

## Supplemental 1: Detailed description of training validation protocol

The complete protocol is depicted in detail in Figure 2, for the case where 10FCV is employed to perform the training, 3FCV is employed to perform the validation and when filtering is employed to select the reporter set. Similar schemes can be constructed for other selection strategies, but the basic principles are exactly the same. Consider the training process, Block 4. First the training data set,  $X_{(j)}$ , is split in ten folds. The training and test sets associated with fold  $i$  are denoted by  $X_{(-i)}$  (all folds but fold  $i$ ) and  $X_{(i)}$  (fold  $i$ ) respectively. During each fold the following steps are performed.

Step 1: Compute a ranking,  $R_{(-i)}$ , of the genes with signal-to-noise ratio (SNR) based on  $X_{(-i)}$ . Employ this ranking to rank both the training and test sets, resulting in the ranked datasets,  $X_{(-i)}^R$  and  $X_{(i)}^R$ .

Step 2: Train the predictor of choice on the top  $n$  ranked genes, producing the trained predictor  $C_n$ .

Step 3: Evaluate the performance of this predictor on the same  $n$  genes in test set  $X_{(i)}^R$ , producing the performance estimate  $t_{i,n}$ .

Repeat Steps 2 and 3 for  $n = 1, 2, \dots, N_g$ , where  $N_g$  is the upper limit of the number of reporters to be evaluated. This produces a curve of performance vs. the number of top ranked genes employed in the predictor.

Steps 1 to 3 are repeated for the other folds of the cross-validation process, i.e. for  $j = 1, 2, \dots, 10$ . The result is a collection of performances,  $\{t_{i,n}\}$ , one for each fold as a function of the number of genes. These are in effect ten performance vs. top-ranked genes curves.

Now we enter Block 2, where the ten results of the cross-validation are combined to estimate the optimal number of reporters. The ten curves are averaged across the different folds, to obtain a single performance curve,  $t_n$ .  $t_n$  is a vector of length  $N_g$ , where the  $i^{\text{th}}$  element is the average

performance of the top  $i$  genes across the ten folds. The maximal value in  $t_n$  is the training performance,  $t_j^*$ , and the number of genes that corresponds to this performance is the optimal number of reporters,  $n^*$ .

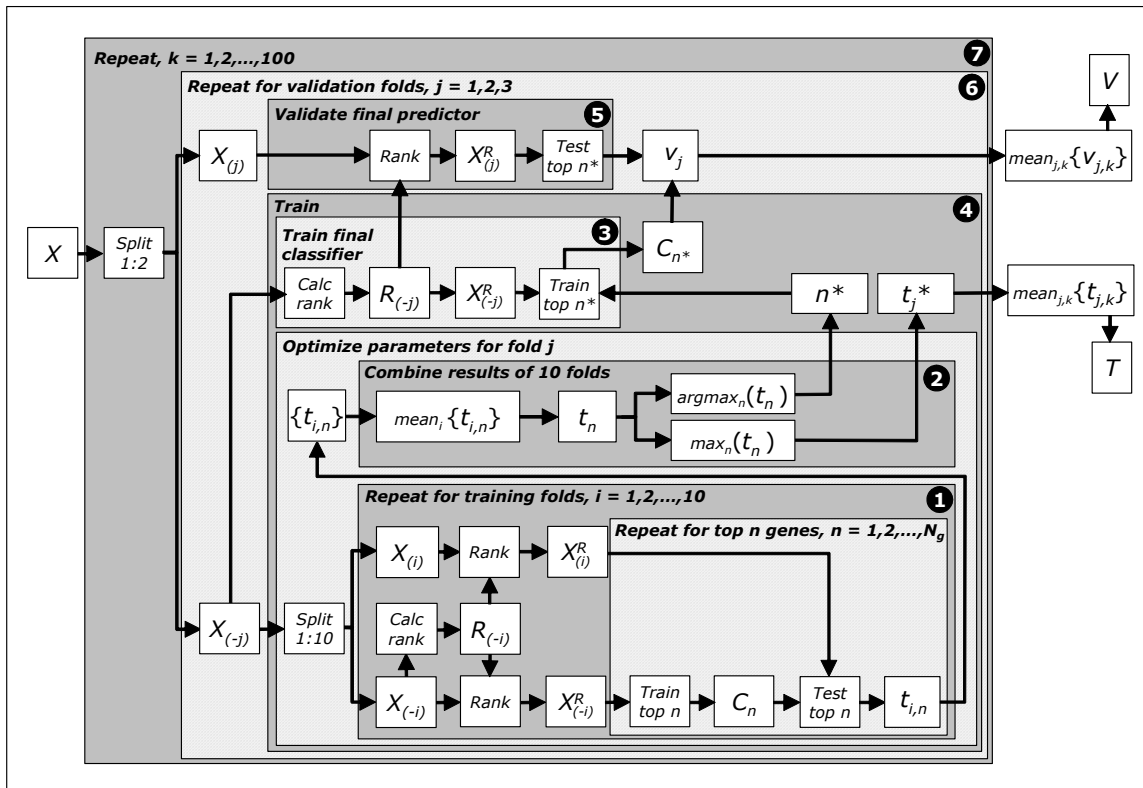
The optimal number of reporters is the input of Block 3, where the final predictor is trained. Here, a ranking,  $R_{(-j)}$ , is computed based on the training set,  $X_{(-j)}$ . This ranking is employed to rank the training set, resulting in the ranked training set,  $X_{(-j)}^R$ . The final predictor,  $C_{n^*}$ , is trained on the top  $n^*$  genes in  $X_{(-j)}^R$ .

Finally, in Block 5, the final predictor,  $C_{n^*}$ , using the same  $n^*$  genes, is applied to the validation set,  $X_{(j)}$  producing the validation performance for the  $j^{\text{th}}$  fold of the validation loop. Note that an analogous procedure is followed for partial least squares. Now the ranking is replaced with a projection from the full set of genes to a subset of 'meta-genes' or 'factors' and the optimization of the number of reporters is substituted with the optimization of the number of factors to employ.

## Supplemental 1: Figure captions

**Figure 2.** A detailed flow diagram of the training validation protocol. Each of the numbered blocks corresponds to the numbered blocks in Figure 1. The reader is referred to Figure 1 and the text for a detailed description.

Figure 2



## Supplemental 2: Datasets employed in this study

**Breast Cancer metastasis dataset.** In (1) a dataset consisting of 78 lymph node negative breast carcinomas was analyzed. This set consisted of a 'GOOD' outcome group of 34 patients with a distant metastasis within five years (mean follow-up of 2.5 years) and a 'POOR' outcome group of 44 with no distant metastasis during a mean follow-up of 8.7 years. This dataset is augmented with 12 POOR outcome tumors (mean follow-up of 2.4 years) and 55 GOOD outcome tumors (mean follow-up 9.7 years) from a larger study (2). This results in a dataset containing 145 tumors in total, 46 in the POOR group and 99 in the GOOD group. Originally all tumors were profiled on cDNA arrays containing 24885 genes. This initial set was reduced to a set of 4912 genes employing the Rosetta error model (8) with  $p_{\min}=0.01$  and  $N_{\min}=3$ .

**Leukemia dataset** (4). This data set contains samples from two variants of Leukemia: 25 samples of *acute myeloid leukemia* (AML) and 47 samples of *acute lymphoblastic leukemia* (ALL), i.e. 72 samples in total. High density oligonucleotide microarrays were employed to measure the expression of 7129 genes for each of the 72 samples. Based on the protocol described in Dudoit *et al.* (9) the number of genes was reduced to 3571.

**Colon cancer dataset** (3). The colon dataset consists of 'tumor' and 'normal' samples collected from colon-cancer patients. The biopsies, consisting of colon epithelial cells, were respectively collected from colon tumors and healthy parts of the colon of the same patient. The dataset consists of 62 samples divided in two diagnostic classes: 40 normal and 22 tumor samples. High-density oligonucleotide microarrays were employed to measure the expression of approximately 6000 genes for each of the 62 samples. From this set, 2000 genes were selected based on confidence measures applied to the expression values (See Alon *et al.* (3)).

***Diffuse large B-cell lymphoma (DLBCL).*** This dataset, which is a subset of the complete dataset first published by Alizadeh et al. (5), contains microarray measurements on two distinct types of diffuse large B-cell lymphoma. There are 47 samples in total, 24 of them are from the "germinal centre B-like" group while 23 are labeled as "activated B-like" group. Each sample is described by 4026 genes.

***Prostate cancer dataset.*** Singh et al. (6) published this dataset which consists of two classes. The 'tumor' class contains 52 prostate tumor samples the 'normal' class contains 50 non-tumor prostate samples. High density oligonucleotide microarrays, containing around 12600 genes, were employed to measure these data. The same preprocessing steps as described in the supplementary information supplied in (6) were employed to reduce the set of genes to 5962 genes.

***Central nervous system (CNS).*** The central nervous system dataset is a subset of a larger study by Pomeroy et al. (7). The subset employed in this study considers the outcome (survival) after embryonic treatment of the central nervous system. Survivors are patients who are alive after treatment while the failures are those who succumbed to their disease. The data set contains 60 samples, 21 are survivors and 39 represent deaths. High density oligonucleotide microarrays were employed to measure the expression of 7129 genes for each of the 60 samples. The data was preprocessed according to the steps outlined in the supplementary information associated with (7) resulting in a dataset consisting of 4458 genes.

## Supplemental 2: References

1. Van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen *et al.* (2002) *Nature* **415**, 530–536.
2. Marc J. van de Vijver, Yudong D. He, Laura J. van 't Veer, Hongyue Dai, Augustinus A.M. Hart, Dorien W. Voskuil, George J. Schreiber, Johannes L. Peterse, Chris Roberts *et al.* (2002) *N. Engl. J. Med.*, **347** 1999-2009
3. Alon, U., Barkai, N., Notterman, D., Gish, K. Mack, S. and Levine, J. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 6745–6750.
4. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J. *et al.* (1999) *Science* **286**, 531–537
5. Alizadeh, A. A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T, Yu, X. *et al.* (2000) *Nature* **403**, 503-511.
6. Singh, D. Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., (2002), *Cancer Cell* **1**, 203-209.
7. Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y.H., Goumnerova, L.C., Black, P.M., Lau, C., et al (2002) *Nature* **415**, 436 – 442.
8. Roberts, C.J., Nelson, B., Marton, M.J., Stoughton, R., Meyer, M.R., Bennett, H.A. (2000) *Science* **287**, 873-880.
9. Dudoit, S., Fridlyand, J. and Speed, T. (2002). *JASA* **97**, 77–87.

## **Supplemental 3:** **Results on Colon, Leukemia, DLBCL, CNS and Prostate datasets**

### **Supplemental 3: Table caption**

**Table 4.** Validation,  $V$ , and training,  $T$ , performance of the different predictors and reporter selection strategies on the Colon dataset. The best predictor-selector combination is indicated in bold. The column marked 'k\*' contains the average number of factors selected during training for the cases where PLS is employed for dimensionality reduction, and the optimal number of genes selected in the remaining cases. The columns 'W','D' and 'L' represent the number of times a particular selector-predictor combination won (better validation performance), drew (same performance) and lost (worse performance) from the best combination across all 300 validation folds.

**Table 5.** Validation,  $V$ , and training,  $T$ , performance of the different predictors and reporter selection strategies on the Leukemia dataset. The best predictor-selector combination is indicated in bold. The column marked 'k\*' contains the average number of factors selected during training for the cases where PLS is employed for dimensionality reduction, and the optimal number of genes selected in the remaining cases. The columns 'W','D' and 'L' represent the number of times a particular selector-predictor combination won (better validation performance), drew (same performance) and lost (worse performance) from the best combination across all 300 validation folds.

**Table 6.** Validation,  $V$ , and training,  $T$ , performance of the different predictors and reporter selection strategies on the DLBCL dataset. The best predictor-selector combination is indicated in bold. The column marked 'k\*' contains the average number of factors selected during training for the cases where PLS is employed for dimensionality reduction, and the optimal number of genes

selected in the remaining cases. The columns 'W','D' and 'L' represent the number of times a particular selector-predictor combination won (better validation performance), drew (same performance) and lost (worse performance) from the best combination across all 300 validation folds.

**Table 7.** Validation,  $V$ , and training,  $T$ , performance of the different predictors and reporter selection strategies on the Prostate dataset. The best predictor-selector combination is indicated in bold. The column marked 'k\*' contains the average number of factors selected during training for the cases where PLS is employed for dimensionality reduction, and the optimal number of genes selected in the remaining cases. The columns 'W','D' and 'L' represent the number of times a particular selector-predictor combination won (better validation performance), drew (same performance) and lost (worse performance) from the best combination across all 300 validation folds.

**Table 8.** Validation,  $V$ , and training,  $T$ , performance of the different predictors and reporter selection strategies on the CNS dataset. The best predictor-selector combination is indicated in bold. The column marked 'k\*' contains the average number of factors selected during training for the cases where PLS is employed for dimensionality reduction, and the optimal number of genes selected in the remaining cases. The columns 'W','D' and 'L' represent the number of times a particular selector-predictor combination won (better validation performance), drew (same performance) and lost (worse performance) from the best combination across all 300 validation folds.

**Table 4**

Colon									
Reporter Selection	Predictor	T(%)		V(%)		k*	S		
		Mean	SD	Mean	SD		W	D	L
Filter	<b>NMC</b>	<b>86.6</b>	<b>3.7</b>	<b>86.7</b>	<b>1.9</b>	<b>24</b>	-	-	-
	DLDC	86.3	3.8	85.9	2.6	29	60	162	78
	SBGC	86.2	3.9	86.1	2	50	52	168	80
	1NN	79.7	5.6	78.7	4.3	44	38	42	220
	5NN	85.5	4.6	85.2	2.7	36	58	148	94
	9NN	86.3	4	86.1	2.2	30	59	180	61
	RFLD(0)	84.8	5.2	84.1	3.4	106	71	91	138
	RFLD(1)	86.4	4.1	86	2.8	38	74	144	82
	RFLD(10)	85.9	3.9	85.8	2.5	30	53	166	81
	linSVC	80.8	5.8	79.9	4.6	50	75	16	209
PLS	NMC	82.3	4.8	82	3.3	13.6	85	22	193
	DLDC	81	4.8	81.2	3.5	13.5	78	19	203
	SBGC	80.1	5	78.9	4.1	10.6	63	22	215
	1NN	81.4	5.5	75.2	5	8.7	41	5	254
	5NN	82.2	5.3	76.5	5.4	8.5	66	6	228
	9NN	81.3	5.6	74.9	5.3	7.5	49	15	236
	RFLD(0)	85.8	4.5	83.9	3.6	9	107	20	173
	RFLD(1)	85.8	4.5	84.2	3.5	8.9	108	21	171
	RFLD(10)	84.5	4.5	84.2	2	13.2	102	31	167
	linSVC	84.3	5.1	83.5	3.5	12.1	108	18	174
SC	SC	86.6	3.8	85	2.5	242	123	26	151
RFE	LinSVC	83	6.3	80.5	4.5	751	72	23	205

**Table 5**

Leukemia									
Reporter Selection	Predictor	T(%)		V(%)		k*	S		
		Mean	SD	Mean	SD		W	D	L
Filter	NMC	96.8	1.7	95.4	1.2	82	68	80	152
	DLDC	96.5	1.9	95.7	1.6	108	77	72	151
	SBGC	92.4	2.5	91.5	3.5	66	49	53	198
	1NN	95.6	2.4	94.5	2	73	56	75	169
	5NN	97.2	1.8	95.9	1.6	123	75	92	133
	9NN	97.1	1.7	96	1.4	89	74	93	133
	RFLD(0)	96.1	2.2	94.3	2.1	154	56	70	174
	RFLD(1)	96.5	2	95.7	1.6	87	77	85	138
	RFLD(10)	96.6	1.8	95.4	1.5	89	74	82	144
	linSVC	95.7	2.4	94.1	2	101	50	69	181
PLS	NMC	97.2	1.8	97.2	0.7	5.1	14	262	24
	DLDC	97.2	1.8	97.2	0.7	4.9	10	274	16
	SBGC	97.1	2	97	0.8	7.1	12	252	36
	1NN	97.7	1.7	96.3	1.7	9.5	41	186	73
	5NN	97.8	1.6	97.1	1.4	6.7	49	214	37
	9NN	97.8	1.6	97	1.5	5.6	52	206	42
	RFLD(0)	97.6	1.7	96.6	1.1	5.2	32	184	84
	RFLD(1)	97.6	1.7	96.8	1	5.3	29	203	68
	<b>RFLD(10)</b>	<b>97.1</b>	<b>1.9</b>	<b>97.3</b>	<b>0.8</b>	<b>5</b>	-	-	-
	linSVC	97.2	2	96.7	1	8.3	24	208	68
SC	SC	97.4	1.7	96.6	1.2	2210	84	96	120
RFE	LinSVC	96.9	2.2	96.3	1.6	2832	69	117	114

**Table 6**

DLBCL									
Reporter Selection	Predictor	T(%)		V(%)		k*	S		
		Mean	SD	Mean	SD		W	D	L
Filter	NMC	95.6	3.2	95	2.9	92	74	120	106
	DLDC	95.5	2.7	94.9	2.8	86	44	174	82
	SBGC	95.5	2.7	94.7	2.3	91	42	165	93
	1NN	95.5	3.1	94.9	2.7	93	75	120	105
	5NN	95.8	2.8	95	2.8	99	71	121	108
	9NN	95.6	2.9	95.2	2.7	93	76	124	100
	RFLD(0)	93.9	3.6	91.8	3.6	149	40	101	159
	RFLD(1)	95.2	2.7	94.4	2.9	54	52	148	100
	<b>RFLD(10)</b>	<b>96.4</b>	<b>2.7</b>	<b>95.7</b>	<b>2.8</b>	<b>80</b>	-	-	-
	linSVC	96.1	3.2	94.2	2.6	85	70	100	130
PLS	NMC	90	3.7	90.4	3.9	16.6	51	58	191
	DLDC	90	3.9	90.7	3.4	17.1	50	57	193
	SBGC	87.9	4.7	87.3	4	16.6	33	38	229
	1NN	86.4	5.1	82.6	3.9	16	16	18	266
	5NN	88	4.2	87.9	4.1	16.6	47	37	216
	9NN	87.8	4	87.4	4.6	16.5	37	34	229
	RFLD(0)	64.3	5.2	58.9	5	6.1	2	2	296
	RFLD(1)	66.1	5	64.9	5.6	10.8	5	2	293
	RFLD(10)	90.5	3.6	91.5	3.8	16.7	56	67	177
	linSVC	85.9	5.3	89.6	4.4	18.4	43	50	207
SC	SC	97.3	1.9	94.7	2.4	380	80	106	114
RFE	LinSVC	94	3.9	90.4	3.1	767	32	72	196

**Table 7**

Prostate									
Reporter Selection	Predictor	T(%)		V(%)		k*	S		
		Mean	SD	Mean	SD		W	D	L
Filter	NMC	89.6	2.1	88.8	2	15	68	25	207
	DLDC	91.8	2.1	91.8	1.8	13	123	26	151
	SBGC	91.5	2.1	91	1.9	25	102	28	170
	1NN	88.7	3.2	87.6	2.4	49	52	21	227
	5NN	91.3	2.1	90.9	1.8	24	88	31	181
	9NN	91.2	2.1	91	1.8	21	85	34	181
	RFLD(0)	91.3	2.1	90.6	2.4	60	101	25	174
	RFLD(1)	91.3	2.1	90.6	2.7	79	98	28	174
	RFLD(10)	92	2	90.9	2.5	71	102	23	175
	linSVC	91.9	3	90.4	2.6	156	78	31	191
PLS	NMC	68.2	3.6	67.9	1.5	13.7	0	0	300
	DLDC	73.7	3.2	74.2	2.7	16.6	0	1	299
	SBGC	69	3.6	65.8	2.5	14.4	0	0	300
	1NN	83.1	3.5	79.5	3	11.8	0	2	298
	5NN	82.7	3	79.6	2.5	11.3	2	3	295
	9NN	83	2.7	81.3	2.2	11.9	4	3	293
	<b>RFLD(0)</b>	<b>94.1</b>	<b>2.1</b>	<b>93.4</b>	<b>1.7</b>	<b>14</b>	-	-	-
	<b>RFLD(1)</b>	<b>94.1</b>	<b>2.1</b>	<b>93.4</b>	<b>1.7</b>	<b>14</b>	-	-	-
	RFLD(10)	94.1	2.1	93.4	1.7	14.1	0	299	1
	linSVC	93.3	2.2	93.2	1.8	17	88	122	90
SC	SC	89.3	1.9	89.1	1.6	41	70	31	199
RFE	LinSVC	93.5	2.5	92.2	1.6	1610	118	38	144

**Table 8**

CNS									
Reporter Selection	Predictor	T(%)		V(%)		k*	S		
		Mean	SD	Mean	SD		W	D	L
Filter	<b>NMC</b>	<b>60</b>	<b>7.4</b>	<b>61.3</b>	<b>5.2</b>	<b>96</b>	-	-	-
	DLDC	59.1	7.6	60.2	4.5	86	112	26	162
	SBGC	56.2	6.5	56.3	5.3	88	98	13	189
	1NN	57	7.8	56.7	6	74	107	16	177
	5NN	57.6	7.1	58.7	5	85	114	23	163
	9NN	57.1	7.2	58	5.3	80	119	16	165
	RFLD(0)	61.2	8.5	58.9	5.8	142	105	24	171
	RFLD(1)	58.6	8.6	56.3	5	118	84	13	203
	RFLD(10)	58.6	8.6	56.6	5	118	87	15	198
	linSVC	60.3	7.9	57.6	5.5	120	112	15	173
PLS	NMC	63.9	4.8	59.6	5.9	8.9	133	8	159
	DLDC	64.8	5	60.1	5.6	9	134	8	158
	SBGC	62.9	5.7	57.6	5	7.1	116	4	180
	1NN	62.1	5.6	54.2	6.1	9.4	90	8	202
	5NN	62.7	5.4	54.9	5.4	7.7	100	4	196
	9NN	60.9	5.2	54.5	4.3	7.2	91	3	206
	RFLD(0)	61.2	4.8	47.1	5.1	4.8	52	4	244
	RFLD(1)	61.2	4.8	47.1	5.1	4.8	52	4	244
	RFLD(10)	61.2	4.8	47.1	5.1	4.8	52	4	244
linSVC	58.3	5.6	53.9	4.3	8.9	87	8	205	
SC	SC	65.4	5.4	61	4.6	966	138	6	156
RFE	LinSVC	64.9	6.7	60.1	5.8	1235	132	10	158