Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

BLOBREC – a blood-based test for breast cancer

Analyses of test properties in the years before and after diagnosis

Note no. SAMBA/27/19

Authors Eiliv Lund, Marit Holden, Lars Holden

Date September 2019

Authors
Eiliv Lund (1)
Marit Holden (2)
Lars Holden (2)

1) UiT The Arctic University of Norway, Tromsø, Norway
2) Norsk Regnesentral, Oslo, Norway

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	**BLOBREC – a blood-based test for breast cancer**
Authors	**Eiliv Lund, Marit Holden, Lars Holden**
Date	September 2019
Year	2019
Publication number	SAMBA/27/19

Abstract

BLOBREC is a test for distinguishing breast-cancer patients from population-based controls described by Dumeaux et al. Previously, we performed a quality control of the methods and procedures used for developing this test, and our analyses confirmed the results obtained by Dumeaux et al.

The aim of these analyses were to study the properties of the BLOBREC test in the years before or after diagnosis, and compare with time of diagnosis. In addition, we looked at effects of parity in controls, and also used a clinical stress study.

We had a case-control design with 539 pairs before diagnosis, 59 at and 429 after diagnosis. In the controls taken from the NOWAC postgenome biobank, we found no difference in percentage false positives (%P) between the pre- and postdiagnostic controls (37% and 34% respectively). The %P were similar to the case-control study at time of diagnosis; 37%. The %P for cases were except for one year before diagnosis similar to the controls pre- and postdiagnostic. Additionally, we found a weak, non-significant increase in %P for controls with many children. The stress data originated from the "second look" at one single centre in the national screening program for breast cancer. The %P (the per cent of positive tests) were lower both for cases and controls than for the original case-control series. These data were collected under more stringent conditions.

Conclusion: the BLOBREC showed higher %P for cases at diagnosis than either before or after, while the %P for controls remained identical. As previous, the %P was higher for metastatic breast cancer.

BLOBREC might be improved through a more stringent sampling of both cases and controls.

Keywords	Gene expression test; Breast cancer; Blood; Screening program; Diagnostic test; Naïve Bayes;
Target group	Clinical medicine; Systems epidemiology
Availability	Open
Project number	NR project number 220 785
Research field	MH, Bioinformatics
Number of pages	17

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]. In the note [2], we performed a quality control of the methods and procedures used for developing the test, and our analyses confirmed the results obtained by Dumeaux et al. The BLOBREC test was developed based on gene expression in blood at the time of diagnosis. In this note, we will use the BLOBREC test on three datasets, one with gene expression in blood before the time of diagnosis, one with gene expression in blood after the time of diagnosis, and one with gene expression in blood at the time of diagnosis.

In Section 2 we describe the BLOBREC test, while the datasets are presented in Section 3. Results are summarized in Section 4.

2 The BLOBREC test

In [1], three datasets with gene expression in blood at the time of diagnosis, CC1, CC2 and CC3, were used for defining the BLOBREC test. First, the CC1 and CC2 datasets were used for finding a set of 345 genes that were differentially expressed in both datasets (FDR q-value < 0.005). The defined set of differentially expressed genes consisted of 345 genes (Appendix 7.1). The CC3 dataset was then used to select 50 of the 345 genes for the predictor that separates cases from controls (Appendix 7.2).

The BLOBREC test predicts cancer or not using a Naïve Bayes method. The predictions made are based on data in the CC3 dataset that consists of N=118 individuals and $M = 50$ genes, where each individual is either a case or a control. In the CC3 dataset, there are 59 cases and 59 controls.

Let x_{ij} be the gene expression data on log-scale, $i = 1, \dots, M$ and $j = 1, \dots, N$. Let group 0 consist of individuals without cancer (controls), and group 1 of individuals with cancer (cases). A new individual with data y_i, $i = 1, \dots, M$, is predicted to be a case, i.e. to have cancer, 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 be a control, i.e. to be without cancer, otherwise, where

- $p = \sum_{j=1,\dots,N}\frac{g_j}{N}$, where g_j =0 if individual j of the CC3 dataset 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 $g_{j_0} = 0$.

Note that if some of the 50 genes are not included in the dataset with the new individual these genes are excluded from the test. Also note that before using the test on data from a new dataset, the mean and standard deviations of the test are adjusted so that the new mean and

standard deviations for the controls in the BLOBREC test becomes equal to the mean and standard deviations for the controls in the new dataset. More precisely, we compute the adjusted mean and standard deviations μ_i^{0*}, μ_i^{1*}, σ_i^{0*} and σ_i^{1*} as:

$$\mu_i^{0*} = \mu_i^0 - \mu_i^0 + \mu_i^{0N}, \quad \mu_i^{1*} = \mu_i^1 - \mu_i^0 + \mu_i^{0N}, \quad \sigma_i^{0*} = \sigma_i^0 \cdot \frac{\sigma_i^{0N}}{\sigma_i^0}, \quad \text{and} \quad \sigma_i^{1*} = \sigma_i^1 \cdot \frac{\sigma_i^{0N}}{\sigma_i^0},$$

where μ_i^{0N} and σ_i^{0N} are the sample mean and standard deviation, respectively, for gene i for the controls in the new dataset.

We will also define a new test, the extended BLOBREC test, which is equal to the BLOBREC test except that it is based on the set of 345 genes instead of only 50 genes.

3 Data

Three breast cancer datasets are available: Two of them were based on the NOWAC postgenome biobank; the prospective dataset and the postdiagnostic dataset. The stress dataset was collected at the "second look" in one clinic. Each of these datasets will be described in more detail below.

3.1 Preprocessing the gene expression datasets

Each dataset was background corrected using negative control probes, $\log_2$ transformed using a variance stabilizing technique [2], and quantile normalized. We retained probes present in at least 70% of the samples. If a gene was represented with more than one probe, the average expression of the probes was used as expression value for the gene. The probes were translated to genes using the lumiHumanIDMapping [3].

3.2 The prospective dataset

After removing technical outliers and controls that were diagnosed with cancer, the dataset consists of data from 539 case-control pairs that are from three different runs / batches. After preprocessing the dataset consists of 8155 genes. The cases are diagnosed with metastases, without metastases or with in-situ tumors. Note that the adjusted mean and standard deviations, μ_i^{0*}, μ_i^{1*}, σ_i^{0*} and σ_i^{1*}, are computed separately for each of the three different runs.

3.3 The postdiagnositic dataset

After removing technical outliers and controls that were diagnosed with cancer, the dataset consists of data from 429 case-control pairs. After preprocessing the dataset consists of 8400 genes. We updated the metastasis and follow-up time information for cases that were diagnosed with metastases (6) or new breast cancer (10) after the initial/first breast cancer diagnosis and before the blood sampling[1]. After removing 7 case-control pairs where the case was diagnosed with other cancers after the initial/first breast cancer diagnosis and before the

[1] For 17 of the 429 cases such information was not available.

blood sampling[2], and 7 cases where the metastases status of the case was unknown, 415 cases remained in the dataset.

3.4 The stress dataset

The stress dataset was produced for examining the effect of stress when the women returned for a "second look" after positive findings in the ordinary mammographic screening. In this note we will use the dataset as a validation set for the BLOBREC test. The blood sampling procedure is such that all cases in the stress dataset 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.

The stress dataset consisted of 40 case-control pairs and 47 323 probes. In this dataset some cases have cancer (12), while the remaining cases (28) and controls (40) are healthy. In addition to the 40 case-control pairs, the stress dataset also contains 16 samples that are selected from a pooled sample based on blood samples from 16 individuals. The data are preprocessed using the procedure described in Section 3.1, but adapted so that the probes present in the CC3 dataset were used when mapping from probes to genes. After preprocessing, the dataset consisted of the same 9 936 genes as the CC3 dataset.

4 Results

Table 1 below shows prediction results for the CC3 dataset. These results are taken from Table 7 in [2]. When 50 of the 341 genes are included in the test, 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*.

Table 1 *Prediction results for the 59 cases and 59 controls in the CC3 dataset using leave-one-out prediction both when selecting genes for the test and when computing mean and standard deviation. The results are taken from Table 7 in [2]."N" is the number of negative tests, i.e. the number of individuals that are classified as not cancer, while "P" is the number of positive tests, i.e. the number of individuals that are classified as having cancer. "%P" is the per cent of positive tests, i.e. P/(P+N).*

	Number of genes included in test	N	P	%P
Controls	The 341 of the 345 genes present in the CC3 dataset	37	22	37
	50 of 341 genes[3], simulation 1	37	22	37
	50 of 341 genes, simulation 2	36	23	39
Cases	The 341 of the 345 genes present in the CC3 dataset	15	44	75
	50 of 341 genes, simulation 1	12	47	80
	50 of 341 genes, simulation 2	15	44	75

Tables 2-6 show prediction results for the prospective, postdiagnostic and stress datasets. For each of the three datasets, we give results for tests with the 50 and 345 genes that were identified in [1]. Only genes that are included in the preprocessed dataset are used in the tests.

[2] For 17 of the 429 cases such information was not available.

[3] The 50 genes have been selected as described in [1] and [2]

Table 2 *Prediction results for the 539 cases and 539 controls in the prospective dataset. "N","P" and "%P" are defined as in Table 1. Table 6 show the prediction results stratified on screening status (screen-detected or clinically detected cancer).*

Controls	N	P	%P	The BLOBREC test (41 genes included in the test)					
	338	201	37						
Cases	In-situ			Without Metastases			With Metastases		
	N	P	%P	N	P	%P	N	P	%P
Year 1	17	4	19	40	24	38	10	12	55
Year 2	7	4	36	47	20	30	19	8	30
Year 3	12	6	33	39	27	41	19	7	27
Year 4	10	10	50	50	17	25	24	11	31
Year 5	4	4	50	27	14	34	15	4	21
Year 6	1	0	0	5	8	62	3	4	57
Year 7	0	0	0	1	3	75	1	0	0
Year 8	0	0	0	1	0	0	0	0	0

Controls	N	P	%P	The extended BLOBREC test (300 genes included in the test)					
	342	197	37						
Cases	In-situ			Without Metastases			With Metastases		
	N	P	%P	N	P	%P	N	P	%P
Year 1	17	4	19	41	23	36	10	12	55
Year 2	6	5	45	45	22	33	19	8	30
Year 3	12	6	33	39	27	41	17	9	35
Year 4	10	10	50	52	15	22	27	8	23
Year 5	4	4	50	27	14	34	14	5	26
Year 6	1	0	0	5	8	62	4	3	43
Year 7	0	0	0	1	3	75	1	0	0
Year 8	0	0	0	0	1	100	0	0	0

The %P of controls were identical to the %P of controls in CC3. In situ had similar %P as the controls. For metastatic cases the %P were highest for metastatic cases. There was no effect of the extended BLOBREC test with regard to test properties.

Table 3 *Prediction results for the 429 cases and 429 controls in the postdiagnostic dataset. "N","P" and "%P" are defined as in Table 1. We know the metastasis status (in-situ, without metastases, with metastases) for 415 of the 429 cases.*

Controls	N	P	%P	The BLOBREC test (42 genes included in the test[4])					
	282	147	34						
Cases	In-situ			Without Metastases			With Metastases		
	N	P	%P	N	P	%P	N	P	%P
Year 1	5	2	29	17	16	48	7	6	46
Year 2	4	3	43	29	25	46	14	10	42
Year 3	6	3	33	21	18	46	10	12	55
Year 4	6	3	33	19	13	41	13	9	41
Year 5	4	0	0	11	8	42	7	6	46
Year 6	4	2	33	11	7	39	11	9	45
Year 7	5	1	17	15	8	35	11	11	50
Year 8	1	0	0	6	4	40	0	2	100

Controls	N	P	%P	The extended BLOBREC test (301 genes included in the test[5])					
	278	151	35						
Cases	In-situ			Without Metastases			With Metastases		
	N	P	%P	N	P	%P	N	P	%P
Year 1	5	2	29	16	17	52	8	5	38
Year 2	5	2	29	27	27	50	12	12	50
Year 3	6	3	33	25	14	36	10	12	55
Year 4	5	4	44	23	9	28	13	9	41
Year 5	4	0	0	11	8	42	9	4	31
Year 6	5	1	17	12	6	33	11	9	45
Year 7	5	1	17	17	6	26	11	11	50
Year 8	1	0	0	6	4	40	0	2	100

The %P for controls in the postdiagnostic dataset were similar to the two previous series. In situ was similar to the controls. There was no relationship between time after diagnosis and percent positives.

[4] These 301 include the 42 genes for the prospective dataset and in addition the gene ZNF266.

[5] These 301 include the 300 genes for the prospective dataset and in addition the gene ZNF266.

Table 4 *Prediction results for controls in the prospective and postdiagnostic dataset stratified on parity. "N"," "P and "%P" are defined as in Table 1.*

41/42 genes included in the test	539			429			968		
	Prospective			Postdiagnostic			Both		
Parity	N	P	%P	N	P	%P	N	P	%P
0	59	24	29	34	19	36	93	43	32
1-3	259	154	37	223	116	34	482	270	36
4-6	24	19	44	25	12	32	49	31	39

The additional analyses of the %P among controls according to parity showed a weak, non-significant trend with higher parity (Fisher's test, p>0.3).

Table 5 *Prediction results for the 12 cases with cancer, 28 individuals with biopsy, but without cancer, and the 16 pooled samples, the 40 controls, in the stress dataset. ."N"," "P and "%P" are defined as in Table 1.*

	The BLOBREC test (48 genes included in the test)			The extended BLOBREC test (327 genes included in the test)		
	N	P	%P	N	P	%P
Pools	15	1	6	16	0	0
Controls	34	6	15	35	5	12
With biopsy, without cancer	19	9	32	20	8	29
Cases with cancer	5	7	58	5	7	58

The stress test differs fro0m the other series as it was collected in one surgical department. The pooled controls showed a %P equal to zero. The controls had a much lower %P than in the previous material, but the number is small. The %P of the cases was lower than in the case-control study, which can be due to the small tumours found in the screening.

5 Conclusion

The analyses has demonstrated a high %P for controls taken from the NOWAC postgenome biobank leaving the test difficult to use. However, the stress test %P could indicate that in a clinical situation the test might prove to work fairly well.

The prediction results (Section 7.3) demonstrates that the BLOBREC primarily is a test for metastatic clinical cancer.

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 Lund. Peripheral blood cells inform on the presence of breast cancer: A population-based case–control study. Int. J. Cancer: 136, 656–667 (2015).

[2] Marit Holden, Clara-Cecilie Günther and Lars Holden. Verification of a blood-based test for breast cancer (BLOBREC): Distinguishing breast-cancer patients from population-based controls. NR note SAMBA/33/15, 2015.

7 Appendix

7.1 The set of 345 differentially expressed gene

ABHD10 ABI3 ABR ACTB ACTG1 AGPAT3 AIFM1 AIMP2 ALG8 ALKBH5 ANXA1 ANXA2 ANXA5 APBB3 APEX2 APOBEC3C APOL3 APP AQP9 ARCN1 ARF3 ARFIP1 ARHGAP1 ARHGAP17 ARHGDIA ARPC5L ASPHD2 ATF5 ATG12 ATP1A1 ATP2B4 ATP5B AXIN1 BHLHE40 BRE C11orf57 C12orf47 C14orf2 C16orf72 C17orf63 C18orf8 C20orf4 C21orf33 C4orf3 CALHM2 CALM1 CAMLG CAPNS1 CAPZA2 CASP4 CBARA1 CCDC86 CCDC92 CCDC97 CCT7 CD74 CDK19 CDKN1C CECR1 CKAP5 CLN5 CLPTM1L CLSTN1 CNDP2 COBRA1 COMT COPB2 COPS7B CPD CPEB3 CRKL CS CSK CSTF2 CTBP1 CTCF CTNNB1 CTNNBL1 CUL4B CX3CR1 DCAF7 DCP1A DCP2 DCPS DCTD DDIT3 DDX19B DENR DHX33 DHX40 DNAJB1 DPH5 DPM1 DSC2 DYNC1H1 DYNC1LI2 DYNLRB1 ECH1 EIF3E EIF4A3 EIF4G1 EIF4H ELAC2 ELMO1 ELMO2 EMP3 ENO1 EP300 ERO1L ERP29 EVI2A EWSR1 EXOC6 EXOSC10 FAM107B FAM108A1 FAM127A FAM127B FAM13AOS FAM13B FAM160B1 FAR1 FAU FIP1L1 FNBP1 FRYL FYN GAR1 GARS GLRX GMFG GNAS GNPDA1 GORASP2 GPBAR1 GPI GPN2 GPR56 GPR68 GSTP1 H2AFX H3F3B HBP1 HELZ HIST1H2BK HK1 HMGCR HNRNPAB HNRNPD HNRNPUL1 HNRPM HPS6 HSBP1 HSDL2 HSF1 HSP90AB1 HSPBAP1 IDH2 IGBP1 IK IL18BP IL2RB ILK ING4 IQGAP1 ITGB2 JAK1 KARS KEAP1 KIAA0930 KIAA1310 KIAA2026 KIF13B KLF13 LAMP1 LARP1 LLGL1 LMNB2 LOC100290936 LOC100510589 LRCH3 LRFN3 LRRC33 LRRFIP1 LY96 MAGED1 MAP3K1 MAPRE1 MARCH7 MARCKSL1 MCM3 MED24 MGAT4A MLLT6 MPP5 MYOF NCOA5 NDUFB3 NKG7 NSMAF NUBP1 NUDC NUP62 NUP93 OSBPL8 PAPOLA PFN1 PGAM1 PHF5A PIGS PITRM1 PJA2 PLAGL2 POGK PORCN POTEKP PPM1B PPM1G PPP1CA PPP1CB PPP1R9B PPP2R5A PPP3CA PPP4R1 PPRC1 PRF1 PRPF19 PSMB10 PSMB2 PSMD1 PTEN PTPN1 PTPN6 PUM2 PYCR2 QRICH1 RALA RARS2 RASA3 RASSF5 RBM15 RBM42 RCC2 RELA RFTN1 RFWD2 RGL2 RHOQ RNF4 RNH1 RNPS1 RPL11 RPL15 RPL21 RPL4 RPL41 RPL5 RPL7 RPRD1A RPS18 RPS29 RPS3A RPS6KA5 RRS1 RSL24D1 RTN1 RUNX3 S100A8 SASH3 SBK1 SCAF11 SCAF4 SDHA SEC23B SEC31A SEPT9 SF3B2 SF3B4 SH2B3 SH3BGRL3 SLC10A3 SMARCA4 SMARCAL1 SNRPB SORT1 SP2 SPTLC1 SQSTM1 SRC SRF SRP68 SRPK1 SRPR SRSF4 ST6GAL1 SURF4 SURF6 SUZ12 TAF15 TAX1BP1 TBC1D15 TCF7L2 TERF2 TH1L THOC4 TICAM1 TM9SF4 TMEM109 TMEM131 TMEM39B TMEM71 TNFSF10 TP53INP1 TPP1 TPST2 TRAF6 TRIM26 TRPV2 TSPAN14 TUBB TXNDC12 UBAC2 UBL3 UCP2 USP9X VDAC1 VMP1 VPS33A VPS52 WBP11 WBP2 WDR1 XPNPEP1 XRCC6 YARS YWHAB ZBTB4 ZDHHC7 ZMPSTE24 ZNF266 ZNF319 ZNF385A ZNF586 ZNF598 ZNF763

7.2 The 50 genes included in the BLOBREC test

ABHD10 AIMP2 APOL3 ARHGAP1 ARHGAP17 C14orf2 C16orf72 C18orf8 CAMLG CD74 CECR1 COPS7B DCP1A DENR DHX40 DYNLRB1 EP300 FIP1L1 FRYL GLRX GMFG GSTP1 HNRNPAB HPS6 IK JAK1 KARS KIAA0930 KIAA1310 LOC100290936 LRFN3 PAPOLA PHF5A PPP2R5A RASSF5 RPL21 RPL5 RPS3A S100A8 SDHA SEC31A SMARCAL1 SP2 TMEM39B TUBB UBAC2 WBP11 YWHAB ZNF266 ZNF319

7.3 Prediction results stratified on screening status

Table 6 *Prediction results for the 539 cases and 539 controls in the prospective dataset. "N"," "P and "%P" are defined as in Table 1.*

Controls	N	P	%P	The BLOBREC test (41 genes included in the test)														
	338	201	37															
Cases	Screening									Clinical								
	In-situ			Without Met.			With Met.			In-situ			Without Met.			With Met.		
	N	P	%P	N	P	%P	N	P	%P	N	P	%P	N	P	%P	N	P	%P
Year 1	12	4	25	33	19	37	8	4	33	5	0	0	8	4	33	2	8	80
Year 2	4	5	56	40	16	29	13	6	32	2	0	0	5	6	55	6	2	25
Year 3	10	5	33	33	25	43	16	8	33	2	1	33	6	2	25	1	1	50
Year 4	10	9	47	40	13	25	24	6	20	1	0	0	12	2	14	3	2	40
Year 5	3	4	57	23	12	34	10	5	33	1	0	0	4	2	33	4	0	0
Year 6	0	0	0	4	6	60	4	2	33	1	0	0	1	2	67	1	0	0
Year 7	0	0	0	1	2	67	1	0	0	0	0	0	1	0	0	0	0	0
Year 8	0	0	0	0	1	100	0	0	0	0	0	0	0	0	0	0	0	0

Controls	N	P	%P	The extended BLOBREC test (300 genes included in the test)														
	342	197	37															
Cases	Screening									Clinical								
	In-situ			Without Met.			With Met.			In-situ			Without Met.			With Met.		
	N	P	%P	N	P	%P	N	P	%P	N	P	%P	N	P	%P	N	P	%P
Year 1	12	4	25	32	20	38	8	4	33	5	0	0	8	4	33	2	8	80
Year 2	5	4	44	43	13	23	13	6	32	2	0	0	4	7	64	6	2	25
Year 3	10	5	33	33	25	43	18	6	25	2	1	33	6	2	25	1	1	50
Year 4	10	9	47	38	15	28	21	9	30	1	0	0	12	2	14	3	2	40
Year 5	3	4	57	23	12	34	11	4	27	1	0	0	4	2	33	4	0	0
Year 6	0	0	0	4	6	60	3	3	50	1	0	0	1	2	67	1	0	0
Year 7	0	0	0	1	2	67	1	0	0	0	0	0	1	0	0	0	0	0
Year 8	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

7.4 Selecting 10 of 50 genes for the predictor

The CC3 dataset is preprocessed as described in [1]. In [2] this preprocessing method is denoted preprocessing method A and differs only slightly from the preprocessing procedure described in Section 3.1.

Table 7 Prediction results for the 118 individuals in the CC3 dataset.

	FN	FP	TN	TP	P-value
50-gene best predictor in [1] (result reported previously in Table 6 a) in [2])	10	22	37	49	2.98e-07
10-gene best predictor (genes selected from the 50-gene best predictor using the same procedure as was used in [1] for the 50-gene best predictor)	9	21	38	50	3.38e-08

Table 8 Prediction results for the 118 individuals in the CC3 dataset using leave-one-out prediction. Results are shown for five different simulations.

	FN	FP	TN	TP	P-value
Naïve Bayes (selected 50 of 345 genes, simulation 1) Result reported previously in Table 7 in [2]	14	23	36	45	3.78e-05
Naïve Bayes (selected 50 of 345 genes, simulation 2) Result reported previously in Table 7 in [2]	18	21	38	41	2.06e-04
Naïve Bayes (selected 10 of 50 genes, simulation 3)	17	27	32	42	4.30e-03
Naïve Bayes (selected 10 of 50 genes, simulation 4)	19	23	36	40	1.49e-03
Naïve Bayes (selected 10 of 50 genes, simulation 5)	19	21	38	40	4.19e-04

Note that the 10 genes that are selected for the predictor (Table 3) define one of many possible «10-gene best predictors» as this predictor is selected as the best of 100 000 randomly chosen predictors (we choose 10 out of 50 genes 100 000 times so that we obtain 100 000 different predictors). Choosing one predictor from 100 000 other randomly chosen predictors, will result in another «10-gene best predictors» (i.e. we cannot expect that the same 10 genes as in Table 3 are selected for the predictor).

Table 9 This table is a simplified version of Additional file 3 from [1] where only rows with the 50 genes of the 50-gene best predictor are included. The 10 genes selected from these 50 genes are highlighted in yellow.

Gene Symbol	Gene Name	Keywords	logFC CC1	logFC CC2	logFC CC3	FDR CC3
DYNLRB1	dynein, light chain, roadblock-type 1	ER vesicles transport, tumorigenesis	-0.18	-0.12	-0.12	0.01
TUBB	tubulin beta	cytoskeleton,cell shape, cell cycle	-0.20	-0.14	-0.12	0.01
DHX40	DEAH (Asp-Glu-Ala-His) box polypeptide 40	regulation transcription, helicase	0.13	0.08	0.07	0.03
PAPOLA	poly(A) polymerase alpha	RNA processing, polyA	0.42	0.21	0.17	0.03
KARS	lysyl-tRNA synthetase	regulation transcription, immune, monocyte/macrophage	-0.09	-0.12	-0.08	0.03
CAMLG	calcium modulating ligand	immune, T cell, signalling	0.26	0.20	0.10	0.03
DCP1A	DCP1 decapping enzyme homolog A (S. cerevisiae)	RNA metabolism, decay, tgfb	-0.22	-0.16	-0.08	0.07
GSTP1	glutathione S-transferase pi 1	metabolism, xenobiotic, tumorigenesis	-0.27	-0.18	-0.10	0.08
ZNF266	zinc finger protein 266	regulation transcription, zinc-finger	0.16	0.20	0.12	0.08
RPS3A	ribosomal protein S3A	translation, erythropoeisis, ribosome	0.87	0.64	0.30	0.08
RPL5	ribosomal protein L5	translation, ribosome	0.37	0.29	0.16	0.09
DENR	density-regulated protein	translation, tumorigenesis	0.11	0.12	0.07	0.10
GLRX	glutaredoxin (thioltransferase)	Metabolism	0.48	0.26	0.17	0.13
LRFN3	leucine rich repeat and fibronectin type III domain containing 3	cell adhesion	-0.23	-0.14	-0.07	0.14
ARHGAP1	Rho GTPase activating protein 1	cell cycle, ras signalling	-0.40	-0.24	-0.10	0.14
APOL3	apolipoprotein L3	intracellular lipid	-0.17	-0.19	-0.08	0.15
SMARCAL1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a-like 1	regulation transcription, helicase	-0.20	-0.11	-0.06	0.22
FRYL	FRY-like		0.21	0.22	0.05	0.25
C18orf8	chromosome 18 open reading frame 8		-0.12	-0.11	-0.04	0.25
RPL21	ribosomal protein L21	translation, ribosome	0.65	0.33	0.11	0.29
PHF5A	PHD finger protein 5A	RNA metabolism, splicing	0.19	0.18	0.05	0.29
HPS6	Hermansky-Pudlak syndrome 6	secretory pathway, platelet	-0.23	-0.18	-0.05	0.29
KIAA1310			-0.18	-0.16	-0.05	0.31
IK	IK cytokine, down-regulator of HLA II		-0.16	-0.18	0.05	0.32
SDHA	succinate dehydrogenase complex, subunit A, flavoprotein (Fp)	energy, mitochondrion	-0.21	-0.20	-0.04	0.35
PPP2R5A	protein phosphatase 2, regulatory subunit B', alpha	cell growth	0.31	0.26	0.06	0.36

CECR1	cat eye syndrome chromosome region, candidate 1	cell proliferation, cell differentiation	-0.28	-0.47	-0.09	0.37
WBP11	WW domain binding protein 11	RNA processing, splicing	-0.11	-0.19	0.05	0.37
ABHD10	abhydrolase domain containing 10		0.28	0.21	0.05	0.38
COPS7B	COP9 constitutive photomorphogenic homolog subunit 7B (Arabidopsis)	protein modification, ubiquitination, pluripotent	-0.14	-0.09	-0.04	0.45
GMFG	glia maturation factor, gamma		0.34	0.33	0.07	0.46
KIAA0930			-0.14	-0.09	-0.03	0.48
S100A8	S100 calcium binding protein A8	pleiotropic, immune,inflammation, cell cycle, cell differentiation, apoptosis	0.76	0.44	0.12	0.49
SP2	Sp2 transcription factor	regulation transcription	-0.32	-0.25	-0.05	0.51
LOC100290936	phosphoglycerate mutase 1-like		-0.27	-0.27	-0.05	0.52
RASSF5	Ras association (RalGDS/AF-6) domain family member 5		-0.17	-0.19	-0.05	0.52
UBAC2	UBA domain containing 2		0.10	0.09	0.03	0.53
ARHGAP17	Rho GTPase activating protein 17	cytsokeleton, Rho signalling	-0.17	-0.13	-0.03	0.53
ZNF319	zinc finger protein 319	regulation transcription, zinc-finger	-0.23	-0.16	-0.03	0.58
TMEM39B	transmembrane protein 39B		-0.18	-0.12	-0.03	0.59
C14orf2	chromosome 14 open reading frame 2		0.34	0.26	0.04	0.61
CD74	CD74 molecule, major histocompatibility complex, class II invariant chain	immune, MHC II, ER transport	-0.37	-0.33	-0.06	0.62
EP300	E1A binding protein p300	transcription regulation, chromatin, cell proliferation, cell differentiation	-0.22	-0.21	0.03	0.68
AIMP2	aminoacyl tRNA synthetase complex-interacting multifunctional protein 2	regulation transcription, RNA processing, tRNA synthetase	-0.13	-0.12	-0.02	0.73
FIP1L1	FIP1 like 1 (S. cerevisiae)	RNA metabolism, polyA	-0.11	-0.11	-0.01	0.77
SEC31A	SEC31 homolog A (S. cerevisiae)	ER vesicles transport	-0.18	-0.11	-0.01	0.86
YWHAB	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, beta polypeptide	signaling, cell cycle, ras signalling	-0.26	-0.21	-0.01	0.90
C16orf72	chromosome 16 open reading frame 72		0.25	0.22	-0.01	0.94
JAK1	Janus kinase 1	immune, ifn	-0.16	-0.15	0.01	0.96
HNRNPAB	heterogeneous nuclear ribonucleoprotein A/B	RNA processing, RNA metabolism, RBP, RRM	-0.16	-0.17	0.00	0.98