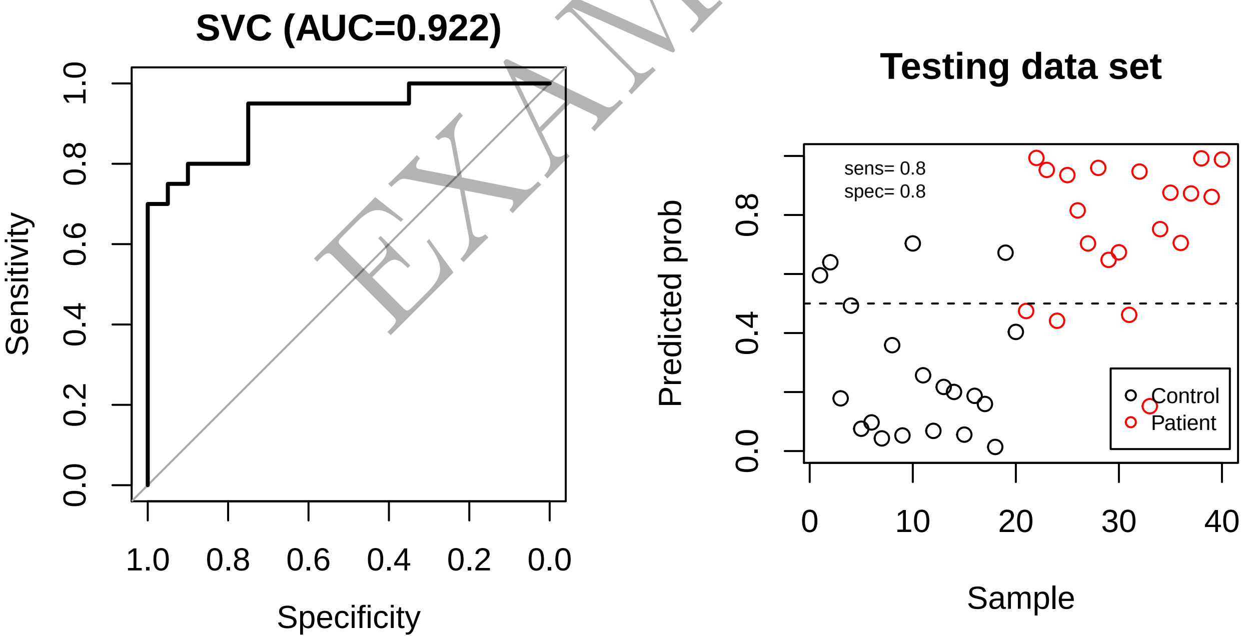


Figure 6: Performance of the support vector classification model of 20 biomarkers in 40 samples

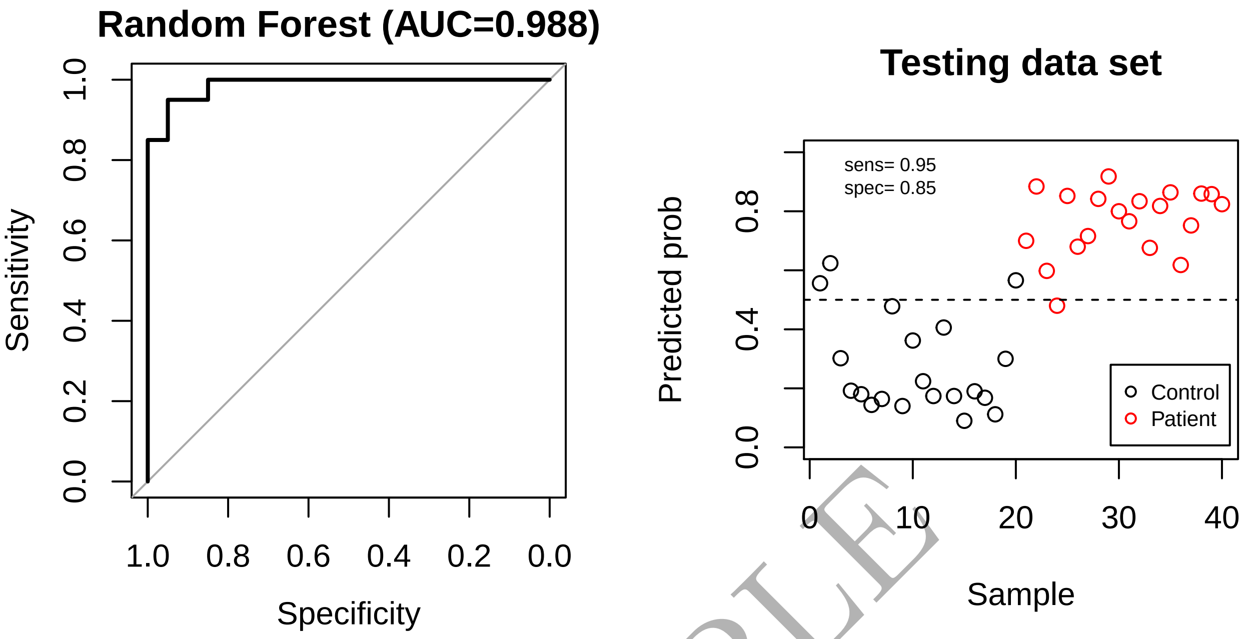


Figure 7: Performance of the random forest model of 20 biomarkers in 40 samples

References

Kuhn, Max, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, et al. 2018. *Caret: Classification and Regression Training.* https://CRAN.R-project.org/package=caret.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. “PROC: An Open-Source Package for R and S+ to Analyze and Compare Roc Curves.” *BMC Bioinformatics* 12: 77.