Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Verification of a blood-based test for breast-cancer (BLOBREC)

Distinguishing breast-cancer patients from population-based controls

Note no. SAMBA/33/15

Authors Marit Holden, Clara-Cecilie Günther, Lars Holden

Date October 2015

NR Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Norsk Regnesentral

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Title	**Verification of a blood-based test for breast-cancer (BLOBREC)**
Authors	**Marit Holden, Clara-Cecilie Günther, Lars Holden**
Date	October 2015
Year	2015
Publication number	SAMBA/33/15

Abstract

BLOBREC is a test for distinguishing breast-cancer patients from population-based controls described by Dumeaux et al. We have performed a quality control of the methods and procedures used for developing this test. Besides reproducing results obtained using exactly the same datasets as Dumeaux et al., we examined how sensitive the test results are to the approach selected when preprocessing the data, and whether the test results are influenced by drug use, smoking, or stress due to a potential cancer diagnosis. We also examined if the test results for the breast-cancer patients depend on whether the patient participated in the screening program. A dataset intended for examining the effect of stress was used as a validation dataset for the test.

Our analyses confirm the results obtained by Dumeaux et al. We obtain comparable results when using different approaches for preprocessing the data. Prediction performance for the datasets used when developing the test is clearly better than for the validation dataset. Batch effects and other differences between the datasets are the most likely explanations for this difference. However, the validation dataset consists of many different subgroups of individuals with a limited number of individuals in each subgroup, making the interpretation uncertain.

We were not able to show that the test is influenced by stress, drug use or smoking, but again, the datasets are too small to draw any firm conclusions.

This work was funded by the European Research Council, project ERC-2014-PoC - 665633_BLOBREC. Norwegian Computing Center (NR) has been subcontractor to University of Tromsø under the management of Eiliv Lund.

NOTE: This is an updated version of the note SAMBA/33/15. We found some errors in the information about screening status that influence two paragraphs in this note, one on page 9 (Section 2.4) , and one on page 18 (Section 4.3, last paragraph). We have updated these two paragraphs using the correct information about screening status.

Keywords	Gene expression test; Breast cancer; Blood; Screening program; Diagnostic test; Naïve Bayes; Fisher test; Preprocessing; Stress; Drug use and smoking;
Target group	Clinical medicine; Systems epidemiology
Availability	Open
Project number	ERC-2014-PoC - 665633_BLOBREC, NR project number 220 732
Research field	MMH, Bioinformatics
Number of pages	36
© Copyright	Norsk Regnesentral

Table of Content

1 Introduction

BLOBREC is a test for distinguishing breast-cancer patients from population-based controls described by Dumeaux et al. in [1]. We will perform a quality control of the methods and procedures used for developing this test.

Besides reproducing results obtained in [1] using exactly the same datasets, we will examine how sensitive the test results are to the approach selected when preprocessing the data, and whether the test results are influenced by drug use, smoking, or stress due to a potential cancer diagnosis. We will also examine if the test results for the breast-cancer patients depend on whether the patient participated in the screening program.

A dataset that can be used for examining the effect of stress is available. This dataset will also be used as a validation dataset for the test developed in [1].

In Section 2 we present the datasets used for developing and verifying the test. Methods are described in Section 3, while results are summarized in Section 4.

2 Data

The datasets used in [1] are described in Section 2.1, while a new dataset that will be used here both for verifying the test in [1] and for examining the effect of stress is presented in Section 2.2. In Sections 2.3 and 2.4 we describe different kinds of background information that is available for the datasets.

2.1 Dataset used for developing the test

After quality control of the data as described in [1], the three datasets CC1, CC2 and CC3 consisted of 55, 49 and 59 case-control pairs, and 39 426, 48 802 and 47 323 probes, respectively. (In [1], the number of probes were 48 803, 48 803 and 47 323, thus for CC1 we have not used the complete set of probes used in [1].)

We have preprocessed the dataset using three different methods A, B and C. Method A is the one used in [1], method B is similar to A, while method C is described in [2]. We denote the dataset obtained using preprocessing methods A, B and C on CC1 as CC1A, CC1B, and CC1C, respectively. Similarly for CC2 and CC3. Methods A, B and C will be described in more detail in the methods section.

2.2 Dataset for examining the effect of stress

The test described in [1] is denoted the BLOBREC test. This test was based on cases diagnosed in different hospital departments and screening units in The Norwegian Breast Cancer Screening program. Controls were sampled randomly from participants in NOWAC matched on time of enrolment and year of birth. They were mailed a letter of invitation together with blood sampling equipment. The design was a hospital based case-control study with matched controls nested in the NOWAC cohort to which the cases belonged. All case-control studies are prone to methodological biases. One claim against the BLOBREC test was that the cases were stressed at time of blood sampling since that was done at the time of the diagnostic biopsy, while the controls had nothing to be anxious about. This could give a systematic difference in gene expression due to stress regulation of the expression.

In order to study this potential bias women arriving at the hospital (NSS) for a second look after the positive findings on a screening mammogram were invited to participate in this methodological study. Women were asked after a second positive mammogram, but before the biopsy to donate blood. At this time the women were under the same stress regardless of the later results from the biopsy. The pathology diagnosis was either normal or malignant. As controls were used women meeting for an ordinary visit in a gynecological out-patient office in the same city.

Blood samples and questionnaires for 40 patients with a biopsy taken were available. Of these 12 women had breast cancer, the others had no malignancies. Forty controls were collected in addition.

On a plate (Illumina) there is 96 positions, one for each individual sample. The cases and controls were matched and they kept together in all laboratory work. The remaining 16 places were filled by a pooled sample based on 16 women with a blood sample collected earlier. Each chip has 12 positions, where ten were used for five case-control pairs and two for the pooled samples.

To summarize, the stress dataset consists of 96 samples. Sixteen samples are pools of many controls, and the remaining 80 samples are from 40 matched case-control pairs. The cases in these pairs are all exposed to stress, while the controls are not exposed. Some of the cases have cancer, the other cases and all the controls are healthy.

The case-control pairs where the cases do not have cancer can be used for measuring the effect of stress, i.e. for identifying genes that are influenced by stress. The entire dataset can be used as a validation set for the test described in [1].

2.3 Information about drug use and smoking for the stress dataset

Information about drug use (Hormone replacement therapy - HRT) and smoking is available for the individuals in the stress dataset. This information is summarized in Table 1 and will be used to examine whether drug use or smoking influence the results of the test.

Table 1 a) The number of individuals that use drug (Yes or No) and the number of individuals that smoke (Yes or No) in the stress dataset. Note that information about drug use is missing for one case with cancer and for one control. b) The number of individuals that use drug (Yes or No) in the CC3 dataset. Note that information about drug use is missing for 28 cases and for one control.

a)	Stressed cases with cancer	12	Drug use	No	11	Smoke	No	8
				Yes	0		Yes	4
				Missing	1		Missing	0
	Stressed cases without cancer	28	Drug use	No	21	Smoke	No	18
				Yes	7		Yes	10
				Missing	0		Missing	0
	Controls	40	Drug use	No	27	Smoke	No	33
				Yes	12		Yes	7
				Missing	1		Missing	0

b)	Cases	59	Drug use	No	29
				Yes	2
				Missing[1]	28
	Controls	59	Drug use	No	46
				Yes	12
				Missing	1

2.4 Information about screening status

All 12 cases in the stress dataset and 29 of the 59 cases in the CC3 dataset participated in the screening program. These 41 individuals belong to the screening group, while the remaining 28 cases in the CC3 dataset belong to the clinical group.[2] For two cases in the CC3 dataset the screening status is unknown. This information can be used to examine whether the test results are influenced by participation in the screening program.

[1] The number of individuals without information about drug use (28) is very high. The files with background information should be examined more closely to check if information about drug use is available for more individuals.

[2] Information about screening status is also available for the CC1 and CC2 datasets. We have not included this information in this note as it has not been used in any analyses.

3 Methods

3.1 Preprocessing the data

The datasets will be preprocessed using three different preprocessing methods. These methods are denoted method A, B and C, respectively.

3.1.1 Method A – original method

Method A is the preprocessing method described and applied in the paper by Dumeaux et al. [1]. Non-present probes are removed, i.e. only probes with detection p-value less than 0.05 in more than 70% of the samples remain in the dataset. The data are transformed using a variance stabilizing technique described in [4], a summary is given in Section 11 (Appendix), and quantile normalized. Finally, the probes are mapped to genes by using the revmap function. When several probes map to the same gene, the average expression of the probes is used. Table 2 in the results section shows the remaining number of probes in the datasets after each preprocessing step.

3.1.2 Method B – altered original method

Method B is an altered version of the original method. It differs from method A in that in method B the data are first background corrected before filtering, transformation and normalization, and the mapping of probes to genes is done differently. The background correction is done using negative control probes, with the nec function in Limma. In method B, we use the function nuID2RefSeqID to obtain the gene symbols for each probe, but as in method A we use the average probe expression value when several probes map to the same gene. Table 3 in the results section shows the remaining number of probes in the datasets after each preprocessing step.

3.1.3 Method C – NR-method

In a previous study [2], we defined a preprocessing procedure that differs from the procedure used in [1]. In our preprocessing method, method C, the data are first background corrected using the nec function in Limma, then probes with poor mapping quality are removed before normalizing between arrays using quantile normalization. The data are thereafter log2-transformed. To remove probes that are not sufficiently present in the dataset, we use a cut-off of 0.01 for the detection p-value, and only probes that are present in at least 40% of the samples (see [2] for details). With a 40% limit, a probe that is present in all cases and not the controls (or vice versa), is included in the dataset which makes it possible to detect probes that are only expressed in one condition. Probes are mapped to genes using the nsFilter function, more details are given in [2], where the probe with the highest interquartile range is chosen if several probes represent the same gene. The gene symbols are found from the function nuID2RefSeqID. Table 4 in the results section shows the remaining number of probes in the datasets after each preprocessing step.

3.2 Finding differentially expressed genes

The Bioconductor R-package Limma (Linear models for microarrays) is used for finding genes that are differentially expressed between two groups, e.g. between cases and controls, between stressed and non-stressed individuals, or between individuals that use drugs and individuals that do not use drug.

3.3 Identifying the 50-gene best predictors

Three datasets, CC1, CC2 and CC3, are used for identifying 50-gene best predictors that separates cases from controls. First CC1 and CC2 are used for finding a set of differentially expressed genes, and then the CC3 dataset is used for defining a predictor from this set of genes.

Genes that are differentially expressed between cases and controls are found using paired linear analysis (Limma, FDR q-value<0.005). The log2-differences of the expression values for each case-control pair were computed and used in the Limma analyses. Genes that are differentially expressed in both CC1 and CC2, and that also are expressed in CC3, are input to the procedure for finding the 50-gene best predictor. In [1], 345 genes were found to be differentially expressed in both CC1 and CC2, and 341 of these were also included in CC3.

The 50-gene best predictor in [1] was found by randomly selecting 100 000 predictors with genes from the 341 genes in CC3. The predictor used is a naïve Bayes classifier (see Section 3.4.1). The predictor with best predictive power, defined as the smallest p-value in a Fisher test, is selected as the 50-gene best predictors. The predictive power of each of the 100 000 predictors is computed using leave-one-out cross validation.

3.4 Predicting group membership

Here we describe two methods for predicting whether an individual belongs to group 0 or group 1. Group 0 can for example consist of individuals without cancer (controls), and group 1 of individuals with cancer (cases). The predictions made by each method are based on data in a training dataset that consists of N individuals and M genes, where each individual is either a case or a control.

Let x_{ij} be the gene expression data on log-scale, $i = 1, \dots, M$ and $j = 1, \dots, N$. In Sections 3.4.1 and 3.4.2 we describe how to predict the group of a new individual with data y_i, $i = 1, \dots, M$.

3.4.1 Naïve Bayes method

A new individual with data y_i, $i = 1, \dots, M$, is predicted to belong to group 1 if

$$\frac{p}{1-p} \prod_{i=1}^{M} \frac{\varphi(y_i; \mu_i^1, \sigma_i^1)}{\varphi(y_i; \mu_i^0, \sigma_i^0)} > 1$$

and to group 0 otherwise, where

- $p = \sum_{j=1,\dots,N} \frac{g_j}{N}$, where g_j =0 if individual j of the training set belongs to group 0, and 1 if individual j belongs to group 1.
- φ is the probability density of the normal density.
- μ_i^1 and σ_i^1 are the mean and standard deviation computed from x_{ij_1}, for all $j_1 \in \{1, \dots, N\}$ such that $g_{j_1} = 1$. Similarly, μ_i^0 and σ_i^0 are the mean and standard deviation computed from x_{ij_0}, for all $j_0 \in \{1, \dots, N\}$ such that $g_{j_0} = 0$.

3.4.2 Method based on standard deviations

This section describes a method for predicting group based on weighted[3] gene expressions (see [5] for a description of the method that includes time). The weights depend on the

[3] Note that these weights can be both positive and negative.

difference in the expected value of the gene expressions relative to the standard deviation in each group. For each gene i we compute the weight as

$$w_i = \frac{\mu_{i,0} - \mu_{i,1}}{\sqrt{\sigma_{i,0}^2 + \sigma_{i,1}^2}}$$

where $\mu_{i,g}$ and $\sigma_{i,g}$ are the expected value and standard deviation for group g of the gene expression $X_{i,j}$. The weight w_i is made such that the sign depends on whether we expect larger/smaller gene expression for group 0 than 1 and the absolute value of w_i is large where we expect the absolute value of this difference to be large.

When predicting the group of a new individual with data y_i, $i = 1, \ldots, M$, we use the variable

$$z = \sum_{i=1}^{m} y_i w_i,$$

where large values indicate group 0 and m is the number of gene expressions that are used. Note that we assume that the variables are sorted such that the sum includes the terms with the largest$|w_i|$ value. The value of m should be at least 20 and may be equal to the number of genes, M. We predict that the new individual belongs to group 0 if $z > 0$. If it is more important to avoid false classification in one of the groups, we may choose another threshold than 0 for z.

3.5 Test difference between two groups based on standard deviation

This method describes a test that finds out whether there is a difference in the gene expression between two groups (see [5] for a description of the method that includes time). If there is a difference, then the gene expressions for the same gene have a smaller standard deviation if all individuals are from the same group than if the individuals are from both groups.

Define τ_i as the sum of the group 0 and group 1 sample standard deviations for the gene expressions for gene i. Further let $\tau_{(i)}$ be the i'th smallest of τ_i. We test the hypothesis:

H0: there is no difference in the gene expression between the groups.

We use $\tau_{(i)}$ as test statistics. The null distribution is obtained by randomizing the data between the different groups.

The test is formed by randomizing the $x_{i,j}$ between the groups, i.e. $x_{i,j}$ is replaced with $x_{i,r(j)}$ where $r(j)$ is a randomization of the individuals. Then we compare the ordered standard deviations $\tau_{(i)}$ in the data relative to the simulated datasets.

4 Results

4.1 Preprocessing the datasets used for developing the test

The datasets have been preprocessed using preprocessing method A (original method), B (altered original method) and C (NR method). See Section 3.1 for more details. Table 2, Table 3 and Table 4 show the remaining number of probes in the datasets after each preprocessing step.

Table 2 Number of remaining probes in datasets CC1, CC2 and CC3 after each step in the preprocessing method A.(original method)

Method A: Preprocessing steps	Remaining probes in dataset		
	CC1	CC2	CC3
1. Remove non-present probes	13 460	10 341	12 519
2. Variance stabilization and normalization			
3. Map probes to genes	9 338	7 898	8 529

Table 3 Number of remaining probes in datasets CC1, CC2 and CC3 after each step in the preprocessing method B (altered original method).

Method B: Preprocessing steps	Remaining probes in dataset		
	CC1	CC2	CC3
1. Background correction			
2. Remove probes not present	13 269	10 342	12 519
3. Variance stabilization and normalization			
4. Map probes to genes	10 260	8 430	9 936

Table 4 Number of remaining probes in datasets CC1, CC2 and CC3 after each step in the preprocessing method C (NR method).

Method C: Preprocessing steps	Remaining probes in dataset		
	CC1	CC2	CC3
1. Background correction			
2. Remove probes with poor mapping quality	30 084	34 361	34 476
3. Quantile normalize between arrays			
4. Remove probes not present	10 336	8 674	10 889
5. Map probes to genes	7 929	6950	7 945

The final datasets after preprocessing are quite different for method A/B and C. Removing probes with poor mapping quality reduces the set of probes before applying the present filtering. Changing the cut-off for the detection p-value or the number of samples for which a probe should be present have a large impact on the number of probes that are left in the dataset. For an overview of the number of common genes for the three different preprocessing methods, see Table 5 in Section 4.3.

4.2 Preprocessing the stress dataset

The stress dataset is preprocessed so that the distribution of the data becomes similar to the distribution of the preprocessed CC3 dataset. This is an advantage as the dataset will be used in a naïve Bayes classifier and as this classifier is based on the mean and standard deviations of the gene expression for each gene for the two groups that are included in the predictor.

For these preprocessing methods B (altered original method) and C (NR method) similar gene-expression distributions to those for CC3 are obtained by first background correcting the data and then keeping the same probes as those that were present in the CC3 dataset before normalization. We then normalize the stress dataset by setting the quantiles of each sample equal to the quantiles obtained for the CC3 dataset in the quantile normalization step for that dataset. For method B, the variance stabilization transform is estimated from the set of probes that are present in the dataset (using the same present criteria as for the CC3 dataset). After normalization we select the same probes as for the CC3 dataset, use the same mapping from probes to genes, and for method B, the expression value for a gene is computed as the average expression of the probes for that gene. We refer to the preprocessed datasets with distributions equal to the CC3 datasets obtained using method B and C, respectively, as the B and C version of stress dataset.

As the quantiles of CC3 is not available for preprocessing method A (original method), we did not preprocess the stress dataset using this method.

The 96 samples of the stress dataset are placed on a plate that contains 8 chips. Two of the 12 arrays on each chip are filled with the pooled samples. The measured gene expression of each of the 16 samples should be similar as the 16 samples are obtained from the same sample. This can be used for examining if the technical variation between chips is too large by comparing the similarity of gene expressions for a pair of pools on the same chip with the similarity of gene expression for two pools on different chips. From the plots in Figure 1 and Figure 2 we conclude that the technical variation between the 8 chips is not too large. Figure 1 shows that pairs including POOL1 or POOL2 have slightly smaller correlations than pairs including only the other 14 pooled samples. Note, however, that all correlations are very close to one. Figure 2 shows that the two first principal components are slightly closer to each other for pools on the same chip.

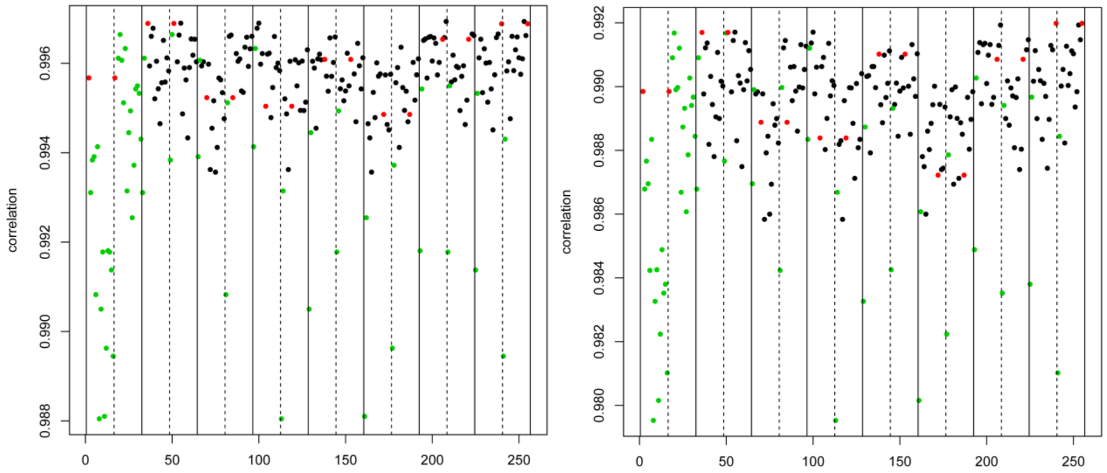


Figure 1 Correlations between gene expression values for each pair of the 16 pooled samples of the stress dataset for data obtained using preprocessing methods B (left panel) and C (right panel). The first column includes all correlations for POOL1, the second all correlation for POOL2 etc. The columns for pairs for pools that are on the same chip are plotted next to each other. Correlations for pools that are on the same chip are plotted in red, while correlations that include one of POOL1 or POOL2 are plotted in green. Each correlation is shown twice, e.g. the correlation between POOL *i* and POOL *j* is shown both in column *i* and in column *j*. Therefore, the red dots with same value are shown next to each other since they are from neighboring POOLs, and only green dots in the first two columns and two green dots in the other columns. We observe that the correlations that include POOL1 or POOL2 tend to be smaller than the other correlations that include only POOL3, …, POOL16.

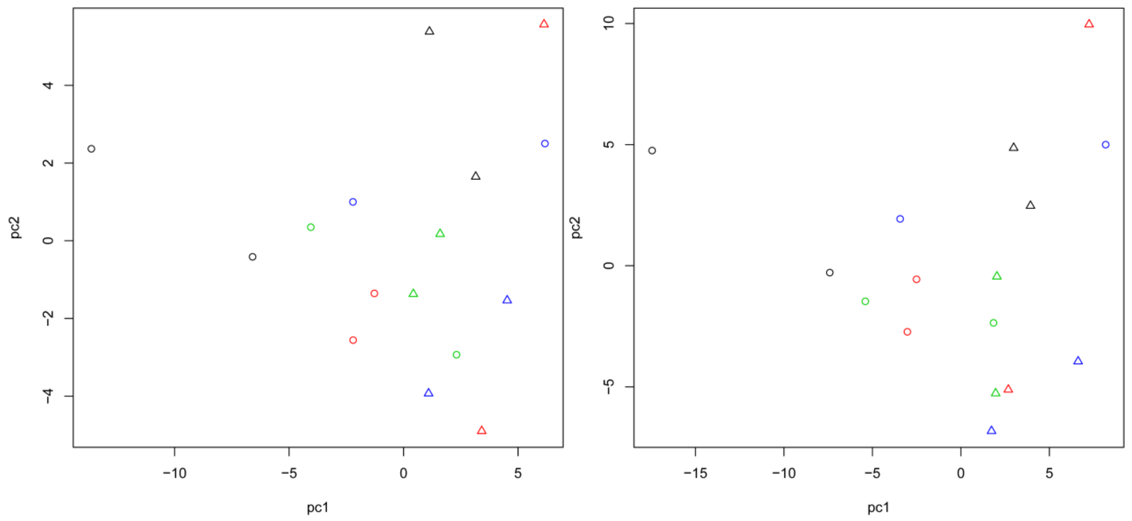


Figure 2 Plots of the first and second principal component of the gene expression for the 16 pooled samples of the stress dataset for data obtained using preprocessing methods B (left panel) and C (right panel). Each pair of pools that are on the same chip is plotted with the same shape and color. The black circles represent POOL1 and POOL2 that are mentioned in Figure 1.

4.3 Differentially expressed genes and the 50-gene best predictor

As described in the data section, we denote the dataset obtained using preprocessing methods A, B and C on CC1 as CC1A, CC1B, and CC1C, respectively. Similarly for the CC2 and CC3 datasets. Repeating the procedure for selecting the differentially expressed genes we found the same 345 genes as were found in [1] when datasets CC1A and CC2A were used. Using datasets CC1B and CC2B, and CC1C and CC2C, we found 369 and 265 genes, respectively. Of these, 317 and 208 were present in the 345-gene set. In CC3B and CC3C, 364 and 263 of the 369 and 265 genes, respectively, were expressed. A summary of common genes for the three different methods of preprocessing the datasets CC1, CC2 and CC3, are given in Table 5.

Table 5 Summary of common genes for the three different methods of preprocessing the datasets CC1, CC2 and CC3. A, B and C denote that datasets are obtained using preprocessing methods A, B and C, respectively.

	Number of genes CC1			Number of genes CC2			Number of genes CC3			Number of genes diff. expr. CC1+CC2			Number of genes diff. expr. CC3		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
A	9338	8667	7304	7898	7374	6364	8529	7966	7067	345	317	208	341	313	205
B	8667	10260	7841	7374	8430	6835	7966	9936	7572	317	369	228	313	364	225
C	7304	7841	7861	6364	6835	6900	7067	7572	7884	208	228	265	205	225	263

From Table 5 we observe that 341 genes are expressed in the CC3A dataset. As the genes are correlated (see Section 7 (Appendix)), we assume that many equally good 50-gene predictors can be built from these 341 genes. The selected 50-gene best predictor is dependent on which genes that are included in the 100 000 predictors that are built by randomly sampling 50 genes from the 341 genes that are expressed in CC3A. We therefore repeated the procedure for selecting the 50-gene best predictor several times and compared the results. More precisely, we repeated the analysis in [1] 99 times with different seeds for sampling the 100 000 predictors from the 341 genes (gene set A1, see Table 6 b)), i.e. 100 analyses in total when including the analysis in [1]. Also, for each of CC3B and CC3C, we repeated the analysis 100 times, once with 364 (gene set B2) and 263 (gene set C2) genes, respectively, and once with 341 genes (gene sets B1 and C1). The results are given in Table 6.

Table 6 a) Prediction results for the 118 individuals in the CC3 dataset when using the same procedure for selecting the 50-gene best predictors as in [1]. Only 15 different prediction results (reported as the number of false negatives (FN), false positives (FP), true negatives (TN) and true positives (TP)) were observed for the 100 different predictors for each of the gene sets A1, B1, B2, C1 and C2. b) Description of the gene sets A1, B1, B2, C1 and C2, i.e. the gene sets the 50-gene best predictors are selected from.

a)		FN	FP	TN	TP	P-value	Number of correct predictions of 100 for gene set				
							A1	B1	B2	C1	C2
		10	20	39	49	4.33e-08	1	2	3	0	0
		11	19	40	48	5.26e-08	0	2	1	0	0
		9	22	37	50	9.00e-08	0	0	0	1	0
		10	21	38	49	1.16e-07	4	7	6	0	1
		11	20	39	48	1.41e-07	26	38	43	1	2
		12	19	40	47	1.65e-07	10	7	14	1	1
		13	18	41	46	1.85e-07	0	0	1	0	0
	Predictor in [1]	10	22	37	49	2.98e-07	13	12	3	8	17
		11	21	38	48	3.66e-07	38	30	29	22	21
		12	20	39	47	4.30e-07	7	2	0	11	25
		13	19	40	46	4.87e-07	0	0	0	3	3
		14	18	41	45	5.30e-07	0	0	0	1	0
		10	23	36	49	7.42e-07	0	0	0	3	8
		11	22	37	48	9.12e-07	1	0	0	43	21
		12	21	38	47	1.08e-06	0	0	0	6	1
	Sum						100	100	100	100	100

b)	Gene sets	A1	B1	B2	C1	C2
	Datasets preprocessed using preprocessing method	A	B	B	C	C
	Gene set = the 341 most significant genes	Yes	Yes	No	Yes	No
	Gene set = the genes with FDR q-value < 0.0005	Yes	No	Yes	No	Yes
	Number of genes in gene set	341	341	364	341	263

We observe that the p-values for A1, B1 and B2 are quite similar. They are also slightly smaller than the p-values for C1 and C2. For a summary of how often each gene is selected for a predictor, see Section 8 (Appendix). The numbers of common genes for each pair of predictors for A1 are shown in Figure 3. With a random selection of 50 genes from the 341 genes, the average overlap between two sets is about seven genes, slightly lower than what is observed in Figure 3. For a summary of the number of times a sample is correctly classified using the 100 different predictors, see Section 10.1 (Appendix).

Table 6 shows that there is a difference in the results between the 100 different predictors, that each is the best of 100.000 simulations. Using more simulations would have resulted in more small p-values. If we for example increased the number of simulations to 1 million simulations, we expect the p-values to vary from 4e-7 to 2e-8, instead of from 1e-6 to 4e-8, at least for gene sets A1, B1 and B2.

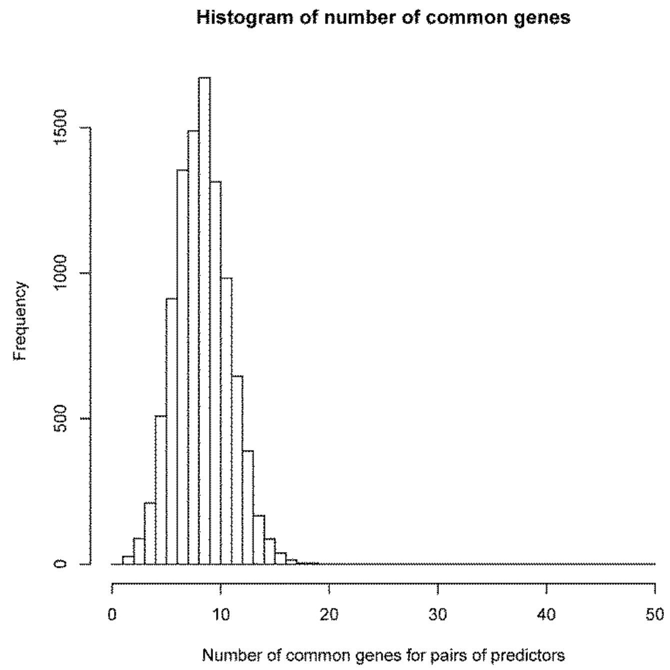


Figure 3 Histogram of number of common genes for each pair of the hundred 50-gene best predictors described in Table 6.

Instead of selecting 50 genes for the predictors we could include all differentially expressed genes in the predictor. Results for this predictor are shown in Table 7 both for the naïve Bayes classifier and for the method based on standard deviations. Note that the data have been normalized to zero mean and standard deviation one for each gene before using the method based on standard deviations. For all five classifiers the leave-one-out approach is used for predicting the status of the individuals in the CC3 dataset. For the naïve Bayes (50 genes) the disease status of an individual *i* is predicted using a 50-gene best predictor that is selected using a dataset consisting of all individuals in the CC3 dataset except individual *i*.

We observe that small p-values are obtained in all the Fisher tests, but that they are larger than the p-values obtained with the 50-gene best predictors shown in Table 6 where we did not use the leave-one-out approach when computing p-values. This result is not surprising as the p-values are computed using the entire CC3 dataset both for estimating the model (i.e. selecting the 50 genes) and for testing the model. Also, a method based on randomly selecting 100 000 predictors and then selecting the predictor with smallest p-value can be over-fitted to the training dataset (CC3). The method based on standard deviations will not result in an over-fitted predictor. Besides, for the naïve Bayes classifier many, equally good 50-gene predictors exist. As we will see later (Section 4.4), these predictors have very different performance for a validation set. If we want to reduce the set of genes included in the predictor from around 250-350 genes to 50 genes, the method based on standard deviation instead of the naïve Bayes classifier, can seem to be a better choice.

Table 7 a) Prediction results for the 118 individuals in the CC3 dataset using leave-one-out prediction. For a description of the gene sets A1, B1, B2, C1 and C2, see Table 6 b). "All genes in gene set" means that all genes in the gene set A1, B1, B2, C1 or C2, respectively, are included in the predictor, while "selected 50 genes" means that 50 genes are selected from the gene set A1, B1, B2, C1 or C2, respectively. For the naïve Bayes method where 50 genes are selected for the predictor, results are shown for two different simulations. b) The number of FN, FP, TN and TP that correspond to each of the p-values reported in table a). The p-values are reported in increasing order.

a)	P-values obtained in a Fisher test	Gene set				
	Method based on	A1	B1	B2	C1	C2
	Naïve Bayes (all genes in gene set)	8.68e-05	4.12e-05	1.90e-04	3.94e-04	3.94e-04
	Naïve Bayes (selected 50 genes, simulation 1)	3.78e-05	2.60e-06	4.09e-04	1.45e-03	3.73e-04
	Naïve Bayes (selected 50 genes, simulation 2)	2.06e-04	8.68e-05	8.05e-04	7.83e-04	2.06e-04
	Standard deviations (all genes in gene set)	2.06e-04	2.06e-04	2.06e-04	7.83e-04	7.83e-04
	Standard deviations (selected 50 genes)	4.09e-04	1.99e-04	1.99e-04	3.94e-04	3.94e-04

b)	p-values from table a) in increasing order and their corresponding number of FN, FP, TN and TP				
	FN	FP	TN	TP	p-value
	12	22	37	47	2.60e-06
	14	23	36	47	3.79e-05
	15	22	37	44	4.12e-05
	15	23	36	44	8.68e-05
	16	23	36	43	1.90e-04
	17	22	37	42	1.99e-04
	18	21	38	41	2.06e-04
	16	24	35	43	3.73e-04
	17	23	36	42	3.94e-04
	18	22	37	41	4.09e-04
	18	23	36	41	7.83e-04
	19	22	37	40	8.05e-04
	18	24	35	41	1.45e-03
	22	22	37	37	4.83e-03

We also tested the method based on standard deviations when including all genes[4] in the CC3 dataset, not only the most differentially expressed as we did in Table 7 above. Then somewhat larger p-values were obtained (4.83e-03 and 8.16e-03). However, an advantage with the approach where all genes in the CC3 dataset are included is that only data from the CC3 dataset is used, while the CC1 and CC2 datasets, or results obtained from these, are not used. This means that less data are needed (one dataset instead of three).

Prediction results summarized with respect to screening status Twenty-nine of the 59 cases in the CC3 dataset participated in the screening program, while the remaining 28 cases did not. For two cases in the CC3 dataset the screening status is unknown. On average (averaging over the 5 x 100 predictors described in Table 6), 21.42 of the 29 cases (74%) that participated in the screening program, and 24.48 of the remaining 28 cases (87%), were correctly classified. The difference in the proportion of correctly classified cases in the two groups is not significant (p-value 0.52 in a Fisher test).

[4] i.e. 8259, 9936 and 7884 genes for preprocessing method A, B and C, respectively.

4.4 Using the stress dataset for verifying the test

We use the B and C version of the stress dataset for the B and C predictors, respectively (see Section 4.2), i.e. the predictors selected from gene sets B1 and B2, and C1 and C2 (see Table 6). For the A predictors, i.e. the predictors selected from gene set A1, we use the B version of stress dataset as preprocessing method B (altered original method) is very similar to preprocessing method A (original method). Note that we mean adjust the CC3A dataset before prediction, so that the mean expression value for each individual is the same as in version B of the stress dataset. Also, note that when using the predictors obtained from the CC3A datasets, there will be less than 50 genes in the predictors as all 50 genes are not present in stress dataset B.

We predict the status (cancer or not cancer) of all individuals/samples in the stress dataset using all 50-gene best predictors described in Table 6. For each predictor the number of FN, FP, TN and TP are found. As before, the predictive power is computed using a Fisher test. As described in the Section 0, the stress dataset consists of the following 96 samples:

- 12 stressed cases with cancer,
- 28 stressed cases without cancer,
- 40 controls and
- 16 pooled samples of controls.

The 16 pooled samples were all obtained from a pooled sample that was based on 16 controls. Hence, these 16 samples are not independent, and therefore their gene expression values cannot be treated as separate measurements in the Fisher test since they depend on each other.

We compute the predictive power for six different subsets of the stress dataset (see Table 8 b)). Summaries of the results for all 50-gene best predictors are found in Table 8, Table 9 and Section 9 (Appendix). For a summary of the number of times a sample is correctly classified using the 100 different predictors, see Section 10.2 (Appendix).

For subset iii), where controls are included, but not the pooled samples and the stressed cases without cancer, we obtain a p-value of 0.02 for the 50-gene best predictor presented in [1]. The median p-values for the other predictors are similar when using the same subset of the stress data set. When we use only the 28 controls that match the stressed cases without cancer (subset iv)) or the 28 stressed cases without cancer (subset v)), significant results are not obtained. Stress due to a potential cancer diagnosis can be a possible explanation for the observed difference between the prediction results for the controls and the stressed cases without cancer.

The results are significant when we include the pooled samples, (subsets i), ii) and vi)). However, these cannot be trusted since the pooled samples are dependent.

Table 8 a) Median (5%-quantile, 95%-quantile) for the p-values of the prediction results for different subsets of the stress dataset using the 50-gene best predictors described in Table 6 (100 predictors for each gene set). For a description of the gene sets A1, B1, B2, C1 and C2, that the 50 genes of each predictor were selected from, see Table 6 b). b) Summary of which samples of the stress dataset that are included in each of the six subdatasets.

		Gene set obtained from the CC3 dataset					Test in [1]
		A1	B1	B2	C1	C2	
a)	i.	0.018 (0.0012, 0.39)	0.033 (0.0035, 0.27)	0.030 (0.0035, 0.27)	0.027 (0.0026, 0.17)	0.020 (0.0035, 0.11)	0.012
	ii.	0.0064 (0.00020, 0.56)	0.011 (0.0015, 0.30)	0.015 (0.00052,0.25)	0.0077 (0.00051, 0.16)	0.0046 (0.00087,0.087)	0.0046
	iii.	0.020 (0.0013, 0.29	0.026 (0.0040, 0.14)	0.025 (0.0028, 0.18)	0.026 (0.0040, 0.13)	0.020 (0.0028, 0.077)	0.020
	iv.	0.072 (0.0090, 0.48)	0.11 (0.017, 0.31)	0.078 (0.017, 0.37)	0.11 (0.025, 0.31)	0.078 (0.025, 0.24)	0.078
	v.	0.12 (0.029, 0.39)	0.22 (0.047, 0.42)	0.18 (0.047, 0.45)	0.22 (0.062, 0.47)	0.19 (0.041, 0.39)	0.12
	vi.	0.0081 (0.00067,0.99)	0.0081 (0.00067,0.91)	0.014 (0.00067,0.81)	0.0025 (0.00016,0.58)	0.0025 (0.00016, 0.67)	0.0025

	Subdataset	
b)	i.	12 stressed cases with cancer, 28 stressed cases without cancer, 40 controls, 16 pooled samples
	ii.	12 stressed cases with cancer, 40 controls, 16 pooled samples
	iii.	12 stressed cases with cancer, 40 controls
	iv.	12 stressed cases with cancer, 28 controls that match a stressed case without cancer
	v.	12 stressed cases with cancer, 28 stressed cases without cancer
	vi.	12 stressed cases with cancer, 16 pooled samples

Table 9 Percent correctly classified samples without cancer in the stress datasets when using the 50-gene best predictors described in Table 6 (100 predictors for each gene set). For a description of the gene sets A1, B1, B2, C1 and C2, that the 50 genes of each predictor were selected from, see Table 6 b).

	Gene set obtained from the CC3 dataset				
Dataset	A1	B1	B2	C1	C2
16 pooled samples	74%	80%	77%	88%	89%
28 stressed cases without cancer	66%	63%	62%	63%	63%
28 controls that match a stressed case without cancer	71%	70%	69%	70%	70%
40 controls	76%	77%	76%	77%	78%

As in Section 4.3 we also test the predictor that includes all differentially expressed genes, i.e. all genes in the gene sets A1, B1, B2, C1 and C2 defined in Table 6. Results are shown in Table 10**.** We observe that the results are significant for method B1 and B2, while the other methods are not able to differentiate between cancer and not cancer. This is difficult to explain. Method B (altered original method) is similar to method A (original method), but for A some genes are omitted and the data are not background corrected. This may explain the difference between A and B. Method C (NR method) uses a log2-transform that gives different normalized values for small values of the gene expressions. This may explain the difference between B and C. It may also be a problem connected with the threshold since all the individuals are classified in the same group for method A and C (NR method).

Table 10 a) Prediction results for the 96 samples in the stress dataset. For a description of the gene sets A1, B1, B2, C1 and C2, see Table 6 b). For a description of the gene sets A1, B1, B2, C1 and C2, see Table 6 b). "All genes in gene set" means that all genes in the gene set A1, B1, B2, C1 or C2, respectively, are included in the predictor, while "selected 50 genes" means that 50 genes are selected from the gene set A1, B1, B2, C1 or C2, respectively. b) The number of FN, FP, TN and TP that correspond to each of the p-values reported in table a). The p-values are reported in increasing order.

	P-values obtained in a Fisher test	Gene set obtained from the CC3 dataset				
	Method based on	A1	B1	B2	C1	C2
a)	Naïve Bayes (all genes in gene set)	1.00	0.019	0.019	1.00	1.00
	Standard deviations (all genes in gene set)	1.00	0.015	0.019	1.00	1.00
	Standard deviations (selected 50 genes)	1.00	0.015	0.024	1.00	1.00

	p-values from table a) in increasing order and their corresponding number of FN, FP, TN and TP				
	FN	FP	TN	TP	P-value
b)	5	19	65	7	0.015
	5	20	64	7	0.019
	5	21	63	7	0.024
	0	84	0	12	1.000
	12	0	84	0	1.000

It is difficult to explain why the results are much weaker for this dataset than for the CC3 dataset. The most likely explanations are probably batch effects between the CC3 and stress dataset or other differences between the two datasets, while effects of stress, drug use or smoking are probably less important. The PCA-plots in Figure 4 show that there is a batch effect.

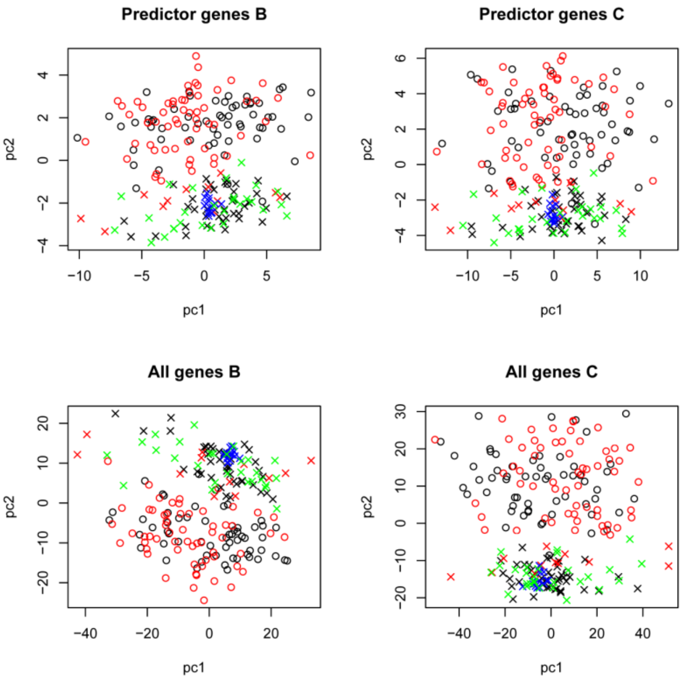


Figure 4 Plots of the first and second principal component of the gene expression for the individuals of the CC3 dataset (circles) and stress dataset (crosses). Cases with cancer are plotted in red, controls in black, cases without cancer in green and pooled samples in blue. In the header of the four plots, B and C denote that the gene expression values are obtained using preprocessing methods B and C, respectively, while "all genes" and "predictor genes" indicate that all genes (9936 genes for B, 7884 genes for C) or only the predictor genes are included when computing the principal components. We define predictor genes to be the genes that are included in at least one of the 50-gene best predictors described in Table 6.

When including all genes[5] in the CC3 dataset, not only the most differentially expressed as in Table 10, for the method based on standard deviation, no significant p-values were obtained.

4.5 Examining the effect of stress

It is possible that stress due to a potential cancer diagnosis can influence the test. In the stress dataset the cases without cancer are exposed to stress, while the controls (without cancer) are not. For examining how stress influence gene expression and whether such an influence by stress has any consequences for the test developed in [1], we will therefore use the case-control pairs of the stress dataset with stressed cases without cancer. We use data obtained with preprocessing method B (altered original method) and C (NR method). See Section 3.1 for more details about the preprocessing methods. We identify genes that are influenced by stress using paired linear analysis (Limma, FDR q-value<0.05). FDR q-values were computed both with respect to all genes[6] and with respect to the genes that are included in at least one of the 50-gene best predictors described in Table 6.

No significant genes were found after multiple testing neither when including all genes nor when including only genes that are present in at least one of the 50-gene best predictors. We cannot conclude that any gene is influenced by stress. However, the dataset is not very large with 28 individuals in each group, so the power of the tests will not be very high.

We observe in Table 9 (Section 4.4) that more stressed cases without cancer than controls are misclassified as individuals with cancer. This may indicate that the gene expression values for some individuals are influenced by stress. However, we are not able to conclude that any gene is significantly differentially expressed between stressed and non-stressed individuals. One possible explanation of this negative result can be that there is a large individual variation in how much the gene expressions are changed due to stress after having received a possible cancer diagnosis. If the individual variation is large, it is also more difficult to find significant differences between the stressed cases without cancer and the (non-stressed) controls.

We have also tested if the two groups are different using a statistic based on standard deviations (Section 3.5). In this case we use data that have been normalized to zero mean and standard deviation one for each gene. No small p-values were observed, except for the 26 genes with lowest standard deviations for stress dataset C. The p-values for these 26 genes are around 0.1. None of the 26 genes are amongst the genes that are included in the 50-gene best predictors described in Table 6.

We also tested if the groups are different based on prediction results obtained using a method based on standard deviation with the leave-one-out approach and a Fisher test (Section 3.4.2). This method was tested both when all genes[7] were included in the dataset, and when only predictor genes were included, where we define predictor genes to be the genes that are included in at least one of the 50-gene best predictors described in Table 6. Results are shown in Table 11. We observe that the p-values are small when all genes are included, but not when only predictor genes are included. This means that the two groups are different, but that we

[5] i.e. 8259, 9936 and 7884 genes for preprocessing method A, B and C, respectively

[6] i.e. 9936 and 7884 genes for preprocessing method B and C, respectively

[7] i.e. 9936 and 7884 genes for preprocessing method B and C, respectively.

are not able to show that they are different for the predictor genes. Also, predictor genes do not tend to have higher absolute weight values than the other genes (see Section 3.4.2 for a definition of the weight of a gene).

Table 11 P-values in a Fisher test for prediction results for stressed cases without cancer from the stress dataset and their matched controls using a method based on standard deviations. B and C denote that datasets are obtained using preprocessing methods B and C, respectively (see Section 3.1 for details). a) Results when only predictor genes are included in the dataset. We define predictor genes to be the genes that are included in at least one of the 50-gene best predictors described in Table 6. b) Results when all genes are included in the dataset, i.e. 9936 genes for dataset B and 7884 genes for dataset C.

a)	Dataset	Number of genes in score for the method based on standard deviations			
		20	50	100	All predictor genes
	B	0.50	0.79	0.70	0.61
	C	0.39	0.86	0.79	0.50

b)	Dataset	Number of genes in score for the method based on standard deviations			
		50	200	2000	All genes (9936 for B and 7884 for C)
	B	0.0076	0.053	0.051	0.085
	C	0.0030	0.141	0.027	0.051

4.6 Examining the effect of drug use and smoking

Drug use and smoking has not been taken into account when developing the test. We have examined whether these exposures influence the results of the test for the stress dataset. For smoking no significant results were obtained (results not shown). The results for drug use, where some are slightly significant, are presented below.

Finding differentially expressed genes First we examined whether there are any differentially expressed genes between those that use and do not use drugs (HRT). We performed separate Limma analyses for each of the groups “Stressed cases with cancer” (12 individuals), “Stressed cases without cancer” (28 individuals), “Controls” (40 individuals) and “Stressed cases without cancer + Controls” (28+40 individuals). No differentially expressed genes were identified except for “Controls” group.

The five genes that were identified as differentially expressed (FDR q-value < 10%, only predictor genes included when computing the FDR q-value) for this group are presented in Table 12. Each of the five genes is included in around 10-25% of the 50-gene best predictors described in Table 6 (100 predictors for each of the five gene set)

Table 12 Differentially expressed genes identified for the 40 controls of the stress dataset. B and C denote that datasets are obtained using preprocessing methods B and C, respectively (see Section 3.1 for details). Note that the HNRNPD gene is not included in predictors selected from the C1 or C2 gene sets (see Table 6 b) for a description of the gene sets C1 and C2).

Gene	FDR q-value	
	B	C
JAK1	0.046	0.058
APP	0.068	0.647
CPEB3	0.078	0.109
KLF13	0.078	0.166
HNRNPD	0.078	-

Prediction results summarized with respect to drug use Table 13 and Figure 5 (left panel) show prediction results for the stress dataset summarized depending on drug use (HRT). The right panel of Figure 5 shows prediction results for CC3. Note that in this case the CC3 dataset has been used both for defining and testing the predictors.

We do not observe any clear tendencies to more or less misclassification due to drug use neither for the stressed cases with cancer, the stressed cases without cancer nor the controls in the stress dataset. There are differences between the groups in the percent of wrongly classified individuals, but the differences are not large. As the dataset is small and few individuals use drugs it is difficult to conclude from these results whether the test can be used independent of drug use and therefore also which individuals that should be excluded from the test due to drug use.

From Figure 5 (left panel) and Table 17 in Section 10.2 (Appendix) we observe that most individuals are either misclassified by most predictors or they are correctly classified by most predictors. This indicates that the ability to predict is not very sensitive to the randomness introduced when selecting genes for the predictor nor to the choice of normalization method.

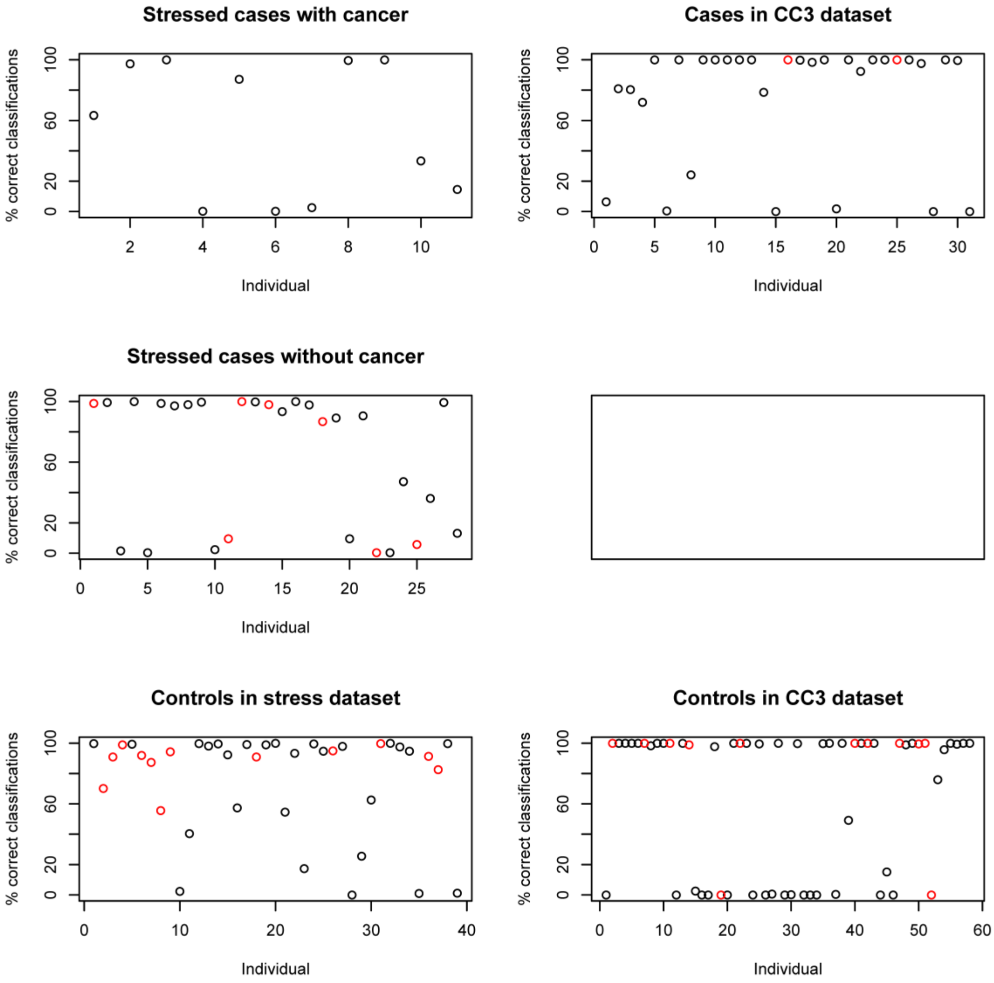


Figure 5 Correspondence between drug use (HRT) and the percent correct classifications (y-axis) as cancer or not cancer of an individual. The individuals (x-axis) are plotted in red if they used drug. The percent correct classification of an individual is computed from the 5 x 100 predictors described in Table 6.

Table 13 Prediction results obtained for the stress dataset summarized depending on drug use. a) Percent wrongly classified using the 50-gene best predictors described in Table 6 (averaged over the 5 x 100 predictors). b) Number of wrongly classified / Number of correctly classified using the 50-gene best predictor in [1]. See Table 1 for a summary of the number drug users for each of the groups "Stressed cases with cancer", "Stressed cases without cancer" and "Controls". c) Percent wrongly classified using the 50-gene best predictors described in Table 6 (averaged over the 5 x 100 predictors) for each of 10 individuals that either use HRT (19), antidepressants (4) or beta blockers (6). Results for antidepressants and beta blockers have been included because use of these drugs was examined in [1].

a)	Percent wrongly classified	Drug use (HRT)	
		No	Yes
	12 stressed cases with cancer	46 %	-
	28 stressed cases without cancer	35 %	48 %
	40 controls	29 %	13 %

b)	Number of wrongly classified / Number of correctly classified	Drug use (HRT)	
		No	Yes
	12 stressed cases with cancer	6 / 5	0 / 0
	28 stressed cases without cancer	4 / 17	3 / 4
	40 controls	6 / 21	0 / 12

c)	Drug	Individuals	Percent wrongly classified
	Hormone replacement therapy (HRT)	Seven stressed cases without cancer	0%, 1%, 2%, 13%, 90%, 94% and 100%
	Hormone replacement therapy (HRT)	Twelve controls	0%, 1%, 5%, 6%, 8%, 9%, 9%, 9%, 13%, 17%, 30% and 44%
	Antidepressants (SSRI)	One stressed case without cancer	60 %
	Antidepressants (SSRI)	Three controls	2 %, 8% and 9%
	Beta blockers (C07AB)	Two stressed cases with cancer	0% and 0%
	Beta blockers C07AB	Four controls	0%, 0%, 0% and 2%

5 Conclusion

In [1] Dumeaux et al. describe a test for distinguishing breast-cancer patients from population-based controls. This note confirms the results in [1]. We are able to predict status for the 118 individuals in the CC3 dataset with a p-value of the order 1e-7. The predictor in [1] is based on simulations, and we therefore repeated the procedure described in [1] so that 100 different predictors were defined. We obtained comparable results for the different predictors, also when using different approaches for preprocessing the data. When using leave-one-out prediction for the 118 individuals in the CC3 dataset, i.e. not defining and testing the predictor on the same data, the p-value increases to the order 1e-4. Also this is significant.

When we predict disease status for a validation dataset (the stress dataset), using the predictors defined based on the CC3 dataset, the results are more varying. Some combinations of preprocessing method, predictor and subset of this dataset are still significant, while others are not. Batch effects and other differences between the two datasets are the most likely explanations for this difference. However, the validation dataset consists of many different subgroups of individuals with a limited number of individuals in each subgroup, making the interpretation uncertain.

We were not able to show that the test is influenced by stress, drug use or smoking, but again, the datasets are too small to draw any firm conclusions.

6 References

[1] Vanessa Dumeaux, Josie Ursini-Siegel, Arnar Flatberg, Hans E. Fjosne, Jan-Ole Frantzen, Marit Muri Holmen, Enno Rodegerdts, Ellen Schlichting and Eiliv Lund1. Peripheral blood cells inform on the presence of breast cancer: A population-based case–control study. Int. J. Cancer: 136, 656–667 (2015).

[2] Clara-Cecilie Günther, Marit Holden, Lars Holden. Preprocessing of gene-expression data related to breast cancer diagnosis. NR note SAMBA/35/14, 2014.

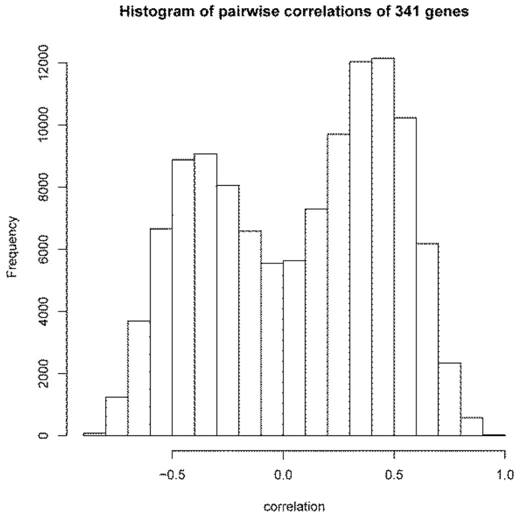
[3] Carlson M. lumiHumanAll.db: Illumina Human Illumina expression annotation data (chip lumiHumanAll). R Package version 1.22.0.

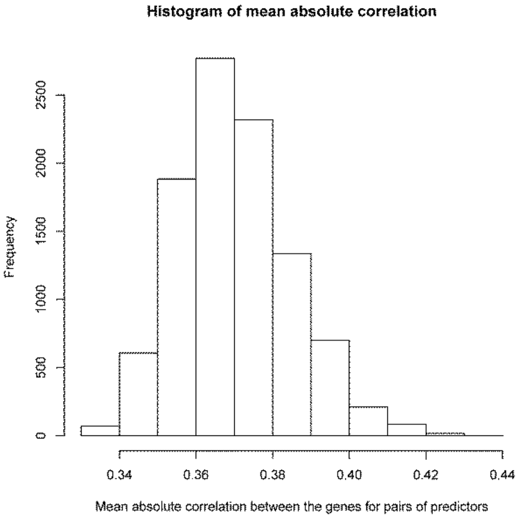
[4] Lin SM, Du P, Huber W, et al. Model-based variance-stabilizing transformation for Illumina microarray data. Nucleic Acids Res 2008;36:e11.

[5] Lars Holden. Time development of gene expression. NR note SAMBA/35/15, 2015.

7 Appendix – correlation between genes (CC3 data)

The mean absolute correlation for the 341 genes is 0.37, while the mean absolute correlation obtained when resampling the data, and thereby removing the correlation, is around 0.08. [8]





[8] The resampled datasets had the same size as the CC3 dataset. Each gene expression value in a resamples dataset was sampled with replacement from the gene expression values of the CC3 dataset. For each resampled dataset the mean absolute correlation was around 0.08

8 Appendix – the 50-gene best predictors (CC3 data)

Table 14 The table shows how many times each of the 341 genes occurs in one of the hundred 50-gene best predictors described in Table 6. For a description of the gene sets A1, B1, B2, C1 and C2, see Table 6 b). * means that the gene is in 50-gene best predictor in [1].

Gene	A1	B1	C1	B2	C2		Gene	A1	B1	C1	B2	C2		Gene	A1	B1	C1	B2	C2	
C18orf8	70	78	63	77	62	*	USP9X	21	22	23	22	29	*	DYNC1LI2	18	4	12	14	26	
GSN	0	0	55	0	62		GMFG	18	18	20	19	29		RBM42	18	24	16	26	0	*
TXNDC12	22	23	56	15	53		TAF15	11	10	29	11	0	*	RPS18	18	24	26	20	23	
ZNF266	41	37	0	47	0		TMEM49	0	22	12	12	29		NDUFB3	9	0	26	7	25	
HIST1H2BK	45	0	28	31	44	*	PYCR2	28	22	16	26	25		CDC2L6	0	13	16	12	26	
EP300	42	26	32	34	29		ENO1	26	22	25	23	28		LOC402057	0	0	23	0	26	
RTN1	22	16	42	16	41		MAP3K1	25	28	17	13	26		HM13	0	0	22	0	26	
TUBA1C	0	41	21	30	25		LY96	22	14	22	16	28		DENR	25	23	0	22	0	
CASP4	24	0	25	23	40	*	CPEB3	21	28	19	22	13		APP	25	19	11	22	6	
ARF3	29	37	34	25	0		C20orf4	21	19	20	23	28		COMT	25	17	14	15	17	
LRFN3	36	23	28	19	0		CTBP1	21	21	20	23	28		PHF5A	19	16	23	18	25	*
RSL24D1	23	26	31	14	36		FAR1	18	20	20	14	28		RPS6KA5	18	11	17	13	25	*
PFN1	32	33	0	35	0		SH2B3	16	15	21	15	28		POGK	17	25	0	15	0	
RBM15	21	22	26	14	35		HNRPM	0	15	20	20	28		PPP1CA	16	12	21	19	25	
AXIN1	14	22	21	18	35		NFATC3	0	0	21	0	28		ARHGDIA	14	8	14	13	25	
TPM3	0	19	19	35	22		JAK1	27	26	20	27	26	*	LOC648024	0	25	0	16	0	
CREB5	0	0	22	0	35		CALHM2	27	24	21	20	26		FLJ10081	0	20	19	15	25	
PPP2R5A	30	21	33	20	29	*	PSMB10	27	16	0	20	0		BHLHB2	0	16	14	16	25	
SUZ12	22	33	16	21	31		RNF4	26	27	26	17	25		ZNF319	24	16	17	12	0	
RNH1	32	18	15	19	21	*	CAPZA2	20	27	20	17	25		ECH1	22	21	15	24	0	
TICAM1	32	19	0	28	0		MAGED1	19	27	14	18	24		LRRFIP1	21	24	0	22	0	
PGAM1	32	17	14	10	14		SLC10A3	17	20	19	13	27		PPM1B	19	24	13	13	0	*
CTNNB1	20	32	0	30	0		EIF3E	17	11	20	15	27		GNAS	19	16	13	24	23	
CTCF	18	20	26	22	32		FNBP1	16	23	17	16	27		RPL5	16	24	16	14	24	
PPP4R1	8	13	23	21	32		EIF4H	14	27	24	22	19		DDIT3	9	12	20	12	24	
SASH3	31	26	0	22	0	*	APEX2	14	12	27	25	25		LOC285900	0	24	0	16	0	
SRC	31	27	23	24	23	*	ATP5B	14	15	14	14	27	*	ZNF20	0	12	18	10	24	
RPS3A	14	17	25	13	31		SRPR	14	19	20	14	27		VCL	0	0	18	0	24	
ASPHD2	30	27	22	25	21		TUBA1B	0	25	22	12	27		KIAA1600	0	0	11	0	24	
CAMLG	22	30	18	21	26	*	HPS6	26	20	18	15	16		CCT7	23	23	11	22	19	
ANXA1	18	23	30	19	29		TAX1BP1	24	23	11	26	21		GPR68	23	15	18	18	20	
HK1	13	19	19	16	30		SP2	22	15	16	13	26		TRAF6	23	23	15	16	14	
CDC40	0	0	30	0	0		GAR1	21	25	0	26	0		SDHA	22	20	18	23	16	*
FRYL	29	18	24	13	27		KEAP1	21	23	9	26	16		CALM1	22	15	14	16	23	
TUBB	29	18	12	22	13		PSMD1	21	21	19	20	26		ERO1L	21	23	0	20	0	

Table continues on next page

Gene	A1	B1	C1	B2	C2		Gene	A1	B1	C1	B2	C2		Gene	A1	B1	C1	B2	C2	
MED24	18	23	13	17	0	*	RPL41	11	15	11	21	16		RHOQ	13	9	18	6	19	
ACTB	17	16	9	23	23	*	RPL11	10	10	14	13	21		BRE	13	0	17	18	19	
ANXA5	16	23	19	15	20		PPP3CA	8	6	21	11	0		ARHGAP17	13	19	9	10	16	
GPR56	15	23	0	17	0		MAPRE1	7	17	17	13	21		ABI3	13	19	7	16	16	
AQP9	14	23	21	17	20	*	LOC650546	0	21	0	20	0		ABR	12	12	12	16	19	
RPL21	14	8	10	10	23		ELMO2	20	13	0	11	0		IK	12	12	7	15	19	
H3F3B	14	21	22	13	23		SF3B2	20	14	9	5	8		FAU	11	6	15	11	19	
AIMP2	13	17	9	16	23		C17orf63	20	11	0	15	0	*	SMARCA4	10	7	14	11	19	
SMAD7	0	0	23	0	12		MYOF	20	0	0	0	0		HNRNPAB	9	10	13	8	19	
FER1L3	0	0	23	0	18		DCAF7	20	10	0	6	0		RRS1	8	18	13	19	17	
CAPRIN1	0	0	19	0	23		TPST2	19	17	17	11	20		HNRNPA3	0	12	10	12	19	
MARCKSL1	22	0	14	16	0		CD74	16	15	20	17	13		CCPG1	0	7	12	14	19	
SMARCAL1	22	16	17	18	0	*	PIGS	15	7	18	14	20		SMAP1	0	0	19	0	0	
SQSTM1	21	15	0	22	0	*	PUM2	15	11	19	14	20		PPFIA1	0	0	19	0	0	
GLRX	21	12	15	20	22		SURF6	14	17	17	20	0		SNORA25	0	0	19	0	18	
HSPBAP1	19	18	22	18	21		COPB2	13	12	19	11	20		MATK	0	0	15	0	19	
NUBP1	17	15	0	22	0		CKAP5	12	18	12	17	20		CUL4B	18	13	15	12	16	
DPH5	16	17	11	7	22		ACTG1	12	20	0	19	0		VMP1	18	0	0	0	0	
ATG12	16	19	0	22	0		HSBP1	11	11	14	9	20		AIFM1	18	14	0	8	0	
MCM3	16	22	0	20	0		OSBPL8	10	7	20	8	19		KARS	18	11	14	9	0	
TMEM131	14	13	19	16	22		MARCH7	7	13	9	11	20		DCTD	18	12	0	14	0	
KIAA2026	13	16	9	10	22		TBC1D15	6	7	12	1	20		S100A8	17	0	18	0	18	
CCDC86	13	19	13	16	22		CSTF2	1	6	14	2	20		RASSF5	17	17	0	18	0	
NKG7	12	12	18	22	0		LOC647856	0	20	0	15	0		GARS	16	18	0	16	0	
C12orf47	9	12	0	22	0		LOC642250	0	20	0	15	0		CRKL	16	18	17	16	14	
U1SNRNPBP	0	22	0	16	0		P15RS	0	15	19	17	20		IL2RB	15	18	0	15	0	
HNRPUL1	0	11	12	13	22		ATP1B3	0	0	20	0	0		SF3B4	15	12	10	13	18	
PCBP1	0	0	22	0	0		ELAC2	0	0	16	0	20		C11orf57	14	12	12	8	18	
MMGT1	0	0	22	0	0		LOC653566	0	0	0	20	0		SURF4	14	18	0	9	0	
LGALS9	0	0	18	0	22		LOC100510589	19	0	0	0	0		IL18BP	13	16	14	7	18	
RPL17	0	0	15	0	22		KIF13B	19	9	12	7	14		SPTLC1	12	18	0	16	0	
TMEM189-UBE2V1	0	0	11	0	22		C16orf72	19	17	15	12	19	*	LRRC33	12	9	11	8	18	
SRP68	21	21	16	18	0		FAM13AOS	19	0	0	0	0		ZNF586	11	18	5	8	16	
ELMO1	18	21	17	10	21		EIF4A3	19	14	10	8	12	*	UBL3	10	18	14	10	16	
TM9SF4	18	9	9	14	21	*	CLPTM1L	18	13	10	17	19		EVI2A	10	11	7	18	13	
ATF5	18	18	17	16	21		MLLT6	18	19	0	10	0	*	FYN	10	11	16	16	18	
DDX19B	15	20	21	17	21		SNRPB	16	7	19	13	0		FAM107B	9	16	10	8	18	
RALA	15	21	15	20	21		ST6GAL1	15	13	13	19	0	*	PTPN6	8	18	6	12	0	
ARHGAP1	14	9	10	9	21		PLAGL2	14	12	13	5	19		RNPS1	8	15	11	18	13	
ARCN1	14	15	17	20	21		LARP1	14	11	19	15	18		SEC23B	6	16	13	9	18	

Table continues on next page

Gene	A1	B1	C1	B2	C2		Gene	A1	B1	C1	B2	C2		Gene	A1	B1	C1	B2	C2	
PPP1CB	5	16	13	3	18		UBE2L6	0	0	11	0	17		HSP90AB1	13	7	15	11	0	
LOC732007	0	18	0	12	0		CD8A	0	0	7	0	17		EWSR1	13	7	0	15	0	
FKSG30	0	10	12	14	18		COPS7B	16	15	0	15	0		CSK	12	10	15	12	15	*
RPL26	0	0	18	0	16		RPL15	16	0	0	0	0		THOC4	11	8	0	15	0	*
EEF1D	0	0	18	0	12		NCOA5	16	0	0	5	0		WBP11	11	14	15	14	13	
ZNF613	0	0	18	0	0		C4orf3	16	0	0	0	0		CLSTN1	10	15	0	2	0	
RPL9	0	0	18	0	16		LOC100290936	16	0	0	0	0		UCP2	10	12	12	4	15	
RPL26L1	0	0	18	0	0		WBP2	15	0	0	16	0		C14orf2	10	0	10	8	15	
CD58	0	0	18	0	18		DHX40	14	15	0	16	0		RARS2	8	0	7	6	15	
ADD1	0	0	15	12	18		ILK	13	0	0	16	0		CAPNS1	8	11	7	15	8	*
RPS29	17	15	9	11	0		RUNX3	13	14	16	10	13		LAMP1	7	15	11	8	0	
SORT1	17	12	11	8	0		GPI	13	15	16	10	14		TNFSF10	7	6	10	15	14	
HNRNPM	17	0	0	0	0		ZMPSTE24	12	14	6	9	16		EIF4A1	0	11	8	13	15	
WDR1	17	14	11	8	15	*	KLF13	12	15	16	16	14		ZCCHC7	0	0	15	0	0	
ALKBH5	17	8	0	10	0		CBARA1	12	14	14	16	0		RPS15A	0	0	15	0	13	
ERP29	17	12	0	17	0		NUP62	11	14	9	7	16		EEF1B2	0	0	15	0	0	
SEC31A	16	17	0	13	0		RPL4	11	10	6	16	16		BAZ2B	0	0	15	0	0	
SEPT9	14	13	0	17	0		CS	10	12	11	11	16	*	RPS14	0	0	10	0	15	
PAPOLA	14	13	10	17	17	*	IGBP1	10	16	14	7	14		SRSF4	14	0	0	0	0	*
C21orf33	13	16	6	16	17		ALG8	8	8	12	12	16		RCC2	14	8	9	12	14	*
TRPV2	13	9	10	9	17		DNAJB1	7	12	13	8	16		GPN2	14	11	0	11	0	
IQGAP1	11	17	13	11	16		IDH2	4	4	5	12	16		PPM1G	14	10	0	8	0	
CCDC92	11	11	14	15	17		RPL13	0	16	8	15	12		YWHAB	14	9	0	11	0	
ZBTB4	11	11	17	15	10		RPL27	0	0	16	0	0		BHLHE40	14	0	0	0	0	
RFTN1	10	12	12	11	17		TPT1	0	0	16	0	15		HNRNPUL1	14	0	0	0	0	*
MGAT4A	8	17	5	4	6		MRPL55	0	0	12	0	16		LLGL1	14	7	9	7	12	*
HNRNPD	8	13	0	17	0		IVNS1ABP	0	0	8	0	16		ATP1A1	13	14	6	8	7	*
NUP93	7	12	12	6	17		ZNF763	15	15	0	6	0		PRPF19	13	7	14	8	12	
LOC402221	0	14	0	17	0		LRCH3	15	14	12	9	11		ATP2B4	12	14	0	9	0	*
SFRS4	0	14	12	12	17		EMP3	15	9	5	12	14		SBK1	12	12	6	12	14	
LOC644063	0	11	0	17	0		RASA3	15	9	10	15	15		MPP5	12	14	0	11	0	
SFRS15	0	10	5	9	17		DHX33	15	13	12	14	13		RGL2	11	13	13	14	14	
ERH	0	0	17	0	0		ANXA2	15	11	15	10	14		VDAC1	11	12	11	14	0	*
NCF1	0	0	17	0	0		HELZ	15	13	9	14	12		ARPC5L	11	10	7	14	11	
ZFHX3	0	0	17	0	0		UBAC2	15	14	12	12	0		CX3CR1	10	7	9	10	14	
RBM4	0	0	17	0	0		PTEN	15	0	0	0	0		EXOC6	10	14	0	10	0	*
ATP5C1	0	0	17	0	0		PSMB2	15	10	8	14	14		CPD	9	6	14	12	11	
CIP29	0	0	17	0	12		PTPN1	15	7	4	11	0		PPRC1	8	0	12	14	0	
RPS17	0	0	17	7	0		TMEM109	14	6	15	9	15		APOBEC3C	8	7	10	5	14	*
RPL23	0	0	14	0	17		CTNNBL1	13	8	0	15	0		XPNPEP1	6	10	7	14	14	

Table continues on next page

Gene	A1	B1	C1	B2	C2	
TMEM39B	5	12	14	6	0	
NSMAF	5	6	8	11	14	
TMEM50B	0	14	0	14	0	
C22orf9	0	13	13	9	14	
RHOT1	0	0	14	0	12	
TROVE2	0	0	14	0	0	
FAM108A1	13	0	0	0	0	
GSTP1	13	0	0	13	0	*
DYNLRB1	13	12	0	11	0	
CECR1	13	9	0	9	0	
APOL3	13	7	0	8	0	
POTEKP	13	0	0	0	0	*
PRF1	12	11	8	13	9	
YARS	11	9	7	6	13	
DSC2	11	0	0	13	0	
TH1L	9	10	0	13	0	
RELA	9	11	6	8	13	
TRIM26	9	13	0	8	0	
ZDHHC7	8	13	0	12	0	*
ZNF598	8	0	0	13	0	
QRICH1	7	13	0	10	0	
ZNF385A	6	5	8	8	13	
LOC643668	0	13	0	7	0	
LOC401152	0	13	0	11	0	
LOC402694	0	13	0	11	0	
LOC650518	0	12	0	13	0	
FOXO1	0	0	13	0	0	
SCRN1	0	0	13	0	0	
sep.06	0	0	13	0	0	
MAPKAP1	0	0	13	0	0	
HNRNPL	0	0	13	0	0	
RAI1	0	0	13	0	0	
KIAA0930	12	0	0	0	0	
NUDC	12	11	0	5	0	
SCAF4	12	0	0	0	0	
GORASP2	12	12	11	7	0	
DPM1	11	7	12	10	0	*
LMNB2	11	6	12	9	8	
DCPS	11	9	7	7	12	
PJA2	10	12	5	8	9	
APBB3	8	12	0	8	0	
HMGCR	7	12	0	8	0	
ITGB2	7	8	0	12	0	*
VPS52	4	11	0	12	0	
LOC648000	0	12	0	8	0	
LOC650276	0	12	0	11	0	
LOC647000	0	12	0	12	0	
LOC731314	0	7	0	12	0	
C5orf5	0	4	12	6	0	
RWDD1	0	0	12	0	0	
PFDN5	0	0	12	0	0	
FAM117B	0	0	12	0	0	
C17orf48	0	0	12	0	0	
LOC391656	0	0	0	12	0	
TPP1	11	10	0	9	0	
RPL7	11	0	0	0	0	
RPRD1A	11	0	0	0	0	*
ING4	11	7	5	9	9	
SRF	10	8	0	11	0	*
RFWD2	10	7	7	6	11	
EXOSC10	8	0	0	11	0	
SRPK1	8	0	11	9	10	
LOC158345	0	11	0	5	0	
LOC644584	0	10	0	11	0	
LOC643446	0	9	0	11	0	
SFRS2IP	0	8	10	6	11	
IPO11	0	0	11	0	0	
GRN	0	0	11	0	0	
RUVBL1	0	0	11	0	11	
LCMT1	0	0	11	0	0	
RALY	0	0	11	0	0	
YPEL5	0	0	9	0	11	
EIF4G1	10	5	7	10	9	
SH3BGRL3	10	8	0	8	0	
H2AFX	10	0	0	0	0	
KIAA1310	10	0	0	0	0	
HBP1	10	9	0	9	0	
FAM127B	9	0	0	10	0	
TERF2	8	6	7	6	10	
CLN5	8	10	0	10	0	
CCDC97	8	9	0	10	0	
GNPDA1	7	7	10	8	8	
ABHD10	5	10	0	6	0	
FAM127A	4	4	10	7	10	
DCP1A	3	6	10	8	0	
GPBAR1	2	4	10	4	9	
LOC642342	0	10	0	7	0	
ETFB	0	0	10	0	8	
ALDOA	0	0	10	0	0	
HMOX1	0	0	10	0	0	
LOC642255	0	0	0	10	0	
XRCC6	8	9	7	9	0	
COBRA1	8	9	4	5	7	
FIP1L1	6	8	8	5	9	
DCP2	6	5	6	3	9	
VPS33A	6	8	0	9	0	
SLC39A3	0	0	9	0	8	
TMEM167B	0	0	9	0	0	
DDX47	0	0	9	0	0	
ZNHIT3	0	0	9	0	0	
MEN1	0	0	9	0	0	
COPE	0	0	9	0	0	
CDK19	8	0	0	0	0	
FAM160B1	8	0	0	0	0	
CNDP2	8	6	0	7	0	
HSDL2	5	3	0	8	0	
TP53INP1	1	7	6	1	8	
SHOC2	0	6	8	4	0	
BAZ1A	0	0	8	0	0	
CENTG3	0	0	8	0	0	
C20orf55	0	0	0	8	0	
SCAF11	7	0	0	0	0	
FAM13B	7	0	0	0	0	
DYNC1H1	5	2	4	3	7	
DAB2	0	0	3	0	7	
HSF1	6	0	0	0	0	
PITRM1	4	5	5	6	5	
PORCN	2	1	3	1	6	
LOC644033	0	6	0	5	0	
STAG2	0	0	6	0	0	
RBM12	0	0	6	0	6	
RPS6KA3	0	0	5	0	6	
TSPAN14	5	0	0	0	0	
TMEM71	5	4	5	4	5	
ARFIP1	5	0	0	0	0	
CDKN1C	1	1	5	2	3	
PLRG1	0	0	3	0	0	

9 Appendix – predicting disease status (stress data)

Table 15 Prediction results for all individuals in the stress dataset, except the stressed cases without cancer, using the 100 different predictors described in Table 6. For a description of the gene sets A1, B1, B2, C1 and C2, see Table 6 b).

					Number of correct predictions of 100										Number of correct predictions of 100				
FN	FP	TN	TP	P-value	A1	B2	C2	B1	C1	FN	FP	TN	TP	P-value	A1	B2	C2	B1	C1
7	3	49	9	3.60E-05	0	0	1	0	0	15	4	41	8	0.011789	3	0	0	0	0
5	4	51	8	6.30E-05	1	0	1	0	1	8	6	48	6	0.012279	1	3	2	4	5
6	4	50	8	0.000137	1	0	0	0	1	20	3	36	9	0.0148	0	1	0	0	0
2	6	54	6	0.000199	1	0	0	0	0	3	8	53	4	0.015466	1	0	0	0	0
10	3	46	9	0.000249	0	1	0	0	0	12	5	44	7	0.015643	2	2	2	0	4
7	4	49	8	0.000275	0	0	1	0	1	16	4	40	8	0.016359	0	2	2	4	0
5	5	51	7	0.000441	1	3	1	3	2	9	6	47	6	0.01859	1	2	3	1	4
8	4	48	8	0.000515	1	2	0	0	1	6	7	50	5	0.019275	0	2	0	4	0
3	6	53	6	0.000545	0	1	0	0	0	21	3	35	9	0.019772	1	0	1	0	1
12	3	44	9	0.000709	0	0	1	0	0	13	5	43	7	0.022023	0	1	0	0	1
6	5	50	7	0.000883	2	2	6	0	4	17	4	39	8	0.022224	1	5	1	0	0
9	4	47	8	0.000907	2	1	3	2	3	27	2	29	10	0.026137	0	1	0	0	0
13	3	43	9	0.001133	0	0	1	0	1	10	6	46	6	0.026986	0	1	2	1	1
4	6	52	6	0.001245	1	0	3	0	2	4	8	52	4	0.027764	2	0	0	0	0
10	4	46	8	0.001522	4	2	2	4	1	18	4	38	8	0.029613	0	5	1	0	0
7	5	49	7	0.001628	3	3	6	2	1	7	7	49	5	0.029785	0	2	0	1	1
14	3	42	9	0.001756	0	0	0	0	1	14	5	42	7	0.030188	0	2	0	3	1
11	4	45	8	0.002448	0	1	2	2	2	11	6	45	6	0.037788	0	1	1	0	2
5	6	51	6	0.002501	5	5	5	7	4	19	4	37	8	0.038761	2	1	0	1	0
8	5	48	7	0.002811	2	2	3	4	1	15	5	41	7	0.040402	0	0	3	1	1
3	7	53	5	0.003169	5	1	0	0	0	24	3	32	9	0.043206	0	0	0	1	0
12	4	44	8	0.003795	4	4	2	3	3	8	7	48	5	0.043579	0	0	0	0	2
21	2	35	10	0.004506	0	1	0	0	0	20	4	36	8	0.049906	1	1	0	3	2
6	6	50	6	0.004561	6	2	8	6	5	12	6	44	6	0.051292	0	2	0	1	0
9	5	47	7	0.004598	1	3	6	2	4	16	5	40	7	0.05292	1	1	0	1	0
17	3	39	9	0.0056	1	0	3	0	0	9	7	47	5	0.060968	0	0	0	0	1
13	4	43	8	0.005693	2	1	0	1	1	21	4	35	8	0.06328	1	2	1	2	0
4	7	52	5	0.006448	5	1	3	1	2	13	6	43	6	0.067745	0	0	0	3	0
10	5	46	7	0.007181	1	1	4	4	5	17	5	39	7	0.067975	0	1	0	0	0
2	8	54	4	0.007378	2	0	0	0	0	26	3	30	9	0.068218	0	0	1	1	1
7	6	49	6	0.007714	3	3	4	6	6	22	4	34	8	0.079101	0	1	2	1	0
18	3	38	9	0.007887	0	1	0	0	1	10	7	46	5	0.082157	0	0	0	0	1
14	4	42	8	0.008299	0	1	2	0	2	18	5	38	7	0.085767	0	0	1	0	2
11	5	45	7	0.010782	1	2	4	3	4	23	4	33	8	0.097564	1	0	0	2	0
19	3	37	9	0.010899	0	0	0	0	1	4	9	52	3	0.098817	0	0	0	1	0
5	7	51	5	0.01165	2	2	0	2	1	19	5	37	7	0.106456	0	0	0	1	0

Table continues on next page

					Number of correct predictions of 100										Number of correct predictions of 100				
FN	FP	TN	TP	P-value	A1	B2	C2	B1	C1	FN	FP	TN	TP	P-value	A1	B2	C2	B1	C1
11	7	45	5	0.107234	1	0	0	0	0	30	4	26	8	0.308496	0	0	1	0	0
15	6	41	6	0.110201	0	0	1	1	1	21	6	35	6	0.312924	0	1	0	0	0
24	4	32	8	0.118835	2	0	0	0	0	31	4	25	8	0.349658	1	1	0	0	0
29	3	27	9	0.124454	1	0	0	0	1	27	5	29	7	0.375806	0	0	1	0	1
20	5	36	7	0.130149	1	1	0	0	0	32	4	24	8	0.392905	1	0	0	2	0
25	4	31	8	0.143037	0	3	0	1	0	38	3	18	9	0.455605	0	1	0	0	0
30	3	26	9	0.148992	1	0	0	0	1	29	5	27	7	0.464526	0	1	0	0	0
21	5	35	7	0.156896	0	1	0	1	1	34	4	22	8	0.483992	1	0	0	0	0
26	4	30	8	0.170244	1	2	1	1	2	30	5	26	7	0.509974	0	0	0	1	0
22	5	34	7	0.186683	0	1	0	0	0	35	4	21	8	0.530857	0	0	0	1	0
27	4	29	8	0.200473	1	0	0	0	1	40	3	16	9	0.555046	1	0	0	0	0
14	7	42	5	0.204912	0	1	0	0	0	32	5	24	7	0.60049	0	1	0	0	0
28	4	28	8	0.23368	2	0	0	1	0	42	3	14	9	0.655715	1	0	0	0	0
33	3	23	9	0.241676	0	1	0	0	0	54	0	2	12	0.676032	1	0	0	0	0
24	5	32	7	0.254969	0	0	1	0	0	39	4	17	8	0.714054	1	0	0	0	0
29	4	27	8	0.26975	2	0	0	0	0	41	4	15	8	0.795088	1	0	0	0	0
39	2	17	10	0.281665	1	0	0	0	0	49	2	7	10	0.810235	1	0	0	0	0
25	5	31	7	0.293086	0	0	0	1	1	42	4	14	8	0.831185	0	0	0	1	0

10 Appendix – correct predictions per sample

10.1 CC3 dataset

Table 16 Number of times a sample in the CC3 dataset is correctly classified using the 100 different predictors described in Table 6. For a description of the gene sets A1, B1, B2, C1 and C2, see Table 6 b).

Sample	A1	B2	C2	B1	C1	Sample	A1	B2	C2	B1	C1	Sample	A1	B2	C2	B1	C1
c_105	0	0	0	0	0	c_158	93	95	97	95	99	bc_124	100	100	100	100	100
c_116	0	0	0	0	0	bc_157	100	95	97	98	98	bc_125	100	100	100	100	100
c_120	0	0	0	0	0	c_122	96	99	98	98	98	c_125	100	100	100	100	100
c_121	0	0	0	0	0	c_112	96	99	98	99	100	bc_126	100	100	100	100	100
c_123	0	0	0	0	0	bc_147	98	97	100	97	100	c_126	100	100	100	100	100
c_124	0	0	0	0	0	bc_143	98	99	99	100	98	bc_127	100	100	100	100	100
c_128	0	0	0	0	0	c_118	98	99	99	99	100	c_127	100	100	100	100	100
c_130	0	0	0	0	0	c_152	100	99	98	100	98	bc_130	100	100	100	100	100
bc_133	0	0	0	0	0	bc_129	99	98	100	99	100	bc_131	100	100	100	100	100
c_133	0	0	0	0	0	c_161	98	99	100	100	100	bc_132	100	100	100	100	100
bc_136	0	0	0	0	0	bc_128	100	100	100	99	99	c_132	100	100	100	100	100
c_136	0	0	0	0	0	c_129	100	100	100	100	98	bc_137	100	100	100	100	100
c_137	0	0	0	0	0	c_154	100	100	98	100	100	bc_138	100	100	100	100	100
c_138	0	0	0	0	0	bc_160	100	100	99	100	99	bc_139	100	100	100	100	100
bc_142	0	0	0	0	0	c_135	99	100	100	100	100	c_140	100	100	100	100	100
c_148	0	0	0	0	0	c_139	100	100	99	100	100	bc_141	100	100	100	100	100
c_150	0	0	0	0	0	bc_146	100	100	100	100	99	c_142	100	100	100	100	100
c_156	0	0	0	0	0	c_106	100	100	100	100	100	bc_144	100	100	100	100	100
bc_158	0	0	0	0	0	c_107	100	100	100	100	100	c_144	100	100	100	100	100
bc_161	0	0	0	0	0	c_108	100	100	100	100	100	bc_145	100	100	100	100	100
bc_163	0	0	0	0	0	bc_109	100	100	100	100	100	c_145	100	100	100	100	100
c_134	0	1	0	0	0	c_109	100	100	100	100	100	c_146	100	100	100	100	100
bc_110	1	0	1	0	0	c_110	100	100	100	100	100	c_147	100	100	100	100	100
c_141	1	0	0	0	1	bc_111	100	100	100	100	100	bc_148	100	100	100	100	100
c_131	0	0	2	0	1	c_111	100	100	100	100	100	bc_150	100	100	100	100	100
bc_149	0	1	3	0	5	bc_113	100	100	100	100	100	bc_151	100	100	100	100	100
c_119	0	4	3	0	6	c_113	100	100	100	100	100	c_151	100	100	100	100	100
bc_105	2	1	15	1	13	bc_114	100	100	100	100	100	bc_153	100	100	100	100	100
c_149	16	25	12	9	14	c_114	100	100	100	100	100	c_153	100	100	100	100	100
bc_112	10	11	48	4	48	bc_115	100	100	100	100	100	bc_154	100	100	100	100	100
c_143	70	72	29	66	9	c_115	100	100	100	100	100	bc_155	100	100	100	100	100
bc_108	73	65	81	65	76	bc_116	100	100	100	100	100	c_155	100	100	100	100	100
c_157	76	91	62	86	65	bc_117	100	100	100	100	100	bc_156	100	100	100	100	100
bc_134	76	85	70	89	73	c_117	100	100	100	100	100	bc_159	100	100	100	100	100
bc_140	86	76	65	91	75	bc_118	100	100	100	100	100	c_159	100	100	100	100	100
bc_135	94	94	56	96	60	bc_119	100	100	100	100	100	bc_162	100	100	100	100	100
bc_107	92	89	70	95	56	bc_120	100	100	100	100	100	c_162	100	100	100	100	100
bc_106	63	78	96	76	92	bc_121	100	100	100	100	100	c_163	100	100	100	100	100
c_160	85	86	80	92	72	bc_122	100	100	100	100	100						
bc_152	96	96	87	95	88	bc_123	100	100	100	100	100						

10.2 Stress dataset

Table 17 Number of times a sample in the stress dataset is correctly classified using the 100 different predictors described in Table 6. For a description of the gene sets A1, B1, B2, C1 and C2, see Table 6 b).

Sample	A1	B2	C2	B1	C1	Sample	A1	B2	C2	B1	C1	Sample	A1	B2	C2	B1	C1
ctrl_38	0	0	0	0	0	POOL12	66	67	81	76	82	ctrl_42	92	97	100	99	100
case_22_syk	1	0	0	0	0	POOL6	70	67	82	72	83	case_35_frisk	95	96	100	98	100
case_27_syk	1	0	0	0	0	POOL14	68	77	83	77	81	case_20_frisk	93	97	100	100	100
case_41_frisk	2	0	0	0	0	POOL9	70	77	89	75	92	case_31_frisk	94	99	99	99	99
case_16_frisk	2	0	0	0	0	ctrl_6	77	68	92	81	95	ctrl_36	92	100	100	100	98
case_43_frisk	1	0	0	1	0	POOL15	75	76	91	79	92	ctrl_22	95	98	100	98	100
ctrl_44	2	1	1	1	0	POOL3	75	79	92	86	89	POOL1	96	100	99	99	100
ctrl_8	6	0	0	0	0	POOL4	80	84	92	86	91	case_1_frisk	97	97	100	100	100
case_11_frisk	7	1	0	0	0	case_38_frisk	82	94	84	90	84	case_17_frisk	97	98	100	99	100
ctrl_19	5	1	1	4	1	case_26_syk	81	85	91	87	92	ctrl_12	98	98	100	99	100
case_23_frisk	10	0	0	0	2	ctrl_16	85	79	93	84	96	ctrl_29	96	99	100	100	100
case_28_syk	6	5	0	2	0	POOL2	88	85	92	90	88	ctrl_27	97	100	100	99	100
case_6_frisk	13	2	3	6	5	POOL7	80	87	91	91	95	case_10_frisk	97	100	100	100	100
case_4_frisk	32	5	5	5	1	POOL13	80	91	95	84	95	case_8_frisk	97	100	100	100	100
case_24_frisk	28	5	3	7	5	POOL5	83	87	96	89	91	ctrl_13	97	100	100	100	100
case_9_frisk	39	10	3	11	3	case_39_frisk	83	89	95	88	91	ctrl_23	98	100	100	100	100
case_5_syk	9	7	26	1	30	POOL10	78	97	92	92	93	case_21_frisk	98	100	100	100	100
ctrl_32	27	12	23	5	20	case_40_frisk	84	80	100	90	99	ctrl_33	98	100	100	100	100
ctrl_39	38	23	21	25	21	ctrl_11	88	92	91	95	89	case_33_syk	99	100	100	99	100
case_42_syk	47	48	19	38	15	ctrl_28	81	93	96	92	93	ctrl_1	99	100	100	100	100
case_7_frisk	52	36	28	39	26	ctrl_5	96	97	91	91	82	ctrl_7	99	100	100	100	100
ctrl_20	48	44	36	42	32	ctrl_14	85	89	100	89	97	ctrl_21	99	100	100	100	100
case_44_frisk	55	45	46	43	47	ctrl_24	89	96	90	98	89	case_30_frisk	99	100	100	100	100
ctrl_30	66	61	48	66	32	ctrl_31	87	96	97	98	89	ctrl_40	99	100	100	100	100
ctrl_17	60	50	51	55	62	case_32_frisk	91	87	100	90	99	ctrl_3	100	100	100	100	100
ctrl_26	50	41	78	39	79	ctrl_18	93	96	93	99	91	case_3_frisk	100	100	100	100	100
POOL16	59	46	74	58	71	ctrl_34	86	94	100	94	100	ctrl_41	100	100	100	100	100
ctrl_4	74	62	56	69	52	ctrl_43	91	90	100	94	99	case_13_frisk	100	100	100	100	100
case_12_syk	63	68	67	61	58	ctrl_35	88	92	100	95	100	case_18_syk	100	100	100	100	100
POOL11	57	52	85	58	78	ctrl_9	93	93	100	95	100	case_29_syk	100	100	100	100	100
ctrl_10	78	64	73	63	73	case_19_frisk	93	95	100	98	100	case_34_frisk	100	100	100	100	100
POOL8	63	65	87	68	82	case_14_syk	94	99	100	98	96	case_36_syk	100	100	100	100	100

11 The variance stabilizing technique

The variance stabilizing technique described in [4] is a transformation of the gene expression from probes defined such that the estimates for the expectation and variance are independent of each other. The transformation is estimated from all the data in an Illumina microarray, which usually have at least 30 replicate measurements for each probe. The transformation is close to a log2-transform except for small values where it results in a larger value. The transform is

$$h(y) = \begin{cases} \frac{1}{c1} \mathrm{arcsinh}(\frac{c2}{\sqrt{c3}} + \mathrm{c1}\ \frac{y}{\sqrt{c3}}) & \text{if } \mathrm{c3} > 0 \\ \frac{1}{c1} \ln(\mathrm{c2}\ + \mathrm{c1}\ \cdot \mathrm{y}) & \text{if } \mathrm{c3} = 0. \end{cases}$$

c3 represent the variance of the background noise that may be estimated from the non-significant measurements. Then c1 and c2 can be estimated from the linear fitting $\sqrt{v(u) - c3} = c1 \cdot u + c2$ where $v(u) = Var(Y)$ is the variance of the data and $u = E(Y)$ is the expectation of the data.