Analysis Methods Supplement for:

Classification and Diagnostic Prediction of Cancers using Gene Expression Profiling and Artificial Neural Networks

Javed Khan, Jun S. Wei, Markus Ringnér, Lao H. Saal, Marc Ladanyi,
Frank Westermann, Frank Berthold, Manfred Schwab,
Cristina R. Antonescu, Carsten Peterson, and Paul S. Meltzer

The data analysis consists of the following steps:

1 Initial Cuts

2 Principal Component Analysis

3 Artificial Neural Network Prediction

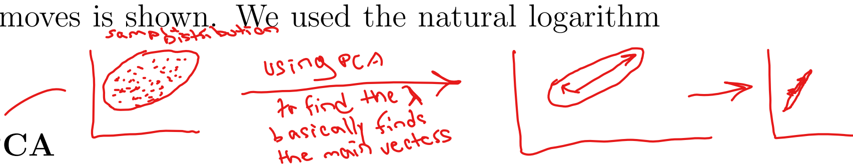
4 Extraction of Relevant Genes

examples

63 samples that are labeled so can be used to train with 2308 features each

1. Initial Cuts

In total, expression levels from 6567 genes are measured for each of the 88 samples, where 63 are labeled calibration samples and 25 represent blind tests. In the analysis we used the red intensity (*ri*) and the relative red intensity (*rri*). Genes are omitted if for any of the samples *ri* is less than 20. With this cut we are left with 2308 genes, which are used below for the analysis. The cut in *ri* mainly removes spots for which the image analysis failed. In Fig. 1 the number of genes each sample removes is shown. We used the natural logarithm of *rri* as a measure of the expression levels.



2. Principal Component Analysis – PCA

To allow for a supervised regression model with no "over-training" (i.e. low number of parameters as compared to the number of samples), we reduce the dimensionality of the samples using PCA [1]. Even though the formal dimension of the problem is given by the number of genes, the effective dimension is just one less than the number of samples. Hence the eigenvalue problem underlying PCA can be solved without diagonalizing 2308×2308 matrices by using singular value decomposition. Thus each sample is represented by 88 numbers, which are the results of projection of the gene expressions using the PCA eigenvectors. In what follows we use the 10 dominant components out of the 88 PCA eigenvectors to represent the expression data.

A potential risk when using PCA on relatively few samples is that components might be singled out due to strong noise in the data. One might then argue that the outputs (labels)

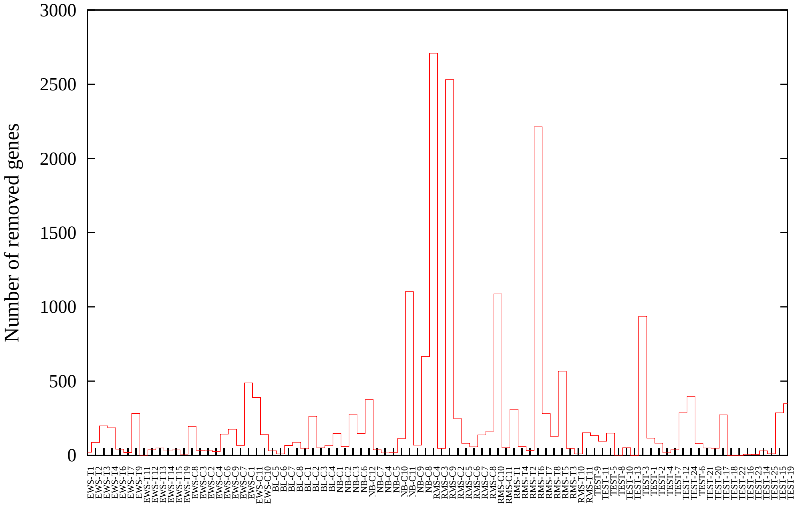


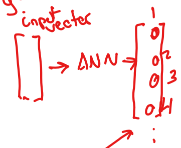
Figure 1: The number of genes (for each sample) which are removed by the cuts.

should be included in the dimensional reduction, using e.g. the Partial Least Squares (PLS) algorithm, in order to promote components with strong relevance for the output. However, based on explorations with similar data sets we strongly feel that this is not optimal; one introduces bias and implicitly "over-trains" already from the outset by including the outputs in the procedure.

3. Artificial Neural Network Prediction

the out put vector -- ie 4 nodes similar to hand written digits

input 2000 × 2000 is pixel matrix which gets vectorized



where the oupt is 10 nodes cosespond to the probability that the input cepsesents a 0–9 number.

Architecture and parameters. For prediction we employ an Artificial Neural Network (ANN) classifier (see e.g. [2]). Due to the limited amount of calibration data and the fact that four output nodes are needed [Ewing's sarcoma (EWS), Burkitt's lymphoma (BL), neuroblastoma (NB) and rhabdomyo sarcoma (RMS)] we limit ourselves to Linear Perceptrons (LP) with 10 input nodes representing the PCA components described above. In other words, the network contains 44 parameters including four threshold units. Using more than 8 PCA components did not improve the classifications of the samples. Since we could use 10 components without risking "over-training" we did not pursue to optimize the number of components to a somewhat smaller number. We have also investigated using all the PCA components as inputs followed by a subsequent pruning of weights to avoid "over-fitting". This resulted in the dominant 4-8 PCA components (depending on the composition of the training set) being the surviving inputs. We concluded that the less dominant PCA components contain variance not related to separating the four cancers, but rather to, for example, experimental conditions (noise) or variance related to sub-groupings within a can-

Final input vector are the PCA components

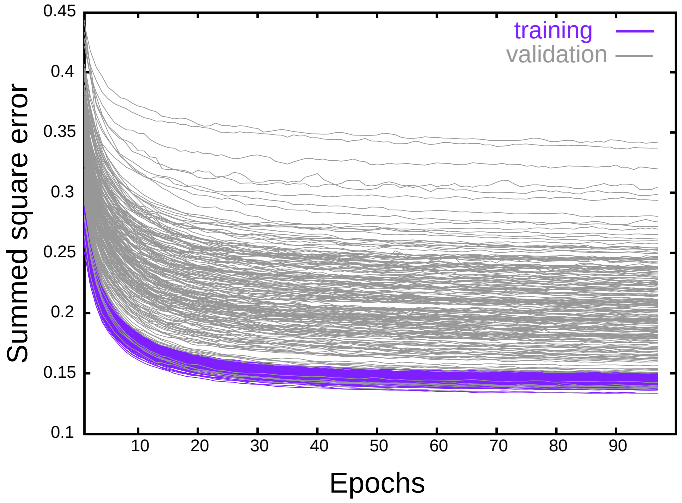


Figure 2: Performance of learning (purple) and validation (grey) sets for 200 models in terms of errors as functions of epochs.

Parameters

cer type. Calibration is performed using JETNET [3], with learning rate $\eta = 0.7$, momentum coefficient $p = 0.3$ and the learning rate is decreased with a factor 0.99 after each iteration. θ Initial weight values are chosen randomly from $[-r, r]$, where $r = 0.1/\max_i F_i$ and the "fan-in" F_i is the number of nodes connecting to node i. The calibration is performed using a training set and it is monitored both for the training set and a validation set, which is not subject to calibration (see below). The weight values are updated after every 10 samples and the calibration is terminated after 100 passes (epochs) through the entire training set. The resulting parameters for a completed training defines a "model".

Due to the limited amount of training data and the high performance achieved, we limited our analysis to linear (*i.e.* no hidden layers) ANN models. However, for other data sets we have extended our methods to use a hidden layer.

Calibration and validation. We use a 3-fold cross validation procedure for our predictions as follows: The 63 known (labeled) samples are randomly shuffled and split into 3 equally sized groups. ANN models are then calibrated as discussed above using two of the groups (training set) and the third group is reserved for testing predictions (validation set). Comparisons with the known answers refer to the results from the validation set (i.e. when using a model, the samples used for training the model are never used in predictions). This procedure is repeated 3 times, each time with a different group used for validation. The random shuffling is redone 1250 times and for each shuffling we analyze 3 ANN models. Thus, in total each sample belongs to a validation set 1250 times and 3750 ANN models have been calibrated.

The performance in terms of how the error of the validation set decreases with epochs is shown for 200 models in Fig. 2. As can be seen, there is no sign of "over-training" and all ANN models extrapolate well for their corresponding validation sets. The 1250 predictions for each validation sample can be used in two different ways. Either one looks at them

4 node
Binary output vector

independently [A] or one uses them as a committee [B]. Each ANN model gives a number between 0 (not this cancer type) and 1 (this cancer type) as an output for each cancer type. In [A] the maximal output is forced to 1 while the other outputs are forced to 0. One then determines how many of the 1250 predictions that are correct. In [B] one takes the average of all the predicted outputs (i.e. they all vote like in a committee) and one then forces this average to 0 or 1. In what follows, we used the average committee vote, [B], to classify samples. For validation samples the committee is based on 1250 models, while for additional test samples all 3750 models are used in the committee.

Assessing the quality of classifications. Each sample is classified as belonging to the cancer type corresponding to the largest average committee vote. In addition, we want to be able to reject the second largest vote as well as test samples which do not belong to any of the four cancer types. To this aim we define a distance d_c from a sample to the ideal vote for each cancer type:

Iterate thru the 4 cancer types

$$d_c = \boxed{\frac{1}{2}} \sum_{i=1}^{4} (\overset{\text{guess}}{o_i} - \overset{\text{truth}}{\delta_{i,c}})^2 \tag{1}$$

where c is a cancer type, o_i is the average committee vote for cancer type i, and $\delta_{i,c}$ is unity if i corresponds to cancer type c and zero otherwise. The distance is normalized such that the distance between two ideal samples belonging to different disease categories is unity. Based on the validation set, we generate for each cancer type an empirical probability distribution of its distances. The empirical probability distributions are built using each ANN model independently (not the average committee vote). Thus, the number of entries in each distribution is given by 1250 multiplied with the number of samples belonging to the cancer type. For a given test sample, we can reject possible classifications based on these probability distributions. This means that for each disease category we define a cutoff distance from an ideal sample within which we, based on the validation samples, expect a sample of this category to be. We have chosen the distance given by the 95th percentile of the probability distribution as a cutoff, which means that if a sample is outside of this cutoff distance it can not be confidently diagnosed. It should be noted that the classification as well as the extraction of important genes (see below) converges using less than 100 ANN models. The only reason we use 3750 ANN models is to have sufficient statistics for these empirical probability distributions.

Blind tests. Finally, these 3750 models are tested on 25 blinded (unlabeled) test samples. These belong to the four cancer types under investigation except 5 ”noise samples” originating from other tissues. The tests are done in two steps. First, we calibrate models using the 63 samples (divided into training and validation sets) as described above. Based on these models we extract the 96 genes which were most important for our classification as described in the next section and using only the 63 samples in these sets. Second, we redo the whole calibration procedure using only these 96 genes. Finally, the models based on these 96 genes were used to make predictions on the test set. Using a committee of the 3750 models calibrated using these 96 genes we correctly classify 100% of the 20 samples out of the 25 blind tests that belong to the four disease categories used in the calibration. The ANN committee predictions for the 25 unlabeled samples, as well as for the 63 validation samples, are given in Table 1.

Table 1: Classification and diagnosis by the committee of ANN models of all the samples. The average vote by the committee for each cancer type is a number between 0 and 1. If a sample falls outside the distance to the ideal vote as given by the 95th percentile (using empirical probability distributions based on the validation samples) a sample is classified but not diagnosed. The non-SRBCT noise samples are denoted in *italic*. The horizontal line separates training/validation samples from blind test samples. ARMS is alveolar RMS and ERMS is embryonic RMS.

Sample	ANN Committee Vote				ANN	ANN	Histological
Label	EWS	RMS	NB	BL	Classification	Diagnosis	Diagnosis
EWS-C1	0.91	0.02	0.27	0.04	EWS	EWS	EWS-C
EWS-C2	0.85	0.03	0.16	0.08	EWS	EWS	EWS-C
EWS-C3	0.89	0.04	0.10	0.08	EWS	EWS	EWS-C
EWS-C4	0.87	0.09	0.08	0.04	EWS	EWS	EWS-C
EWS-C6	0.93	0.11	0.03	0.05	EWS	EWS	EWS-C
EWS-C7	0.94	0.06	0.08	0.04	EWS	EWS	EWS-C
EWS-C8	0.98	0.05	0.04	0.04	EWS	EWS	EWS-C
EWS-C9	0.94	0.10	0.03	0.05	EWS	EWS	EWS-C
EWS-C10	0.81	0.22	0.03	0.06	EWS	EWS	EWS-C
EWS-C11	0.93	0.05	0.03	0.07	EWS	EWS	EWS-C
EWS-T1	0.99	0.04	0.03	0.06	EWS	EWS	EWS-T
EWS-T2	0.95	0.08	0.06	0.04	EWS	EWS	EWS-T
EWS-T3	0.97	0.10	0.05	0.03	EWS	EWS	EWS-T
EWS-T4	0.93	0.14	0.11	0.02	EWS	EWS	EWS-T
EWS-T6	0.97	0.12	0.04	0.04	EWS	EWS	EWS-T
EWS-T7	0.99	0.04	0.03	0.04	EWS	EWS	EWS-T
EWS-T9	0.95	0.13	0.03	0.03	EWS	EWS	EWS-T
EWS-T11	0.99	0.03	0.06	0.03	EWS	EWS	EWS-T
EWS-T12	1.00	0.02	0.03	0.03	EWS	EWS	EWS-T
EWS-T13	0.67	0.28	0.16	0.04	EWS	-	EWS-T
EWS-T14	0.99	0.02	0.04	0.05	EWS	EWS	EWS-T
EWS-T15	0.99	0.03	0.06	0.03	EWS	EWS	EWS-T
EWS-T19	0.93	0.06	0.09	0.04	EWS	EWS	EWS-T
RMS-C2	0.06	0.81	0.11	0.03	RMS	RMS	ERMS-C
RMS-C3	0.04	0.84	0.05	0.03	RMS	RMS	ARMS-C
RMS-C4	0.00	0.88	0.11	0.05	RMS	RMS	ARMS-C
RMS-C5	0.01	0.91	0.09	0.04	RMS	RMS	ARMS-C
RMS-C6	0.00	0.87	0.07	0.07	RMS	RMS	ARMS-C
RMS-C7	0.01	0.88	0.09	0.03	RMS	RMS	ARMS-C
RMS-C8	0.03	0.86	0.07	0.03	RMS	RMS	ERMS-C
RMS-C9	0.05	0.86	0.03	0.05	RMS	RMS	ARMS-C
RMS-C10	0.01	0.90	0.14	0.03	RMS	RMS	ARMS-C
RMS-C11	0.07	0.77	0.08	0.03	RMS	RMS	ERMS-C
RMS-T1	0.02	0.93	0.03	0.06	RMS	RMS	ARMS-T
RMS-T2	0.06	0.86	0.03	0.04	RMS	RMS	ARMS-T
RMS-T3	0.08	0.80	0.07	0.02	RMS	RMS	ERMS-T
RMS-T4	0.07	0.93	0.03	0.03	RMS	RMS	ERMS-T
RMS-T5	0.05	0.84	0.08	0.03	RMS	RMS	ARMS-T
RMS-T6	0.04	0.93	0.05	0.03	RMS	RMS	RMS-T
RMS-T7	0.10	0.75	0.05	0.05	RMS	RMS	ERMS-T
RMS-T8	0.06	0.90	0.05	0.02	RMS	RMS	RMS-T
RMS-T10	0.02	0.92	0.06	0.03	RMS	RMS	RMS-T
RMS-T11	0.03	0.76	0.06	0.03	RMS	RMS	ERMS-T
NB-C1	0.00	0.08	0.93	0.03	NB	NB	NB-C
NB-C2	0.03	0.10	0.70	0.08	NB	NB	NB-C
NB-C3	0.01	0.26	0.64	0.04	NB	NB	NB-C
NB-C4	0.02	0.03	0.85	0.06	NB	NB	NB-C
NB-C5	0.02	0.02	0.92	0.06	NB	NB	NB-C
NB-C6	0.02	0.02	0.89	0.09	NB	NB	NB-C
NB-C7	0.07	0.05	0.80	0.08	NB	NB	NB-C
NB-C8	0.00	0.06	0.96	0.04	NB	NB	NB-C

continued on the next page

Sample	ANN Committee Vote				ANN	ANN	Histological
Label	EWS	RMS	NB	BL	Classification	Diagnosis	Diagnosis
NB-C9	0.06	0.04	0.85	0.04	NB	NB	NB-C
NB-C10	0.00	0.12	0.91	0.03	NB	NB	NB-C
NB-C11	0.06	0.01	0.95	0.05	NB	NB	NB-C
NB-C12	0.02	0.24	0.41	0.06	NB	NB	NB-C
BL-C1	0.03	0.06	0.08	0.90	BL	BL	BL-C
BL-C2	0.04	0.12	0.04	0.82	BL	BL	BL-C
BL-C3	0.07	0.09	0.02	0.89	BL	BL	BL-C
BL-C4	0.04	0.06	0.08	0.80	BL	BL	BL-C
BL-C5	0.10	0.04	0.04	0.87	BL	BL	BL-C
BL-C6	0.10	0.02	0.09	0.87	BL	BL	BL-C
BL-C7	0.09	0.04	0.02	0.93	BL	BL	BL-C
BL-C8	0.20	0.03	0.03	0.89	BL	BL	BL-C
TEST-1	0.01	0.07	0.76	0.06	NB	NB	NB-C
TEST-2	0.67	0.06	0.08	0.09	EWS	EWS	EWS-C
TEST-3	0.11	0.17	0.16	0.11	RMS	-	*Osteosarcoma-C*
TEST-4	0.00	0.95	0.06	0.03	RMS	RMS	RMS-T
TEST-5	0.11	0.11	0.25	0.10	NB	-	*Sarcoma*
TEST-6	0.98	0.04	0.10	0.03	EWS	EWS	EWS-T
TEST-7	0.05	0.02	0.05	0.93	BL	BL	BL-C
TEST-8	0.00	0.05	0.94	0.04	NB	NB	NB-C
TEST-9	0.22	0.60	0.03	0.06	RMS	-	*Sk. Muscle*
TEST-10	0.10	0.68	0.11	0.04	RMS	-	RMS-T
TEST-11	0.39	0.04	0.28	0.15	EWS	-	*Prostate Ca.-C*
TEST-12	0.89	0.05	0.14	0.03	EWS	EWS	EWS-T
TEST-13	0.20	0.70	0.03	0.05	RMS	-	*Sk. Muscle*
TEST-14	0.03	0.02	0.90	0.07	NB	NB	NB-T
TEST-15	0.06	0.03	0.05	0.91	BL	BL	BL-C
TEST-16	0.03	0.02	0.93	0.05	NB	NB	NB-T
TEST-17	0.01	0.90	0.05	0.03	RMS	RMS	RMS-T
TEST-18	0.06	0.04	0.04	0.88	BL	BL	BL-C
TEST-19	0.99	0.02	0.04	0.05	EWS	EWS	EWS
TEST-20	0.40	0.30	0.10	0.06	EWS	-	EWS-T
TEST-21	0.81	0.19	0.12	0.04	EWS	EWS	EWS
TEST-22	0.01	0.88	0.09	0.04	RMS	RMS	RMS-T
TEST-23	0.07	0.08	0.70	0.06	NB	NB	NB-T
TEST-24	0.05	0.87	0.06	0.03	RMS	RMS	RMS-T
TEST-25	0.05	0.02	0.89	0.06	NB	NB	NB-T

For each sample several quantities are given in Table 1. The primary choice of the committee is our classification of the test sample. However, a sample is only diagnosed if its distance to the ideal vote falls inside the cutoff distance given by the 95th percentile of the empirical probability distribution for the validation samples. That is, a sample is only diagnosed if it is sufficiently similar to the samples used in the training. For completeness, we give the average vote by the committee for each cancer type. These averages can be interpreted as probabilities and they should sum up to one. If this is not the case, either the sample is outside the domain of validity of the training set or the training procedure is not appropriate. In our case the former alternative is the case.

For each disease category we calculate the sensitivity and specificity for our diagnosis (see Table 2). Both the sensitivity and the specificity are very high for all categories. It should be noted, that they depend on the kind of samples that are used as test samples. For example, using normal muscle samples as tests makes it harder to separate out RMS samples. If we only would have used samples from the four categories as blind tests distance cutoffs could easily have been designed such that both the sensitivity and the specificity would have been 100% for all diseases. We feel it is important that our method has been tested using a variety of blind tests. If one wants to improve rejection of for example normal muscle samples, one

Category	Sensitivity	Specificity	ROC curve area
EWS	93%	100%	1.0
BL	100%	100%	1.0
NB	100%	100%	1.0
RMS	96%	100%	1.0

Table 2: Sensitivities, specificities and ROC curve areas. The values were calculated using all the 88 samples, i.e. both validation and test samples were used.

could incorporate them as a fifth category in the training process. However, using more samples of all four categories in the training is initially probably the best way to improve the diagnostic separation.

The Receiver Operator Characteristic (ROC) curve area is identical to another more intuitive and easily computed measure of discrimination: the probability that in a randomly chosen pair of samples, one belonging to and one not belonging to the disease category, the one belonging to the category is the one with the closest distance to the ideal for that particular category. Since the ROC curve areas are unity for all disease categories (see Table 2), it is possible to define cutoff distances such that both the sensitivity and the specificity are 100% for all diseases. However, based on the training and validation sets it is difficult to motivate such cutoff distances.

4. Relevant Gene Extraction

Finding relevant variables for given outputs can in principle be done in two ways; (1) model-independent and (2) model-dependent analysis respectively. Due to the relatively few samples, we have chosen the latter using the ANN models.

We define the sensitivity (S) of the outputs (o) with respect to any of the 2308 input variables (x_k) as:

$$S_k = \frac{1}{N_s}\frac{1}{N_o}\sum_{s=1}^{N_s}\sum_{i=1}^{N_o}\left|\frac{\partial o_i}{\partial x_k}\right| \tag{2}$$

where N_s is the number of samples (63 or 88) and N_o is the number of outputs (4). The procedure for computing S_k involves a committee of 3750 models. In addition we have defined a sensitivity for each output i (S_i), which is analogous to Eq. (2) but without the sum over outputs. For these latter sensitivities we have also defined a sign of the sensitivity, which signals if the largest contribution to the sensitivity stems from positive or negative terms. A positive sign implies that increasing the expression rate of the gene increases the possibility that the sample belongs to this cancer type, while a negative sign means that decreasing the expression rate of the gene increases the same possibility. In other words, the sign does not tell whether a gene is up- or down-regulated but if it is more or less expressed in this cancer type as compared to the others. This means that we not only rank the genes according to their importance for the total classification but also according to their importance for the different disease categories separately. In Table 3 the total rank as well as the separate rank for each disease category is shown for the 96 top ranked genes. Based on these ranks we

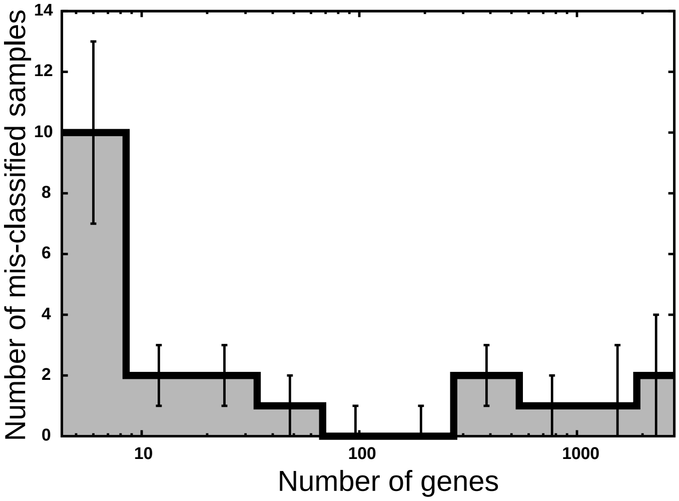


Figure 3: Validation set performance (number of mis-classifications) for 6, 12, 24, 48, 96, 192, 384, 768, 1536 and 2308 genes respectively. Here the committee vote is not used, instead each ANN model is used as an independent classifier. The average and standard deviation (rounded to integers) of the performance for the 1250 models used for each sample is shown.

have classified each gene according to in which disease category it is highly expressed.

Once we have established a ranking list among the in-going 2308 genes, the question of how many of these are really needed to produce the classification results naturally arises. We have explored this issue by selecting the top 6, 12, 24, 48, 96, 192, 384, 768 and 1536 genes and for each choice redone the entire calibration procedure. The results in terms of the number of mis-classified samples in the validation set are shown in Fig. 3. In this figure we did not use the average committee vote. Instead, each ANN model is used as an independent classifier. Using the committee vote always gives equal or better results than this type of classification. However, using this classification method to optimize the number of genes is more conservative. When employing the average committee vote one may risk using a smaller subset of genes which does not work perfectly for some random partitions. As can be seen from Fig. 3, 100% correct classification is obtained using only 96 genes. One could optimize this further, but we feel using significantly less than 96 genes is not optimal with respect to noise in the data (and in future test data).

Table 3: The top 96 ranked genes. For each gene its total rank and its rank for each category separately is given. Sign is the sign of the sensitivity for each category. Based on the separate ranks we have classified the genes according to in which category they are highly expressed (Gene Class).

Rank	Image Id.	Gene	EWS		RMS		NB		BL		Gene Class
			Rank	Sign	Rank	Sign	Rank	Sign	Rank	Sign	
1	296448	IGF2	8	−	1	+	918	−	19	−	RMS
2	207274	IGF2	19	−	2	+	1152	−	11	−	RMS
3	841641	CCND1	11	+	38	−	118	+	6	−	EWS/NB
4	365826	GAS1	25	+	69	+	22	−	9	−	EWS/RMS
5	486787	CNN3	130	−	39	+	14	+	17	−	RMS/NB
6	770394	FCGRT	3	+	186	−	79	−	18	−	EWS
7	244618	EST	22	−	3	+	273	−	86	−	RMS
8	233721	IGFBP2	148	+	43	+	598	+	1	−	Not BL
9	43733	GYG2	4	+	261	−	99	−	21	−	EWS
10	295985	EST	1	−	51	+	9	+	522	+	Not EWS
11	629896	MAP1B	360	−	893	+	1	+	23	−	NB
12	840942	HLA-DPB1	1161	+	383	−	6	−	12	+	BL
13	80109	HLA-DQA1	226	−	1589	−	20	−	3	+	BL
14	41591	MN1	257	+	18	+	4	−	169	−	EWS/RMS
15	866702	PTPN13	2	+	74	−	230	−	62	−	EWS
16	357031	TNFAIP6	5	+	119	−	103	−	60	−	EWS
17	782503	EST	26	+	219	−	104	+	14	−	EWS/NB
18	377461	CAV1	6	+	91	−	90	−	101	−	EWS
19	52076	NOE1	7	+	33	−	1673	+	37	−	EWS
20	811000	LGALS3BP	24	+	246	−	257	+	13	−	EWS/NB
21	308163	EST	49	+	88	+	191	−	22	−	RMS/EWS
22	812105	AF1Q	670	−	934	−	2	+	51	−	NB
23	183337	HLA-DMA	317	−	1574	−	24	−	8	+	BL
24	714453	IL4R	208	−	20	+	8	−	238	+	RMS/BL
25	298062	TNNT2	43	−	4	+	95	−	475	−	RMS
26	39093	MNPEP	46	+	224	+	21	−	103	−	EWS/RMS
27	212542	EST	62	+	993	+	1086	+	2	−	Not BL
28	204545	EST	471	+	49	+	1455	+	5	−	Not BL
29	383188	RCV1	478	−	808	+	13	+	42	−	NB
30	82225	SFRP1	160	−	264	+	17	+	85	−	NB
31	44563	GAP43	693	−	191	−	3	+	166	−	NB
32	289645	APLP1	41	+	102	−	107	+	61	−	EWS/NB
33	324494	HSPB2	1605	−	13	+	7	−	420	−	RMS
34	563673	ATQ1	35	+	1527	−	523	+	7	−	Not BL
35	1473131	TLE2	10	+	1884	−	16	−	217	−	EWS
36	1416782	CKB	134	+	416	+	851	+	4	−	Not BL
37	417226	MYC	63	+	222	−	29	−	110	+	EWS/BL
38	878280	CRMP1	602	−	1522	+	12	+	45	−	NB
39	812965	MYC	23	+	296	−	11	−	308	+	EWS/BL
40	122159	COL3A1	791	+	29	+	1062	−	16	−	RMS
41	609663	PRKAR2B	198	−	55	−	550	+	29	+	BL
42	461425	MYL4	98	−	7	+	80	−	419	−	RMS
43	1469292	PIM2	1007	+	242	−	53	−	36	+	BL
44	809910	1-8U	52	+	168	+	159	−	56	−	RMS/EWS
45	824602	IFI16	336	+	149	−	33	−	89	+	EWS/BL
46	245330	IGF2	65	−	6	+	147	−	434	−	RMS
47	135688	GATA2	354	+	155	−	37	+	88	−	NB
48	1409509	TNNT1	141	−	8	+	153	−	313	−	RMS
49	788107	AMPHL	74	−	14	+	817	+	108	−	RMS
50	784593	EST	224	−	299	+	39	+	68	−	RMS/NB
51	756556	C1NH	90	+	238	+	284	−	38	−	RMS/EWS
52	208718	ANXA1	12	+	827	−	1202	−	33	−	EWS
53	308231	EST	524	−	1015	+	10	+	117	−	NB
54	486110	PFN2	1554	+	1500	+	31	+	31	−	NB
55	21652	CTNNA1	104	+	117	+	2245	−	15	−	Not BL

continued on the next page

Rank	Image Id.	Gene	EWS		RMS		NB		BL		Gene Class
			Rank	Sign	Rank	Sign	Rank	Sign	Rank	Sign	
56	377671	ITGA7	1044	+	24	+	66	−	135	−	RMS
57	745343	REG1A	166	+	93	−	40	−	153	+	EWS/BL
58	241412	ELF1	882	−	1473	−	60	−	27	+	BL
59	504791	GSTA4	276	+	2003	+	108	+	24	−	Not BL
60	841620	DPYSL2	51	+	100	−	366	+	70	−	EWS/NB
61	859359	PIG3	58	−	28	+	288	+	152	−	RMS/NB
62	45542	IGFBP5	991	+	89	+	1661	−	10	−	RMS
63	80338	SELENBP1	20	+	1316	+	42	−	151	−	EWS
64	45291	DRPLA	532	+	81	+	872	−	28	−	Not BL
65	323371	APP	1689	−	90	+	594	+	65	−	Not BL
66	897788	PTPRF	59	+	1358	−	734	+	20	−	Not BL
67	377731	GSTM5	13	+	310	−	34	−	381	−	EWS
68	784224	FGFR4	36	−	5	+	431	−	604	−	RMS
69	293500	EST	262	−	9	+	1084	−	138	−	RMS
70	767183	HCLS1	1481	−	1424	−	50	−	32	+	BL
71	297392	MT1L	1361	−	483	−	113	−	30	+	BL
72	325182	CDH2	590	−	919	−	5	+	260	−	NB
73	1435862	MIC2	14	+	518	−	371	−	97	−	EWS
74	377048	EST	733	−	560	+	23	+	102	−	NB
75	814260	FVT1	9	+	61	−	330	−	335	−	EWS
76	784257	KIF3C	577	+	1099	−	64	+	44	−	NB
77	42558	GATM	379	−	12	+	25	−	1020	−	RMS
78	814526	HSRNASEB	164	−	198	+	98	−	105	+	RMS/BL
79	839736	CRYAB	516	+	67	+	51	−	183	−	EWS/RMS
80	395708	DPYSL4	1269	+	591	−	28	+	91	−	NB
81	416959	NFIB	1420	−	86	+	160	+	72	−	RMS/NB
82	364934	DAPK1	42	+	1481	+	707	−	40	−	EWS
83	868304	ACTA2	1286	−	151	−	122	−	71	+	BL
84	755599	IFI17	16	+	177	−	30	−	918	−	EWS
85	246377	EST	719	−	36	+	641	+	75	−	RMS
86	291756	TUBB5	17	+	31	−	1325	+	245	−	EWS
87	809901	COL15A1	1516	−	23	+	35	−	385	−	RMS
88	769959	COL4A2	1575	+	66	+	1786	−	26	−	RMS
89	796258	SGCA	30	−	10	+	521	−	758	−	RMS
90	854899	DUSP6	774	+	150	+	838	+	39	−	Not BL
91	755750	NME2	1840	+	26	+	591	−	82	−	RMS
92	292522	EST	221	−	667	+	32	+	189	−	NB
93	308497	EST	27	+	1971	−	43	−	231	−	EWS
94	813266	FHL1	1045	+	1610	−	91	+	46	−	NB
95	200814	MME	639	−	1081	+	78	−	66	+	BL
96	768370	TIMP3	547	+	1132	+	606	+	25	−	Not BL

References

[1] I. T. Jollife, "Principal Component Analysis", *Springer-Verlag* (New York) 1986.

[2] C. M. Bishop, "Neural Networks for Pattern Recognition", *Clarendon Press* (Oxford) 1995.

[3] C. Peterson, T. Rögnvaldsson and L. Lönnblad, "JETNET 3.0 - A Versatile Artificial Neural Network Package", *Computer Physics Communications* **81**, 185-220 (1994).