

Outcome prediction in breast cancer

M.H. Van Vliet F. Reyal, H. Horlings, M.J.T.Reinders and
L.F.A.Wessels

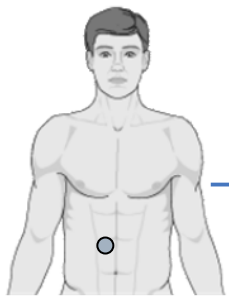
Bioinformatics and Statistics group
The Netherlands Cancer Institute
Amsterdam

Bioinformatics group
Faculty of EEMCS, TU Delft
Delft

Outline

- Gene expression signatures (classifiers)
 - ML, bias
- How to be cautious?
 - Keep it simple...
- (Breast) cancer outcome signatures
 - Bias, variability
- More data (~1000 samples)
 - Revisit variability, bias, algorithmic choice
 - Evaluate signature discordance
 - Reveal biological processes

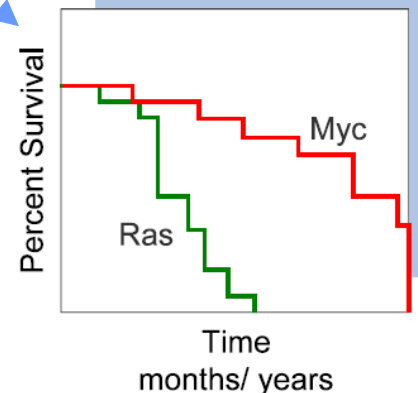
Gene signatures for outcome prediction



Measure:

- Gene expression
- Pathology
- DNA aberration
- Imaging
- ...

Subtype



Gene expression signatures

Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 6745–6750, June 1999
Cell Biology

Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays

U. ALON^{*†}, N. BARKAI^{*†}, D. A. NOTTERMAN^{*}, K. GISH[‡], S. YBARRA[‡], D. MACK[‡], AND A. J. LEVINE^{*§}

Departments of ^{*}Molecular Biology and [†]Physics, Princeton University, Princeton, NJ 08540; and [‡]EOS Biotechnology, 225A Gateway Boulevard, South San Francisco, CA 94080

REPORTS

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

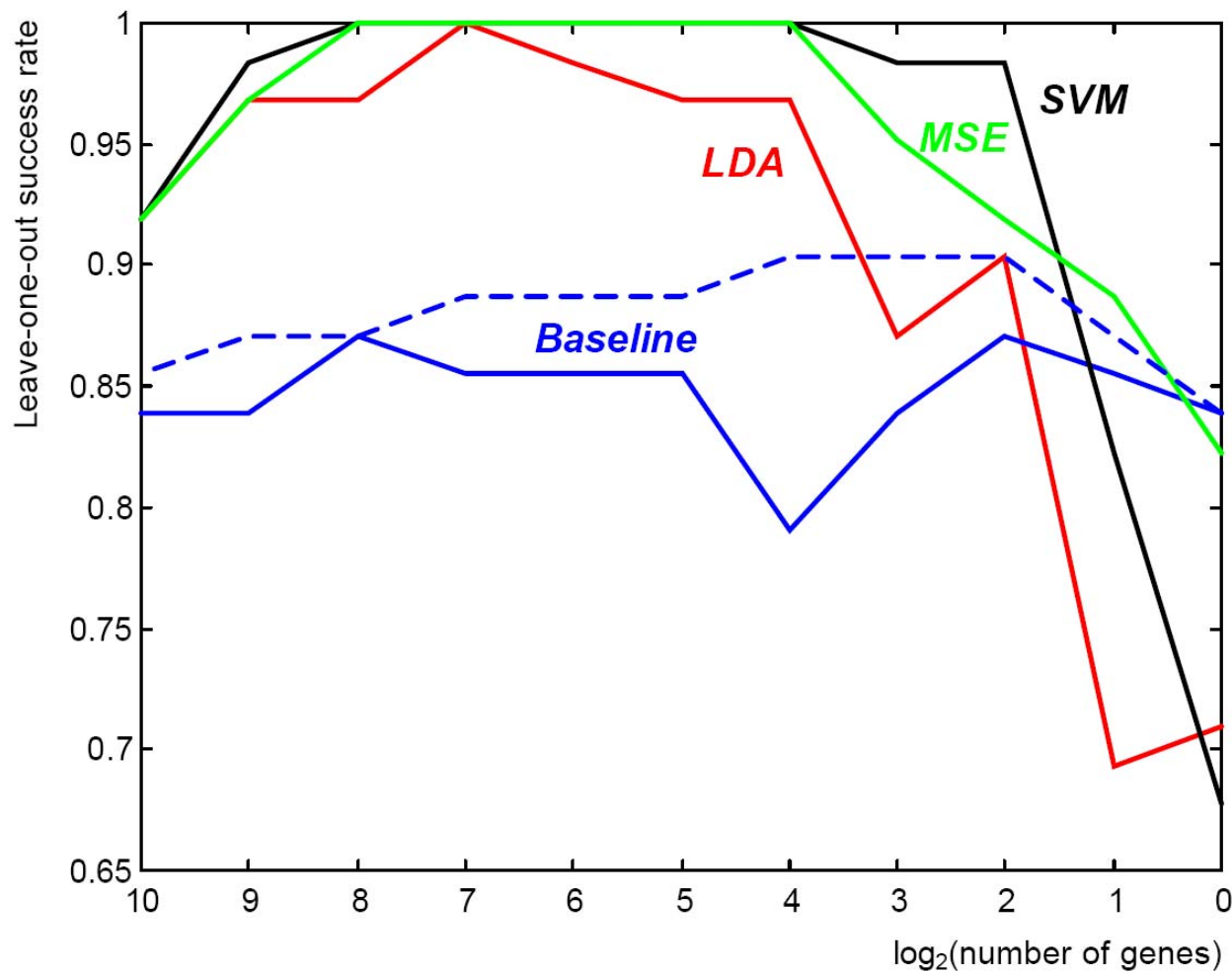
T. R. Golub,^{1,2*†} D. K. Slonim,^{1†} P. Tamayo,¹ C. Huard,¹
M. Gaasenbeek,¹ J. P. Mesirov,¹ H. Coller,¹ M. L. Loh,²
J. R. Downing,³ M. A. Caligiuri,⁴ C. D. Bloomfield,⁴
E. S. Lander^{1,5*}

acute leukemia have been found to be associated with specific chromosomal translocations—for example, the t(12;21)(p13;q22) translocation occurs in 25% of patients with ALL, whereas the t(8;21)(q22;q22) occurs in 15% of patients with AML (7).

Although the distinction between AML and ALL has been well established, no single test is currently sufficient to establish the diagnosis. Rather, current clinical practice involves an experienced hematopathologist's interpretation of the tumor's morphology, histochemistry, immunophenotyping, and cytogenetic analysis, each performed in a separate, highly specialized laboratory. Although usually accurate, leukemia classification remains imperfect and errors do occur.

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery)

Guyon et al. (2002) *Mach. Learn.* **46**, 389–422.



Other cases of selection bias

1. Xiong *et al.* Genome Research, 2001;
 - 10.7% vs. ~15% Colon;
 - 0% vs. ~5% Leukemia
2. Zhang *et al.* PNAS, 2001;
 - 2% vs. ~15% Colon
3. Guyon *et al.* Machine Learning, 2002;
 - 0% vs. ~15% Colon
4. Grate *et al.* WABI, 2002.
 - 1% vs. ~35% 78 BC tumors!

Selection bias in gene extraction on the basis of microarray gene-expression data

Christophe Ambroise[†] and Geoffrey J. McLachlan^{*§}

[†]Laboratoire Heudiasyc, Unité Mixte de Recherche/Centre National de la Recherche Scientifique 6599, 60200 Compiègne, France; and ^{*}Department of Mathematics, University of Queensland, Brisbane 4072, Australia

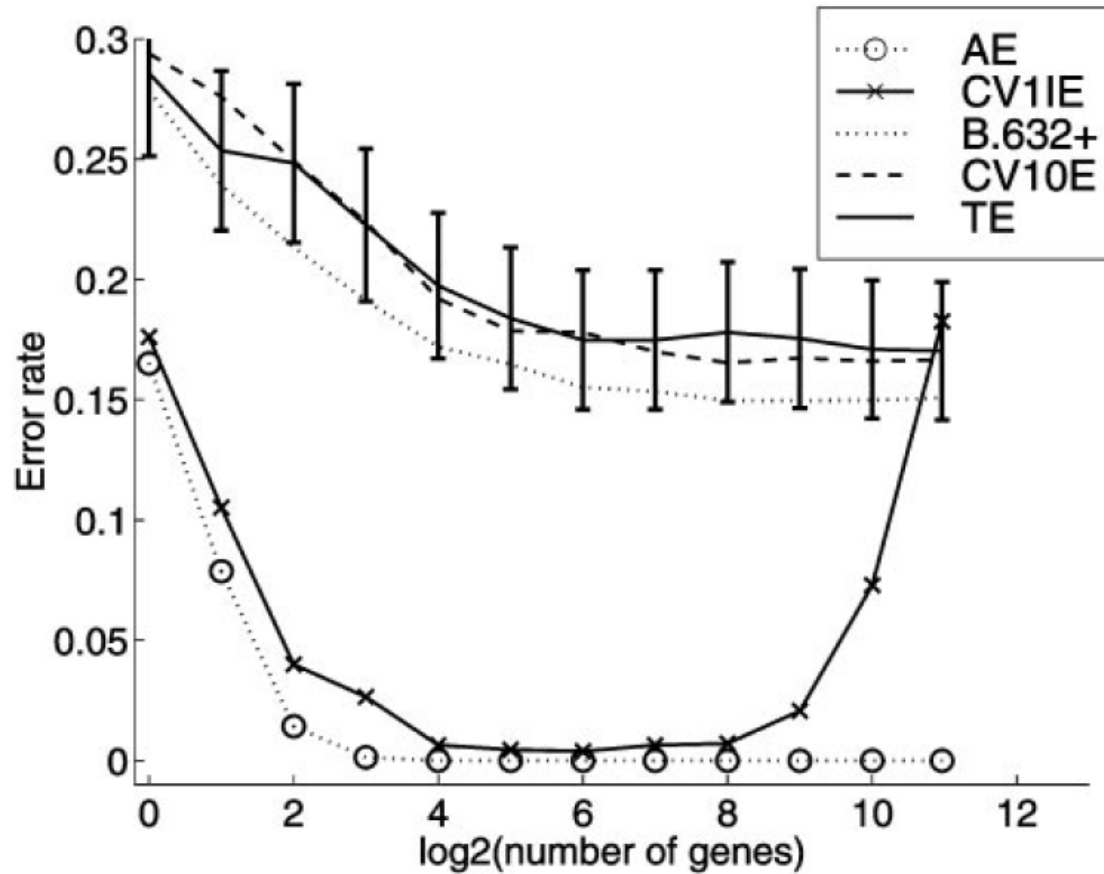
Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved March 21, 2002 (received for review February 20, 2002)

In the context of cancer diagnosis and treatment, we consider the problem of constructing an accurate prediction rule on the basis of a relatively small number of tumor tissue samples of known type containing the expression data on very many (possibly thousands) genes. Recently, results have been presented in the literature suggesting that it is possible to construct a prediction rule from only a few genes such that it has a negligible prediction error rate. However, in these results the test error or the leave-one-out cross-validated error is calculated without allowance for the selection bias. There is no allowance because the rule is either tested on tissue samples that were used in the first instance to select the genes being used in the rule or because the cross-validation of the rule is not external to the selection process; that is, gene selection is not performed in training the rule at each stage of the cross-validation process. We describe how in practice the selection bias can be assessed and corrected for by either performing a cross-validation or applying the bootstrap external to the selection process. We recommend using 10-fold rather than leave-one-out cross-validation, and concerning the bootstrap, we suggest using the so-called .632+ bootstrap error estimate designed to handle overfitted prediction rules. Using two published data sets, we

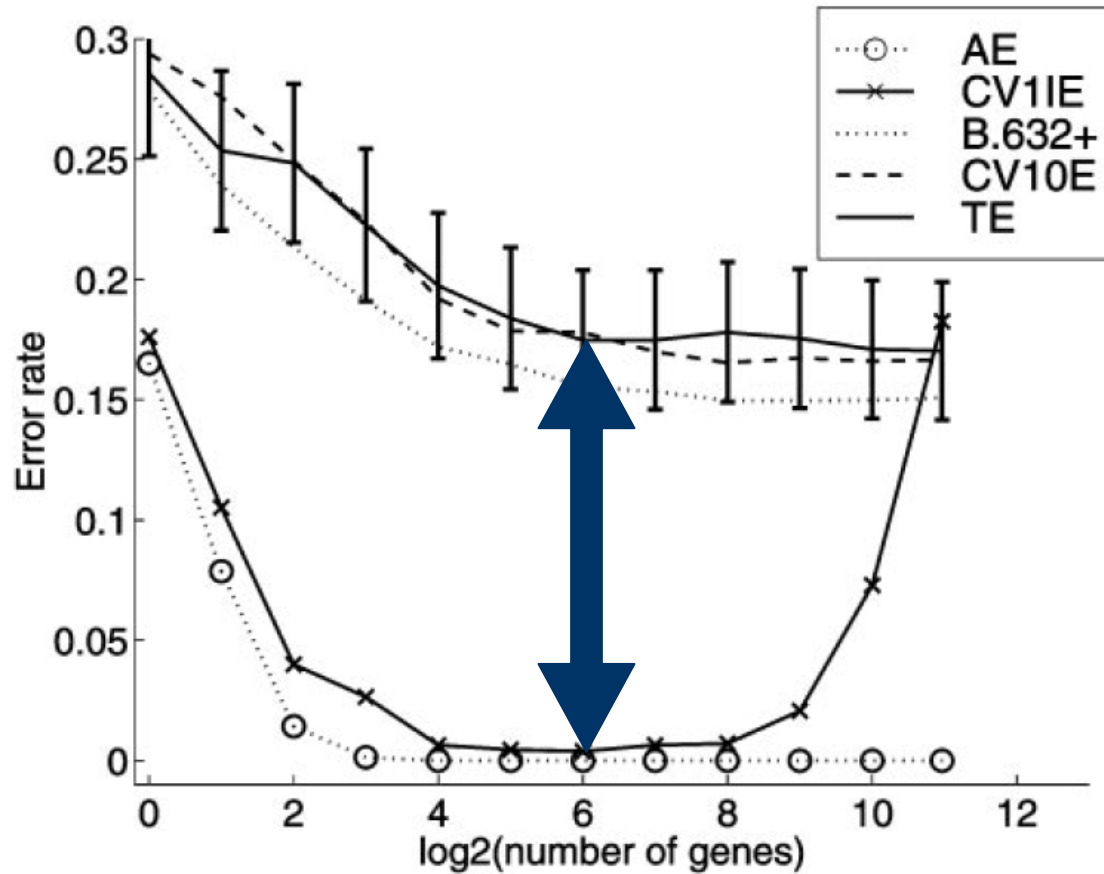
Fisher's linear discriminant function is singular if $n < g + p$. Second, even if all the genes can be used as, say, with a Euclidean-based rule or a support vector machine (SVM; refs. 9, 16, and 17), the use of all the genes allows the noise associated with genes of little or no discriminatory power, which inhibits and degrades the performance of the rule R in its application to unclassified tumors. That is, although the apparent error rate (AE) of the rule R (the proportion of the training tissues misallocated by R) will decrease as it is formed from more and more genes, its error rate in classifying tissues outside of the training set eventually will increase. That is, the generalization error of R will be increased if it is formed from a sufficiently large number of genes. Hence, in practice consideration has to be given to implementing some procedure of feature selection for reducing the number of genes to be used in constructing the rule R .

A number of approaches to feature-subset selection have been proposed in the literature (18). All these approaches involve searching for an optimal or near optimal subset of features that optimize a given criterion. Feature-subset selection can be classified into two categories based on whether the criterion depends on the learning algorithm used to construct the rule

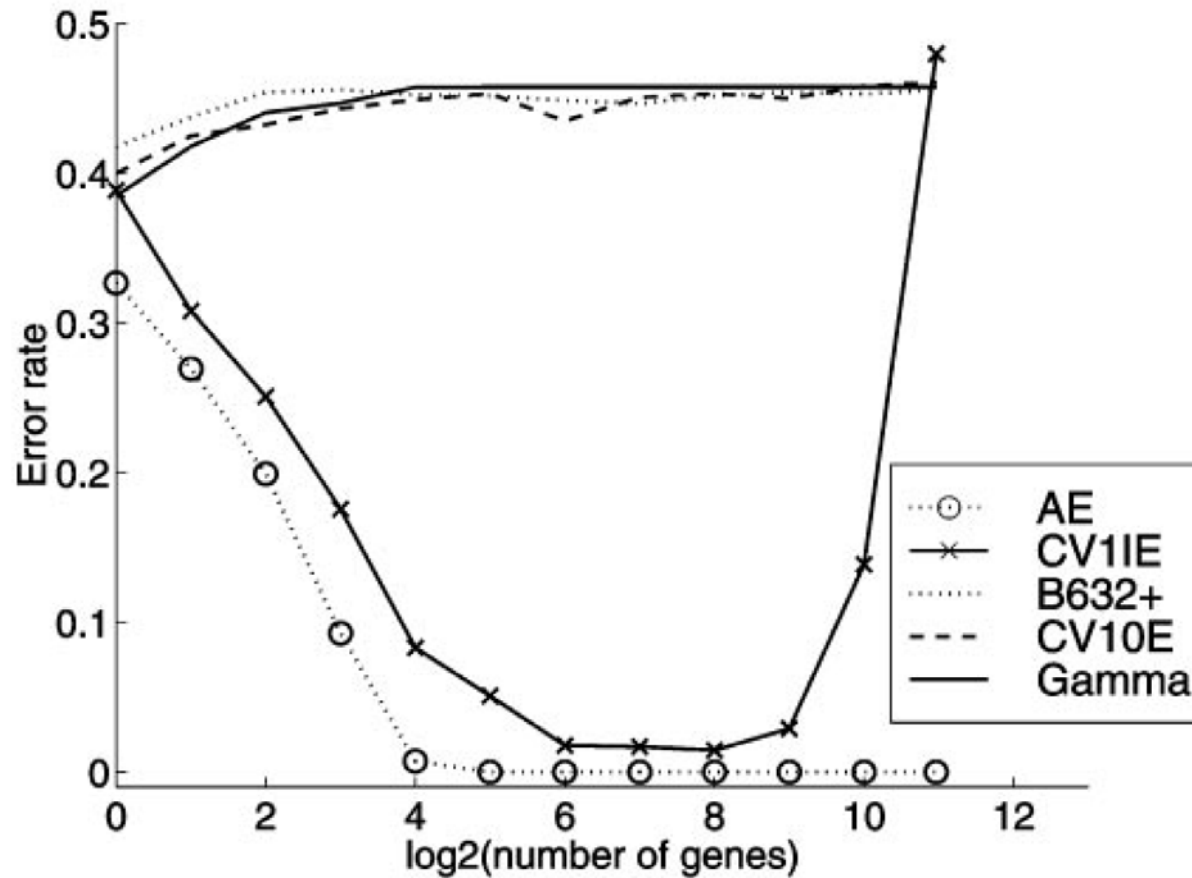
'Out-of-the-loop' gene selection



'Out-of-the-loop' gene selection



On a 'no-information' dataset



NKI 70 gene signature

Gene expression profiling predicts clinical outcome of breast cancer

Laura J. van 't Veer^{*,†}, Hongyue Dai^{‡,§}, Marc J. van de Vijver^{*,†}, Yudong D. He[‡], Augustinus A. M. Hart^{*}, Mao Mao[‡], Hans L. Peterse^{*}, Karin van der Kooy^{*}, Matthew J. Marton[‡], Anke T. Witteveen^{*}, George J. Schreiber[‡], Ron M. Kerkhoven^{*}, Chris Roberts[‡], Peter S. Linsley[‡], René Bernards^{*} & Stephen H. Friend[‡]

^{*} Divisions of Diagnostic Oncology, Radiotherapy and Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands

[‡] Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034, USA

[†] These authors contributed equally to this work

Breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome. The strongest predictors for metastases (for example, lymph node status and histological grade) fail to classify accurately breast tumours according to their clinical behaviour^{1–3}. Chemotherapy or hormonal therapy reduces the risk of distant metastases by approximately one-third; however, 70–80% of patients receiving this treatment would have survived without it^{4,5}. None of the signatures of breast cancer gene expression reported to date^{6–12} allow for patient-tailored therapy strategies. Here we used DNA microarray analysis on primary breast tumours of 117 young patients, and applied supervised classification to identify a gene expression signature strongly predictive of a short interval to distant metastases ('poor prognosis' signature) in patients with-

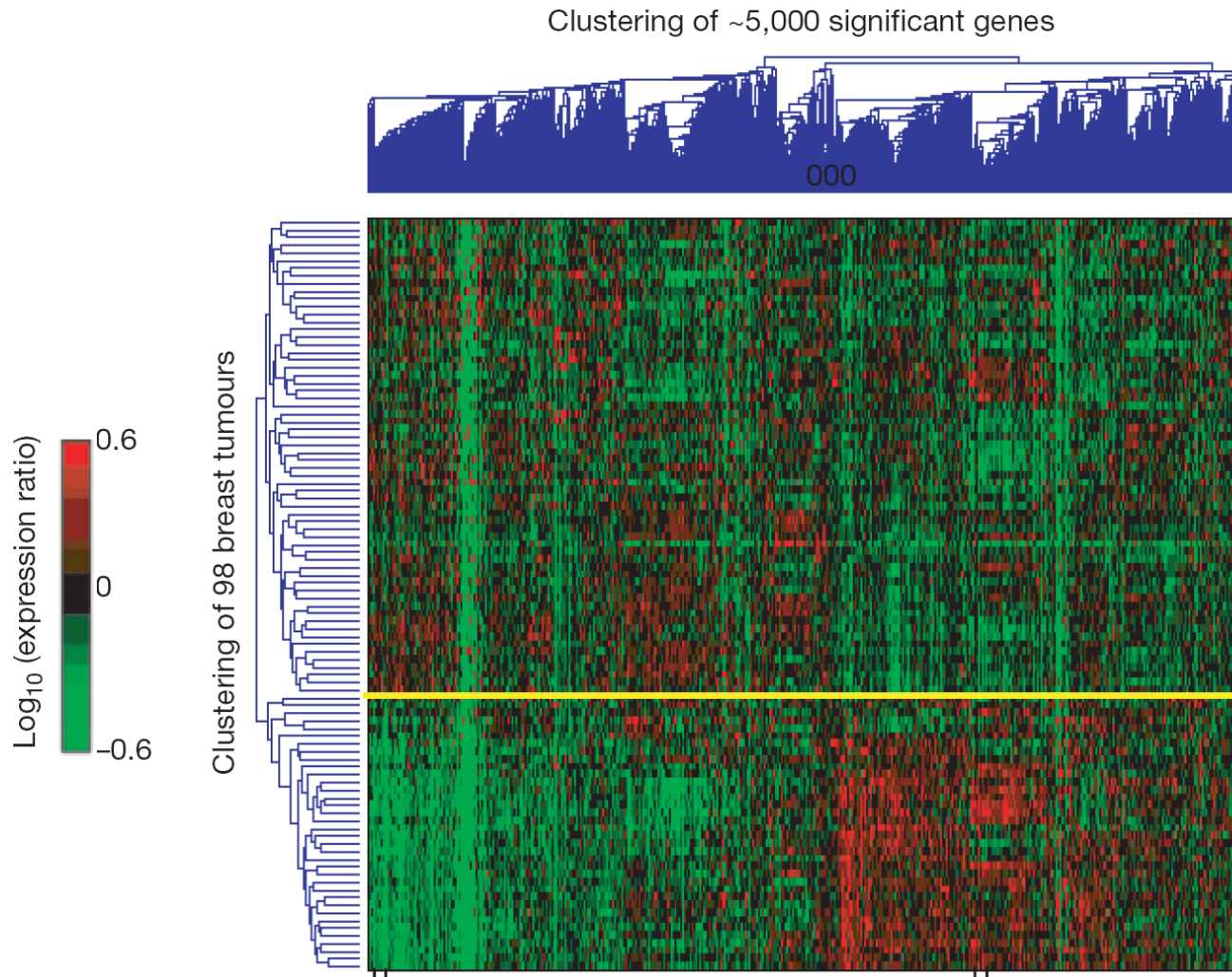
tumours are the dominant feature in this two-dimensional display (top and bottom of plot, representing 62 and 36 tumours, respectively), suggesting that the tumours can be divided into two types on the basis of this set of ~5,000 significant genes. Notably, in the upper group only 34% of the sporadic patients were from the group who developed distant metastases within 5 years, whereas in the lower group 70% of the sporadic patients had progressive disease (Fig. 1b). Thus, using unsupervised clustering we can already, to some extent, distinguish between 'good prognosis' and 'poor prognosis' tumours.

To gain insight into the genes of the dominant expression signatures, we associated them with histopathological data; for example, oestrogen receptor (ER)- α expression as determined by immunohistochemical (IHC) staining (Fig. 1b). Out of 39 IHC-stained tumours negative for ER- α expression (ER negative), 34 clustered together in the bottom branch of the tumour dendrogram. In the enlargement shown in Fig. 1c, a group of downregulated genes is represented containing both the ER- α gene (*ESR1*) and genes that are apparently co-regulated with ER, some of which are known ER target genes. A second dominant gene cluster is associated with lymphocytic infiltrate and includes several genes expressed primarily by B and T cells (Fig. 1d).

Sixteen out of eighteen tumours of *BRCA1* carriers are found in the bottom branch intermingled with sporadic tumours. This is consistent with the idea that most *BRCA1* mutant tumours are ER negative and manifest a higher amount of lymphocytic infiltrate¹⁵. The two tumours of *BRCA2* carriers are part of the upper cluster of tumours and do not show similarity with *BRCA1* tumours. Neither high histological grade nor angiogenesis is a specific feature of either of the clusters (Fig. 1b). We conclude that unsupervised clustering detects two subgroups of breast cancers, which differ in ER status and lymphocytic infiltration. A similar conclusion has

NKI 70 gene signature

a

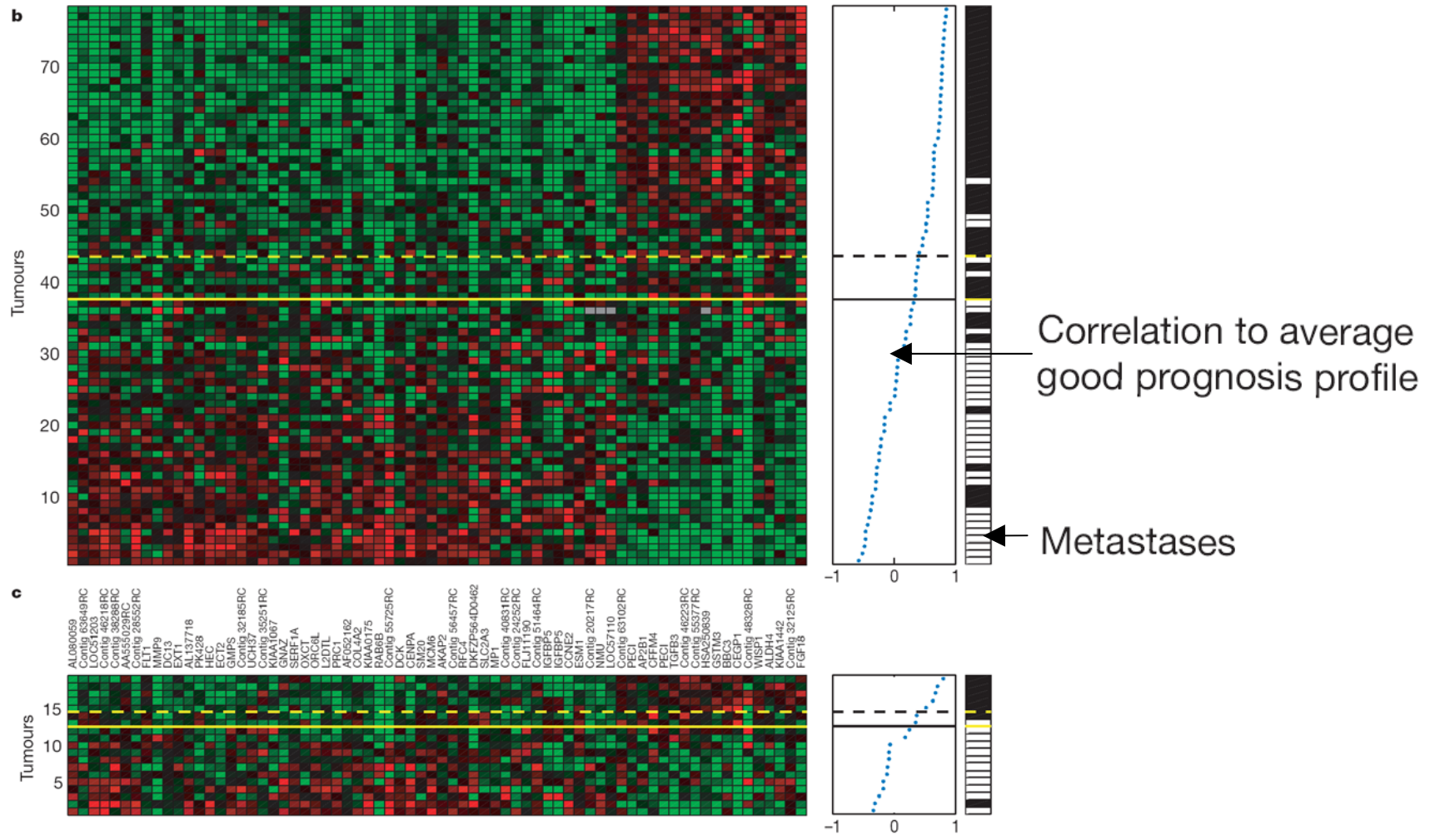


b



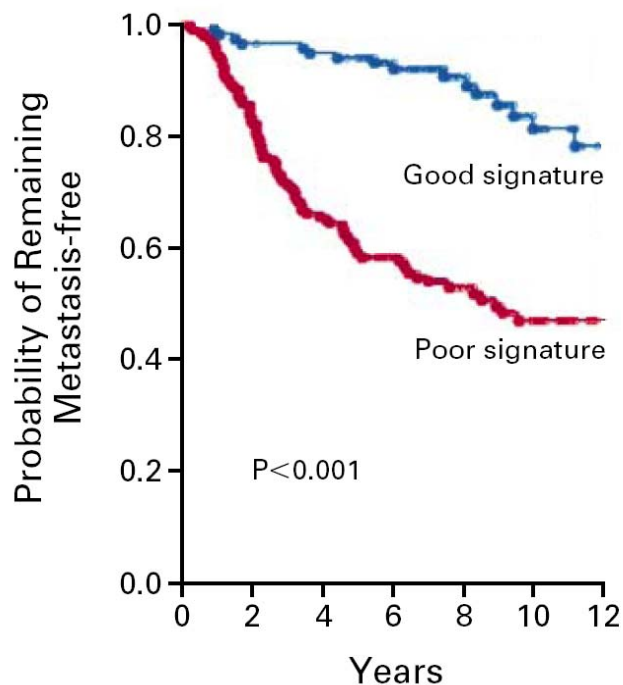
NKI 70 gene signature

70 genes; Nearest centroid classifier (cosine)



Validation on series of 295 patients

All Patients

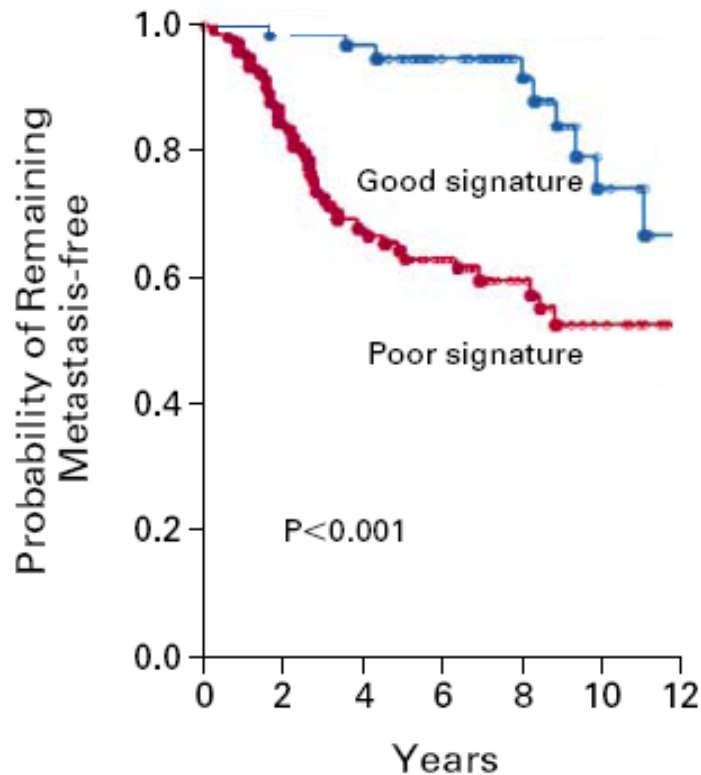


CHARACTERISTIC	POOR-PROGNOSIS	GOOD-PROGNOSIS	P VALUE
	SIGNATURE (N=180)	SIGNATURE (N=115)	
	no. of patients (%)		
Age			<0.001
<40 yr	52 (29)	11 (10)	
40-44 yr	41 (23)	44 (38)	
45-49 yr	55 (31)	43 (37)	
≥50 yr	32 (18)	17 (15)	
No. of positive nodes			0.60
0	91 (51)	60 (52)	
1-3	63 (35)	43 (37)	
≥4	26 (14)	12 (10)	
Tumor diameter			0.012
≤20 mm	84 (47)	71 (62)	
>20 mm	96 (53)	44 (38)	
Histologic grade			<0.001
I (good)	19 (11)	56 (49)	
II (intermediate)	56 (31)	45 (39)	
III (poor)	105 (58)	14 (12)	
Vascular invasion			0.38
Absent	108 (60)	77 (67)	
1-3 Vessels	18 (10)	12 (10)	
>3 Vessels	54 (30)	26 (23)	
Estrogen-receptor status			<0.001
Negative	66 (37)	3 (3)	
Positive	114 (63)	112 (97)	
Surgery			0.63
Breast-conserving therapy	97 (54)	64 (56)	
Mastectomy	83 (46)	51 (44)	
Chemotherapy			0.79
No	114 (63)	71 (62)	
Yes	66 (37)	44 (38)	
Hormonal therapy			0.63
No	157 (87)	98 (85)	
Yes	23 (13)	17 (15)	

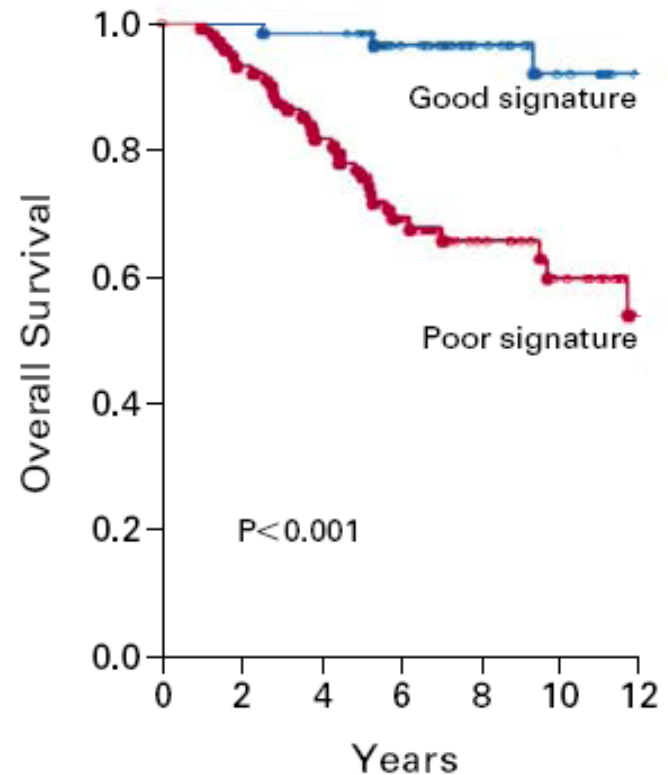
van de Vijver M *et al.* N. Engl. J. Med. 2002;
347: 1999-2009.

Validation on series of 295 patients

E Lymph-Node-Positive Patients



F Lymph-Node-Positive Patients



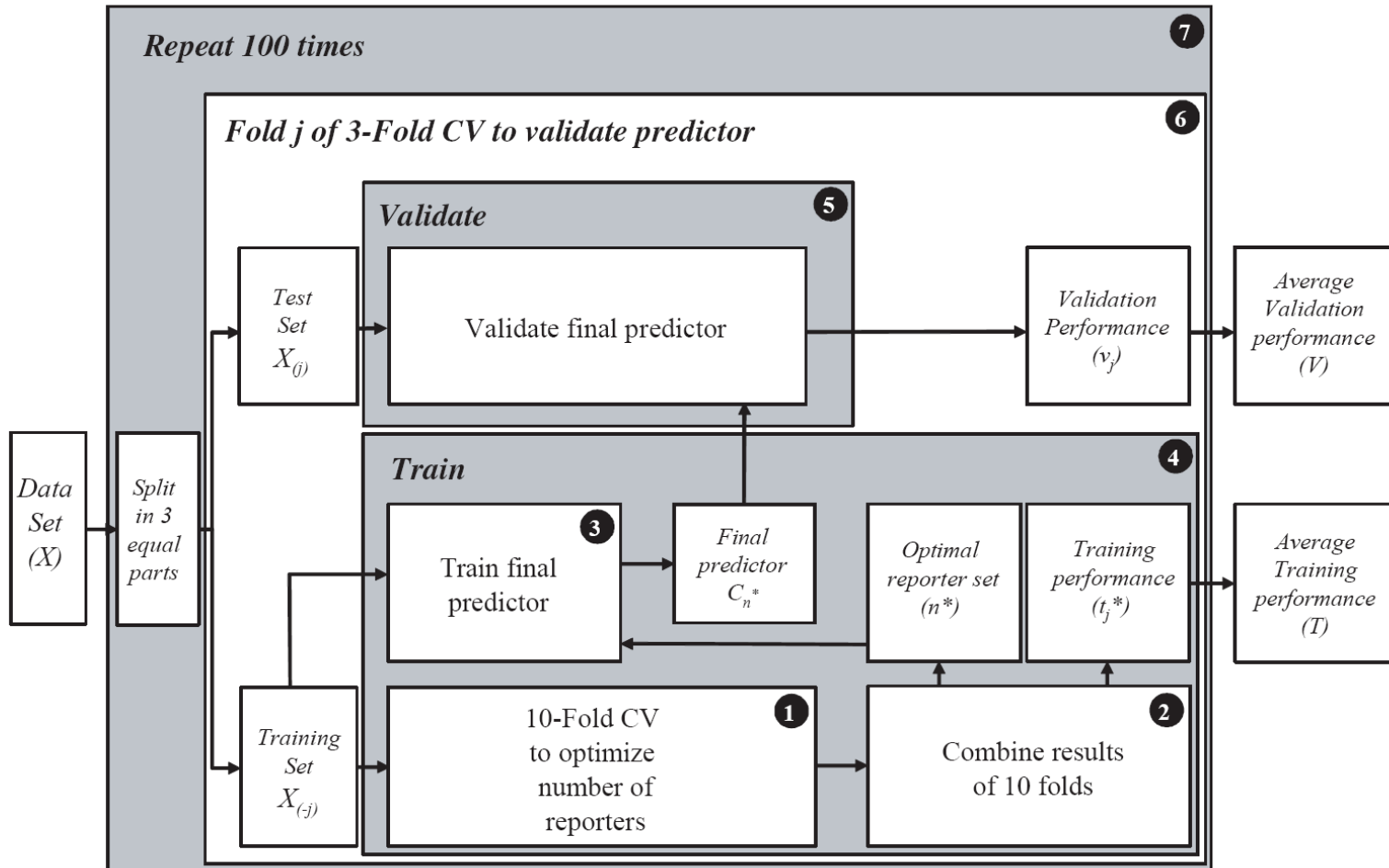
NO. AT RISK

Good signature	55	54	53	42	28	14	7
Poor signature	89	74	56	43	26	16	8

NO. AT RISK

Good signature	55	55	54	43	30	19	11
Poor signature	89	81	68	50	29	19	9

Comparison (Nested Cross-validation)



Simple approaches perform best

Reporter selection	Predictor	<i>T</i> (%) Mean	<i>T</i> (%) SD	<i>V</i> (%) Mean	<i>V</i> (%) SD	<i>k</i> *	W	D	L
Filter	NMC	64.3	3.8	62.7	3.1	92	152	6	142
	DLDC	62.3	3.9	60.6	3	88	119	9	172
	SBGC	60.8	4.2	59.7	3	69	103	6	191
	1NN	61.2	4	60.3	3.5	95	109	4	187
	5NN	61.2	4	59.3	3.5	102	108	4	188
	9NN	60.7	4.2	58.8	3.3	88	90	4	206
	RFLD[0]	61.1	5.6	59.2	4	96	97	3	200
	RFLD[1]	61.8	5.6	60.7	3.8	93	117	4	179
	RFLD[10]	63.4	4	61.8	3.1	86	132	4	164
	LinSVC	60.6	4.3	60.6	3.6	102	111	2	187
	PLS	NMC	62.5	3.1	61.7	2.2	12.4	138	4
DLDC		61.6	3.6	60.1	3.2	12.4	116	3	181
SBGC		61.2	3.9	58	3.6	10.7	92	2	206
1NN		56.6	2.9	51.9	3.6	10.1	29	1	270
5NN		56.7	3	52.7	3	8.9	29	1	270
9NN		56	3	52.5	2.7	8.4	30	0	270
RFLD[0]		59.5	3.5	55.7	3.5	9.2	67	2	231
RFLD[1]		59.5	3.5	55.7	3.4	9.2	66	3	231
RFLD[10]		60.8	3.6	58.6	3.2	11.6	89	4	207
LinSVC		59.2	3.7	56.4	3.2	12.1	64	2	234
SC		SC	65	3.4	62.9	1.9	909	—	—
RFE	LinSVC	62.8	3.8	59.8	3.4	648	107	4	189

Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer

Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, John A Foekens

Lancet 2005; 365: 671-79

See [Comment](#) page 634

Veridex LLC, a Johnson & Johnson Company, San Diego, CA, USA (Y Wang PhD, Y Zhang PhD, F Yang MSc, D Talantov MD, J Yu PhD, T Jatkoe BSc); Veridex LLC, a Johnson & Johnson Company, Warren, NY, USA

Summary

Background Genome-wide measures of gene expression can identify patterns of gene activity that subclassify tumours and might provide a better means than is currently available for individual risk assessment in patients with lymph-node-negative breast cancer.

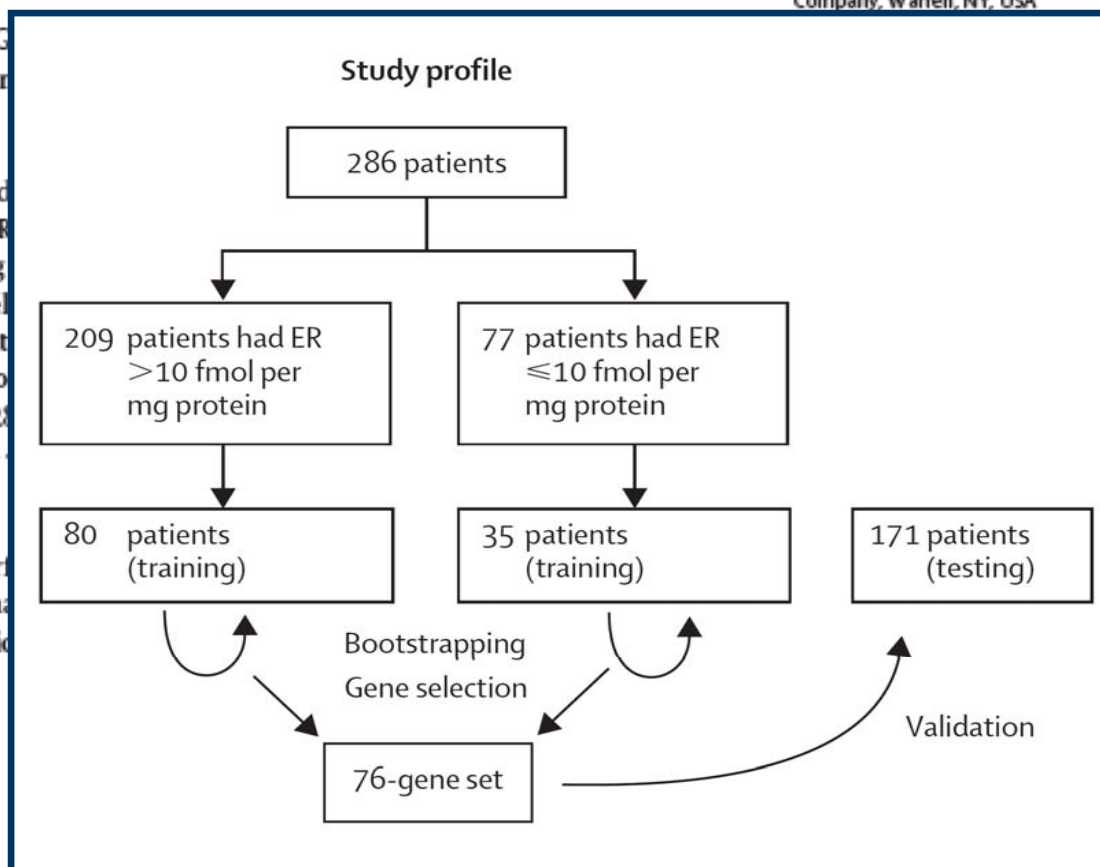
Methods We analysed, with Affymetrix Human U133a C RNA of frozen tumour samples from 286 lymph-node-negative breast cancer patients before and after treatment.

Findings In a training set of 115 tumours, we identified a 76-gene profile that was positive for oestrogen receptors (ER) and 16 genes for ER-negative tumours. This profile had 80% sensitivity and 48% specificity in a subsequent independent testing set. The 76-gene profile also represented a strong prognostic factor in the subgroups of 84 premenopausal patients (9-60 [2-22] months) and 79 patients with tumours of 10-20 mm (14-1 [3-22] months). Prognosis is especially difficult.

Interpretation The identified signature provides a powerful prognostic tool. The ability to identify patients who have a high risk of distant recurrence, allow clinicians to avoid adjuvant systemic therapy.

Introduction

About 60-70% of patients with lymph-node-negative breast cancer are cured by local or regional treatment alone.^{1,2} The most widely used treatment guidelines are based on the extent of disease at diagnosis, the extent of



Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer

Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, John A Foekens

Lancet 2005; 365: 671-79

See [Comment](#) page 634

Veridex LLC, a Johnson & Johnson Company, San Diego, CA, USA (Y Wang PhD, Y Zhang PhD, F Yang MSc, D Talantov MD, J Yu PhD, T Jatkoe BSc); Veridex LLC, a Johnson & Johnson Company, Warren, NY, USA

Summary

Background Genome-wide measures of gene expression can identify patterns of gene activity that subclassify tumours and might provide a better means than is currently available for individual risk assessment in patients with lymph-node-negative breast cancer.

Methods We analysed, with Affymetrix Human U133a GeneChip, the RNA of frozen tumour samples from 286 lymph-node-negative breast cancer patients before treatment.

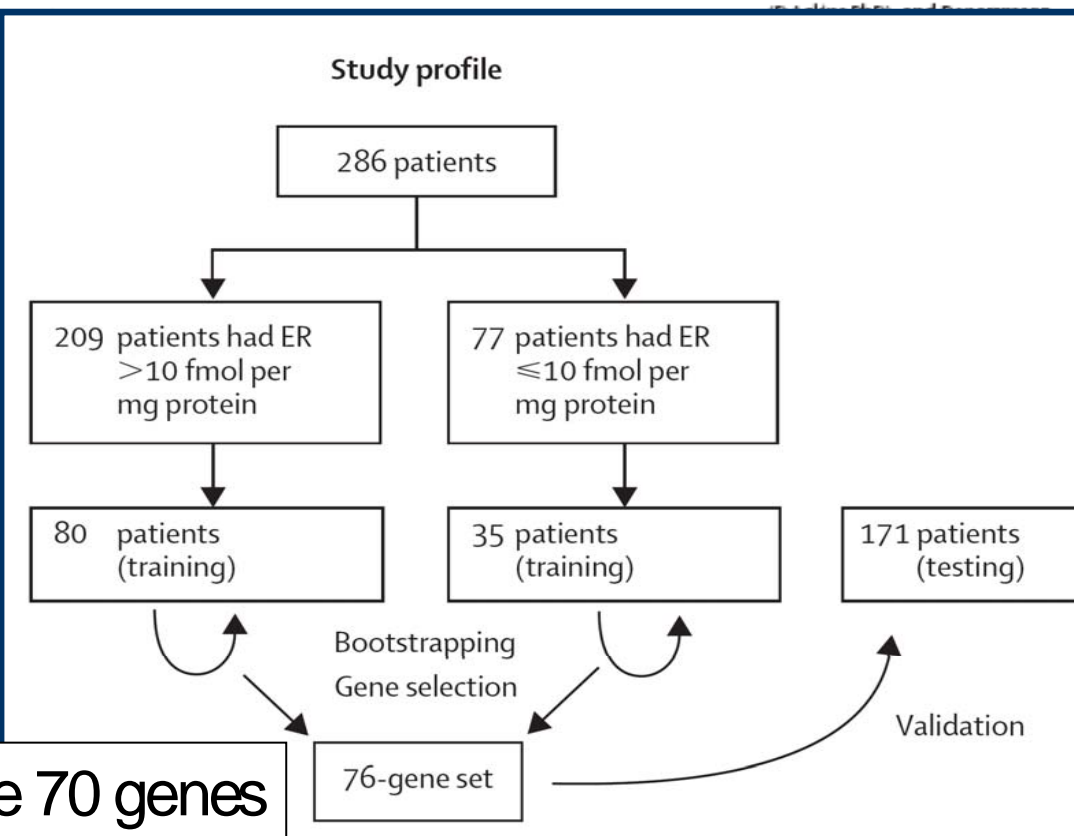
Findings In a training set of 115 tumours, we identified 76 genes that were positive for oestrogen receptors (ER) and 16 genes for ER- and 48% specificity in a subsequent independent testing set. This 76-gene profile was highly informative in identifying patients who developed distant metastasis [95% CI 2.59-12.4], even when corrected for traditional prognostic factors [2.46-12.5]. The 76-gene profile also represented a strong prognostic factor in the subgroups of 84 premenopausal patients (9.60 [2.28-39.1]) and 79 patients with tumours of 10-20 mm (14.1 [3.3-57.1]). Distant metastasis prognosis is especially difficult.

Interpretation The identified signature provides a powerful prognostic tool for breast cancer recurrence. The ability to identify patients who have a high risk of distant metastasis, allow clinicians to avoid adjuvant systemic therapy.

Introduction

About 60-70% of patients with lymph-node-negative breast cancer develop distant metastasis.

alone.



Prediction of cancer outcome with microarrays: a multiple random validation strategy

Lancet 2005; 365: 488–92

Stefan Michiels, Serge Koscielny, Catherine Hill

See [Comment](#) page 454

Biostatistics and Epidemiology Unit (S Michiels MSc, S Koscielny PhD, C Hill PhD), Functional Genomics Unit (S Michiels), and Inserm U605 (S Koscielny), Institut Gustave Roussy, Villejuif, France

Correspondence to: Dr Serge Koscielny, Biostatistics and Epidemiology Unit, Institut Gustave Roussy, 39 rue Camille Desmoulins, 94805 Villejuif, France
koscielny@igr.fr

Summary

Background General studies of microarray gene-expression profiling have been undertaken to predict cancer outcome. Knowledge of this gene-expression profile or molecular signature should improve treatment of patients by allowing treatment to be tailored to the severity of the disease. We reanalysed data from the seven largest published studies that have attempted to predict prognosis of cancer patients on the basis of DNA microarray analysis.

Methods The standard strategy is to identify a molecular signature (ie, the subset of genes most differentially expressed in patients with different outcomes) in a training set of patients and to estimate the proportion of misclassifications with this signature on an independent validation set of patients. We expanded this strategy (based on unique training and validation sets) by using multiple random sets, to study the stability of the molecular signature and the proportion of misclassifications.

Findings The list of genes identified as predictors of prognosis was highly unstable; molecular signatures strongly depended on the selection of patients in the training sets. For all but one study, the proportion misclassified decreased as the number of patients in the training set increased. Because of inadequate validation, our chosen studies published overoptimistic results compared with those from our own analyses. Five of the seven studies did not classify patients better than chance.

Interpretation The prognostic value of published microarray results in cancer studies should be considered with caution. We advocate the use of validation by repeated random sampling.

Introduction

The expression of several thousand genes can be studied simultaneously by use of DNA microarrays. These microarrays have been used in many specialties of medicine. In oncology, their use can identify genes with different expressions in tumours with different outcomes.^{1–9} These gene-expression profiles or molecular signatures are expected to assist in the selection of

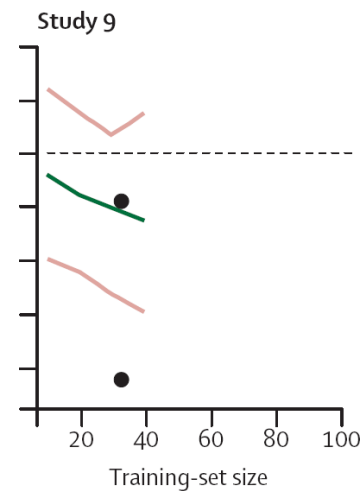
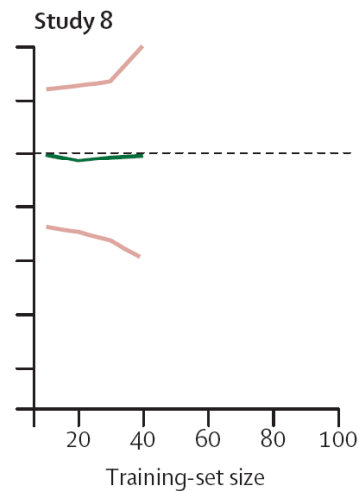
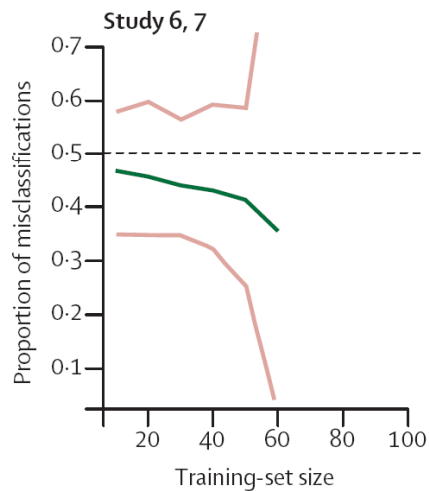
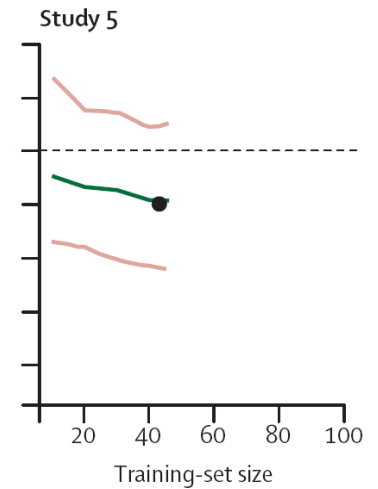
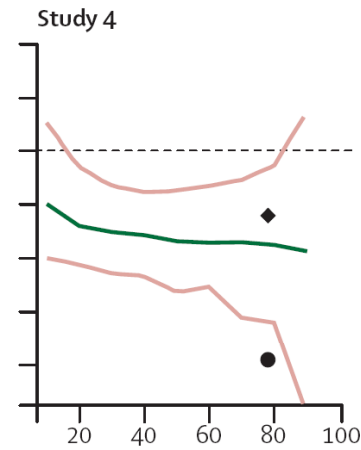
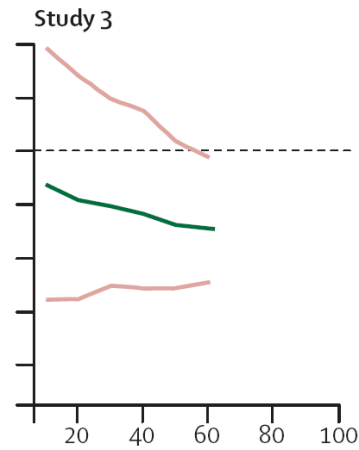
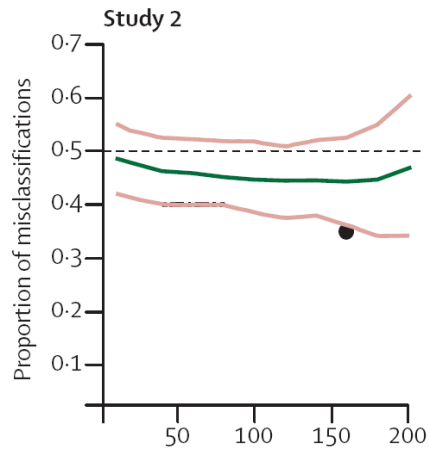
guidelines (Minimum Information About a Microarray Experiment¹⁵). This approach offers an opportunity to propose alternative analyses of these data. We have taken advantage of this opportunity to analyse different datasets from published studies of gene expression as a predictor of cancer outcome. We aimed to assess the extent to which the molecular signature depends on the constitution of the training set, and to study the

Studies evaluated

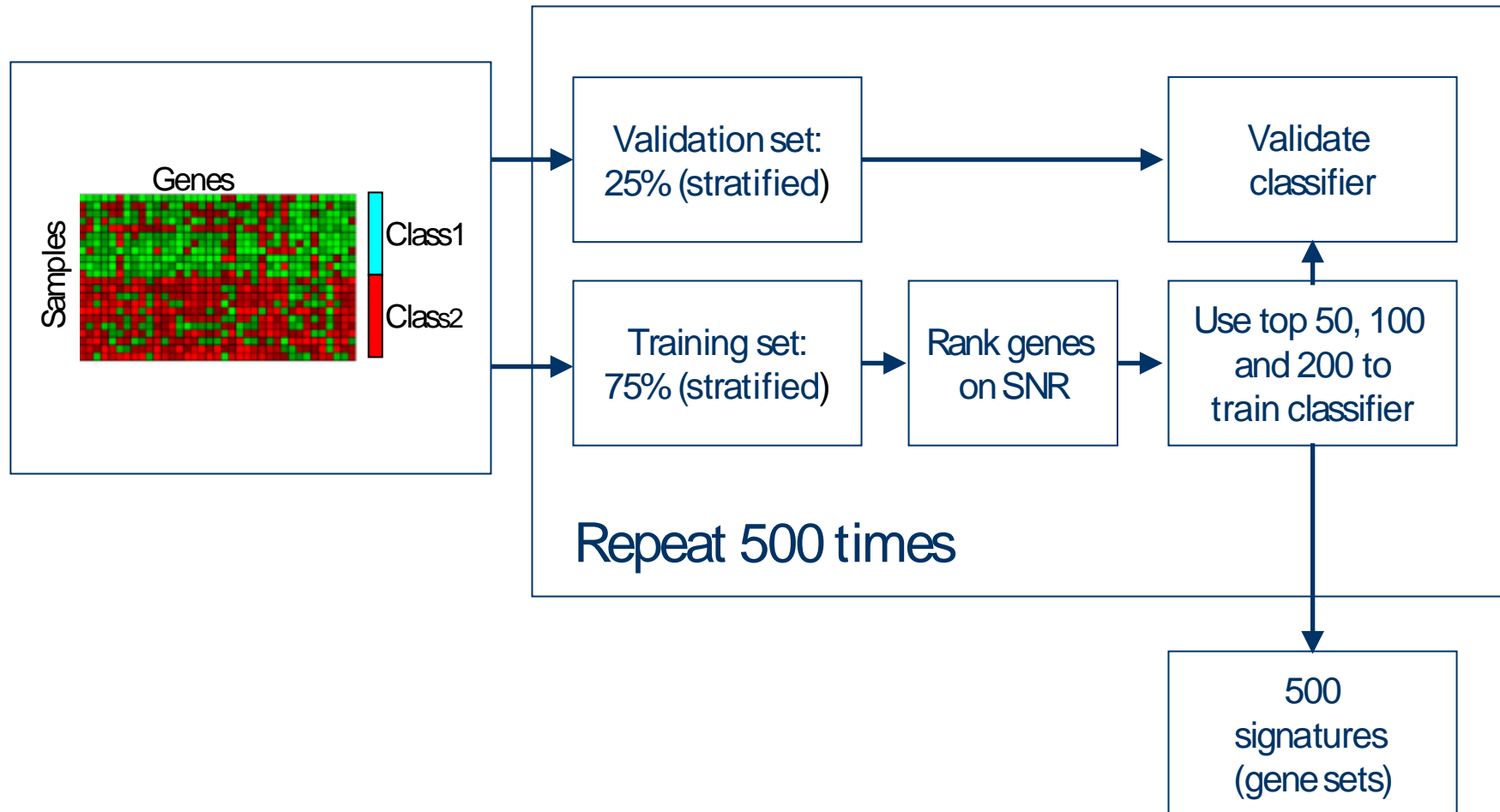
Outcome prediction based on gene expression

Author	Tumor type	Classlabel
Rosenwald	Non-Hodgkin lymphoma	Survival
Yeoh	Acute lymphocytic leukaemia	Relapse-free survival
van 't Veer	Breast cancer	5-year metastasis-free survival
Beer	Lung adenocarcinoma	Survival
Bhattacharjee/Ramaswamy	7 Lung adenocarcinoma	4-year survival
Pomeroy	Medulloblastoma	Survival
Iizuka	Hepatocellular carcinoma	1-year recurrence-free survival

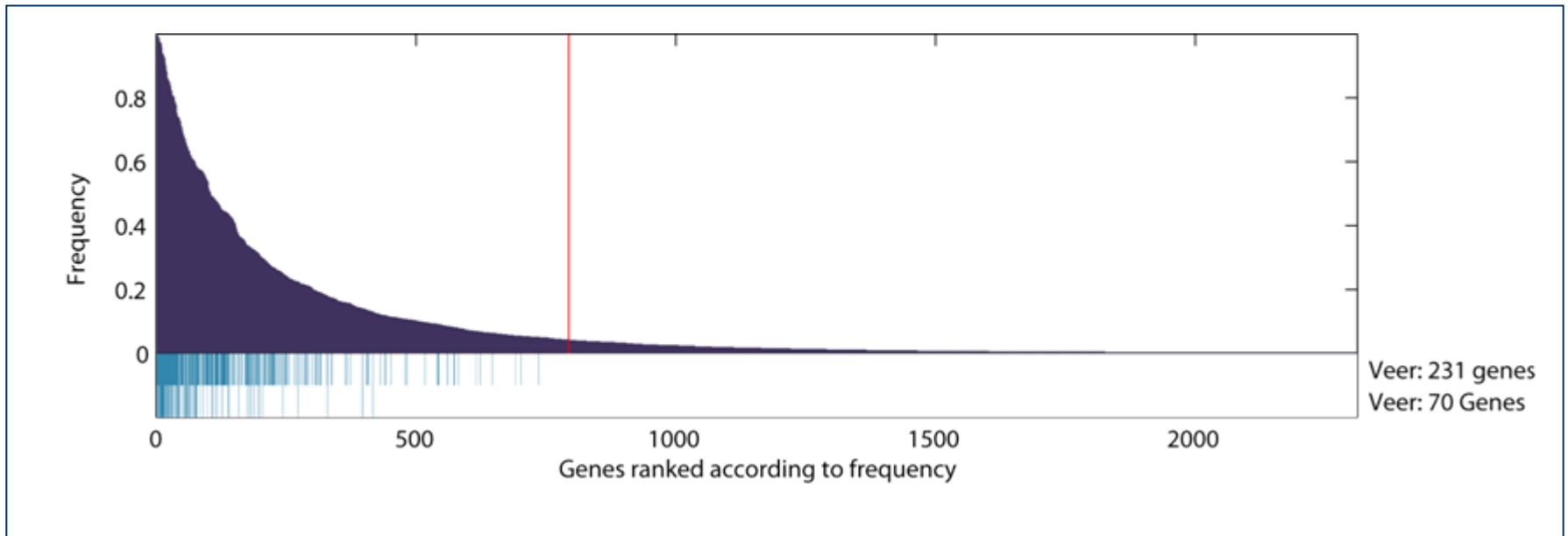
Michiels *et al.* Lancet 2005. Results I



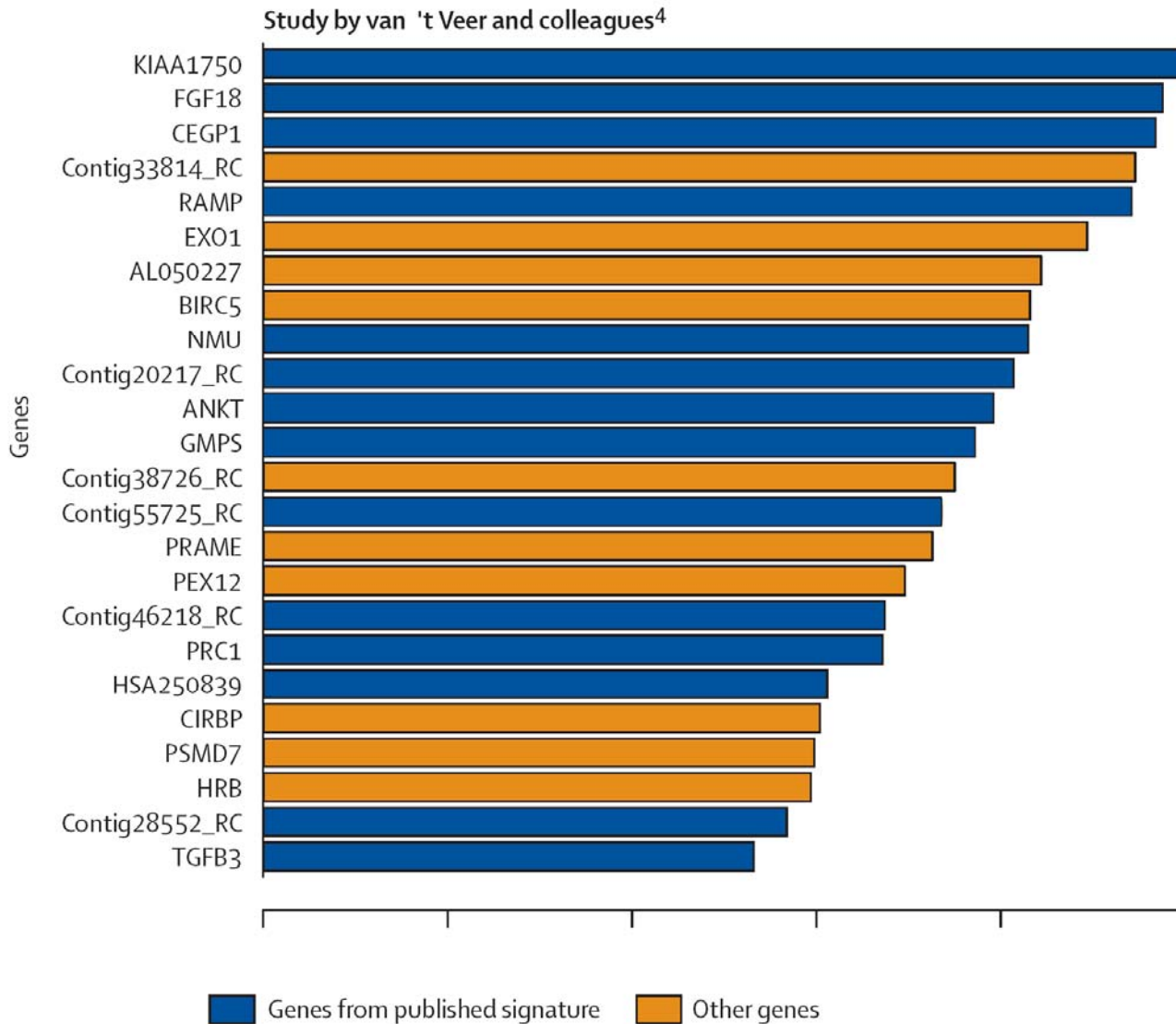
Repeated Random Resampling (1)



Repeated Random Resampling (2)



Michiels *et al.* Lancet 2005. Results II



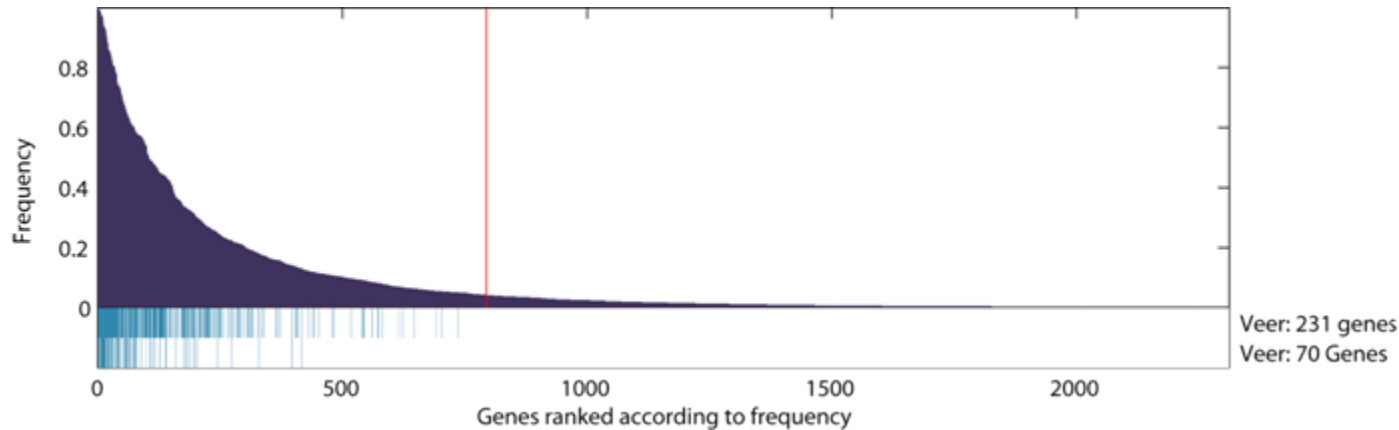
Michiels *et al.* Lancet 2005. Findings:

- The list of genes was highly unstable
- signatures strongly depended on the selection of patients
- proportion misclassified decreased as # patients increased
- published overoptimistic results (inadequate validation)
- 5/7 did not classify patients better than chance.

Michiels *et al.* Lancet 2005. Conclusions:

- ‘The prognostic value of published microarray results in cancer studies should be considered with caution.’
- ‘We advocate the use of validation by repeated random sampling’.

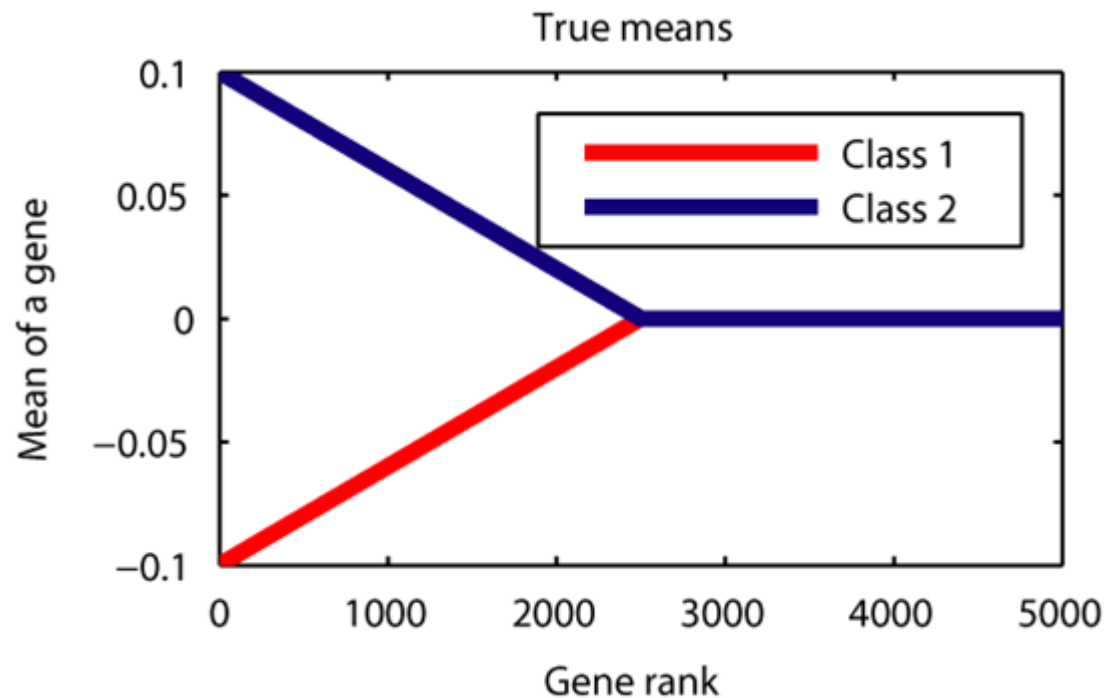
How good is the gold standard?



- True gene set not known
- Design artificial dataset with known ranking

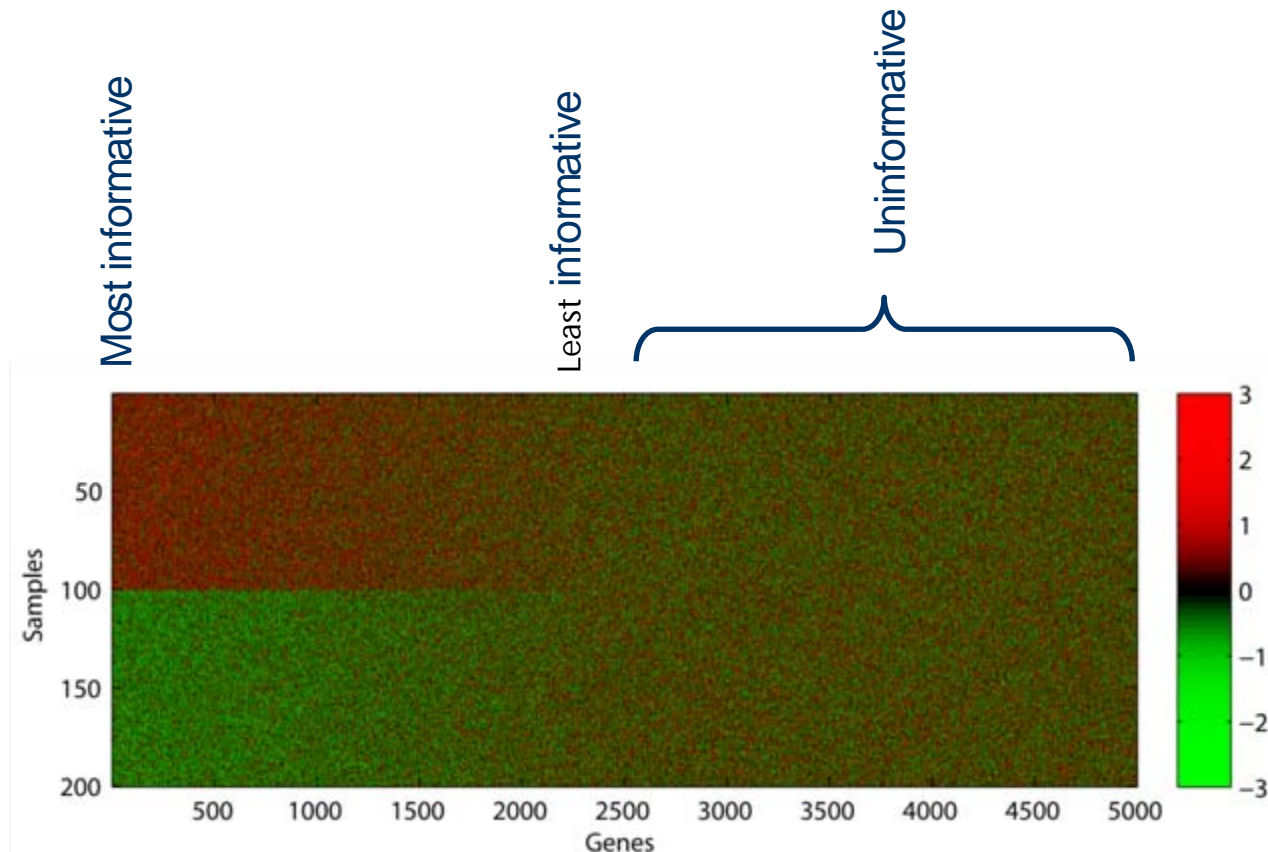
Artificial data

- 2-classes
- class conditional: Normal distributions, $N(\mu, 1)$
- genes assumed to be independent

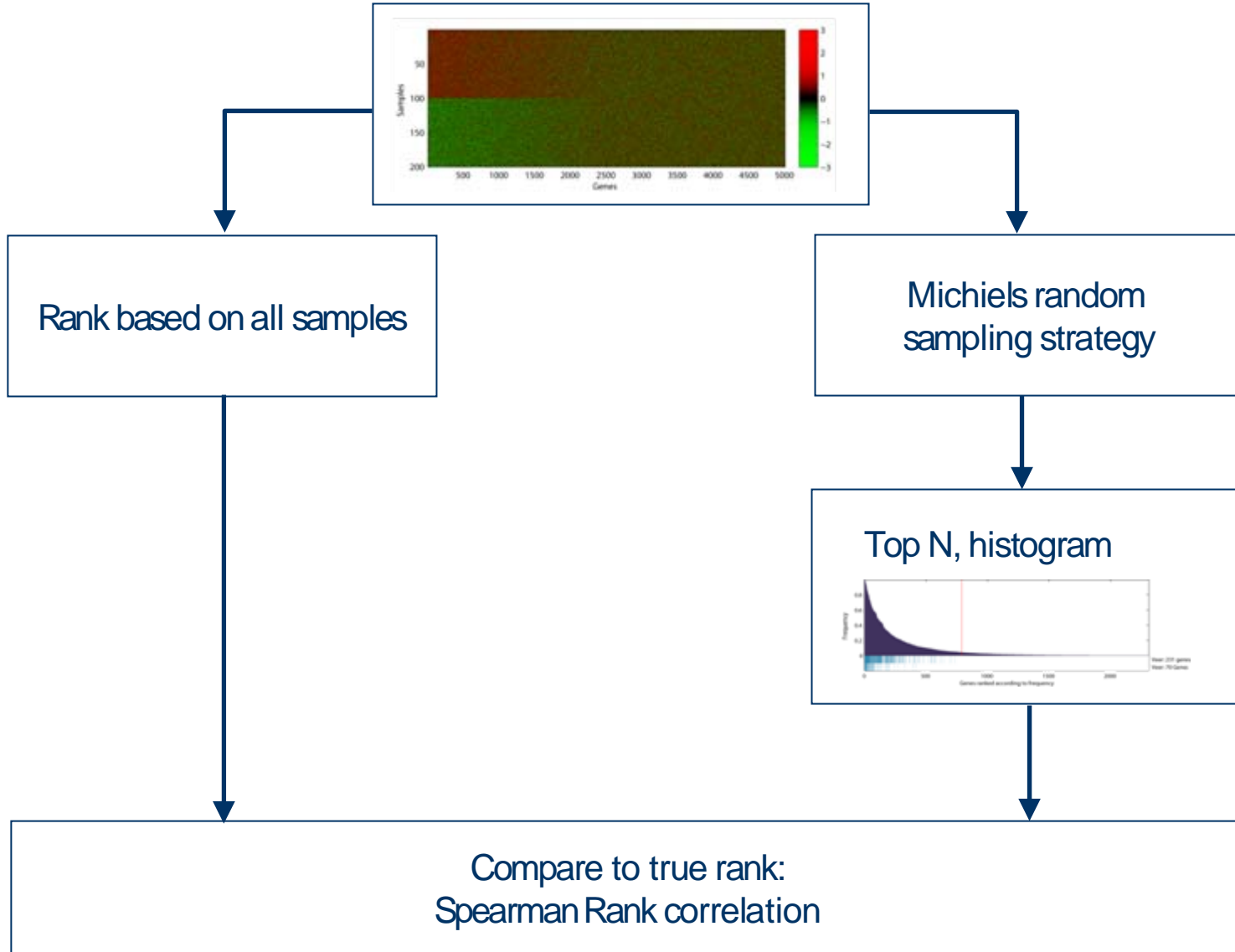


Experiment on artificial data (2)

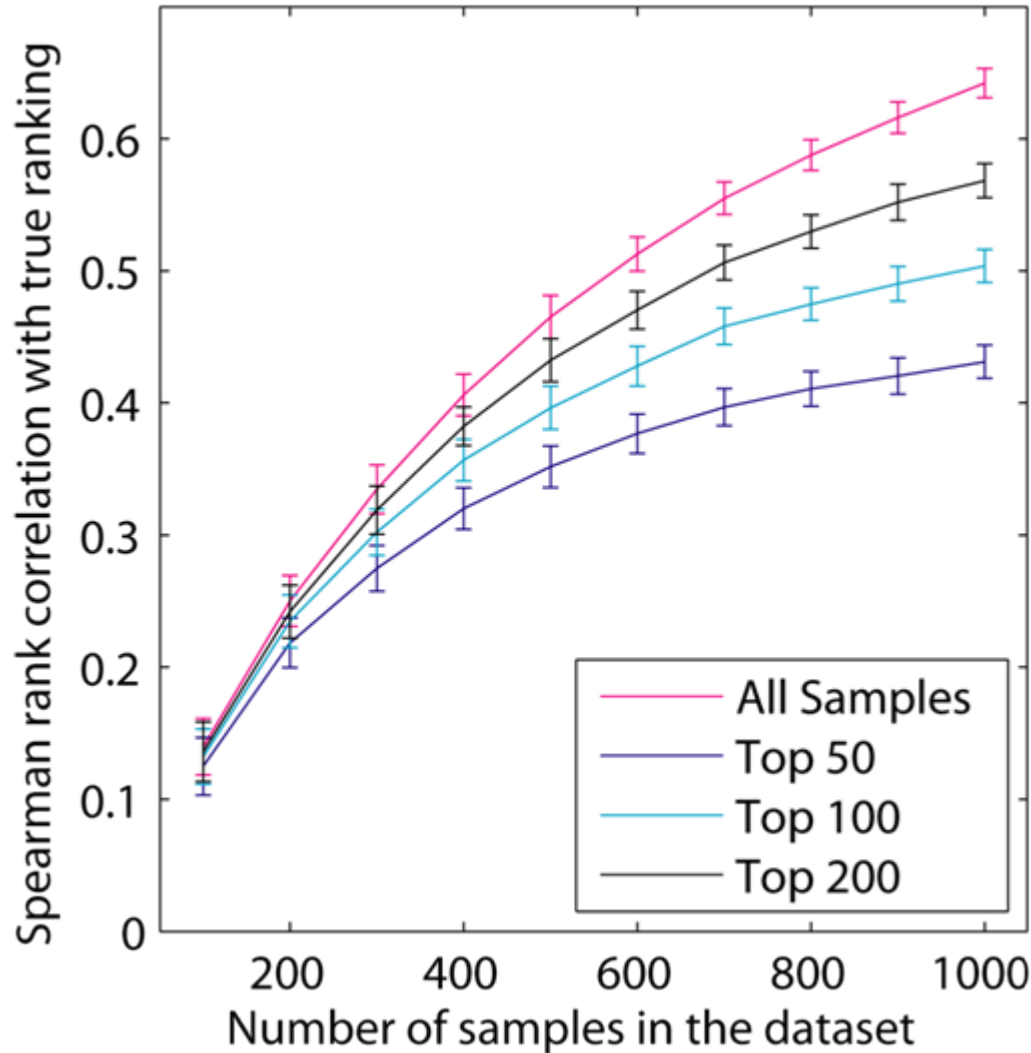
- $p=5000$ genes (2500 informative)
- $n=200$ samples (100 per class)



Experiment on artificial data (3)



Artificial data: gene selection results

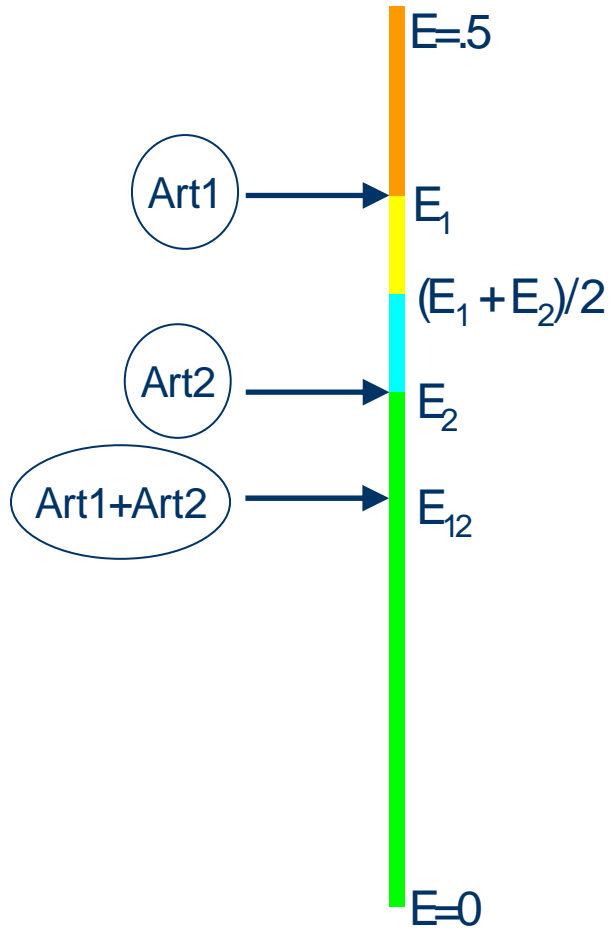


From subsets to datasets

- When sampling from one BC dataset, Michiels observed signature instability
- We currently have 6 BC datasets totaling 947 samples
- BC datasets are resamplings from the BC population.
- Limited signature overlap = signature instability when subsampling
- Given 947 samples:
 - ML: pool the data
 - Heterogeneity of data may be detrimental
 - Investigate effects of pooling on performance, signature stability
- First look at pooling of 6 artificial datasets sampled from the artificial model

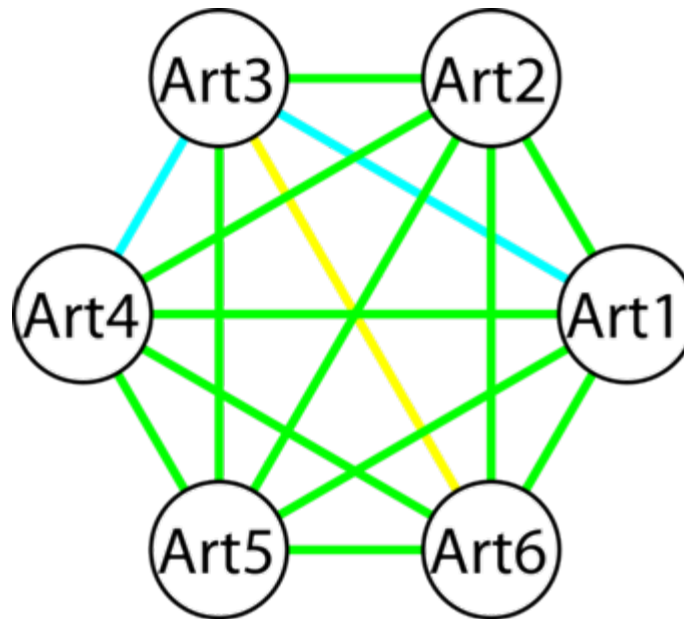
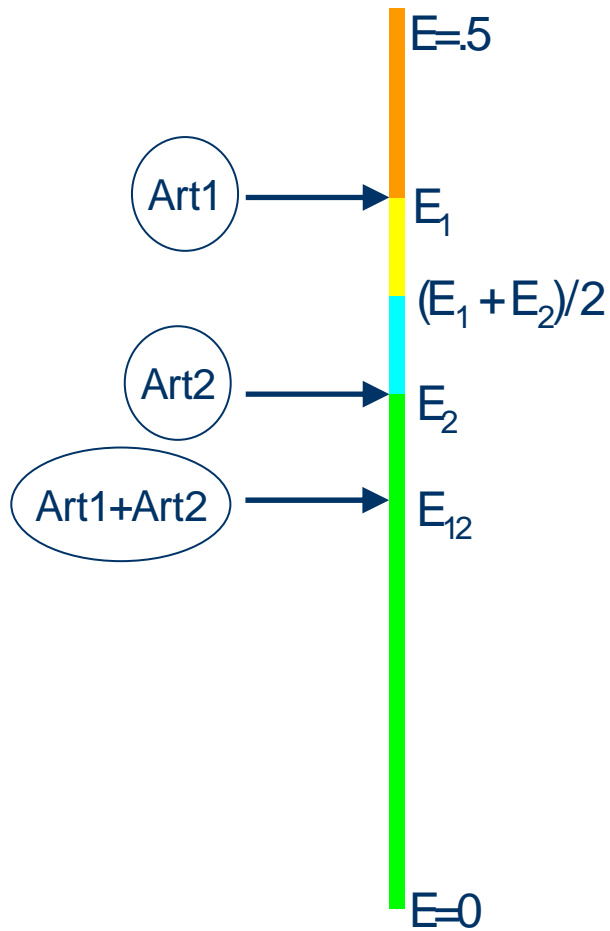
Pooling 2 artificial datasets





Double Loop Cross
Validation Error



Pooling 2 artificial datasets

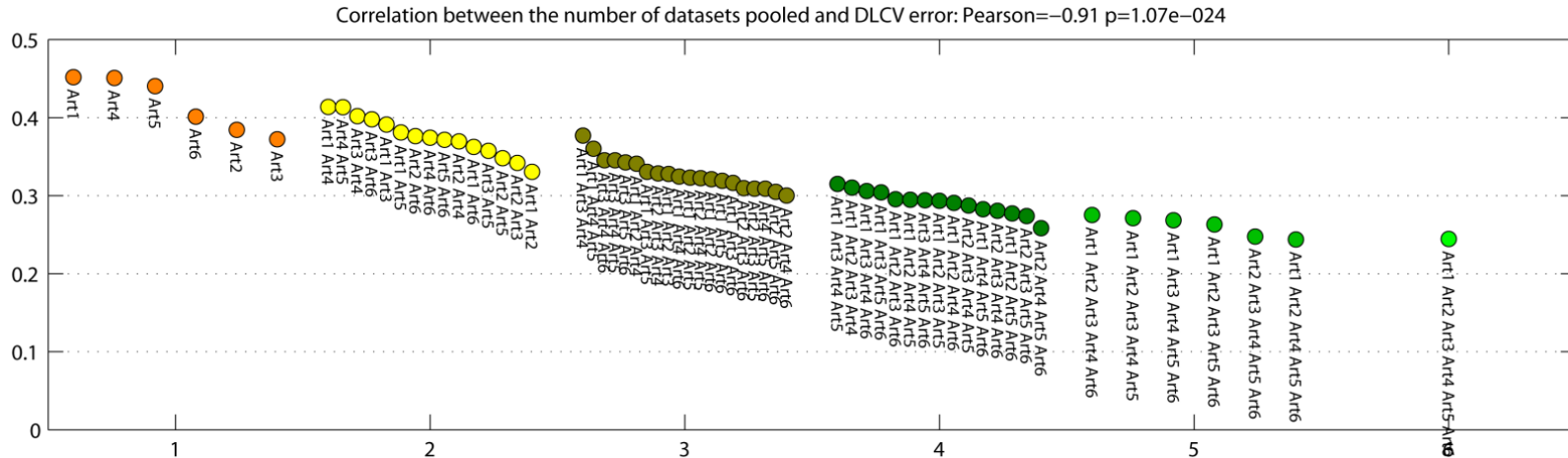
Double Loop Cross
Validation Error



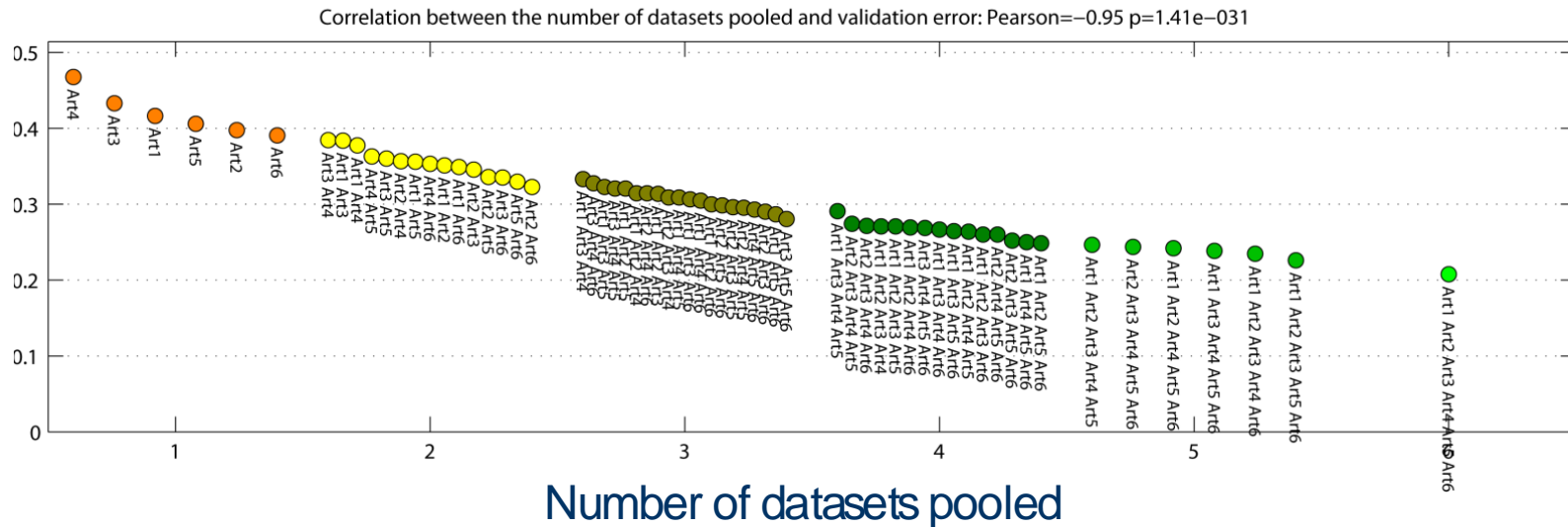
-  Synergy
-  Marginal Synergy
-  Marginal Anti-Synergy
-  Anti-Synergy

Pooling 1 to 6 artificial datasets

DLCV error

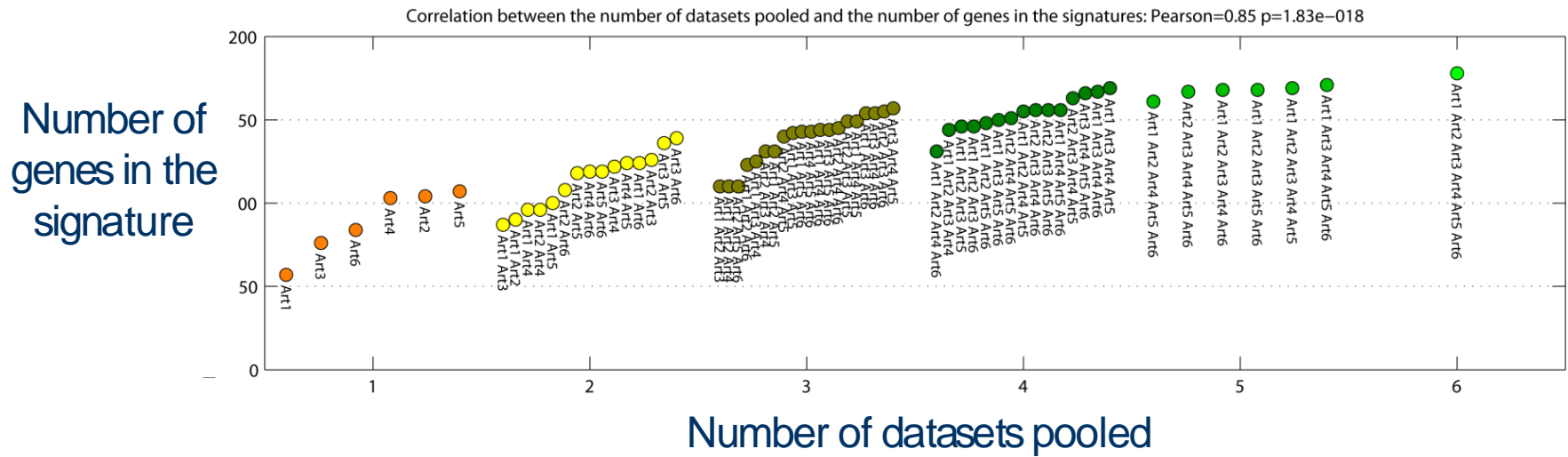


Large validation set

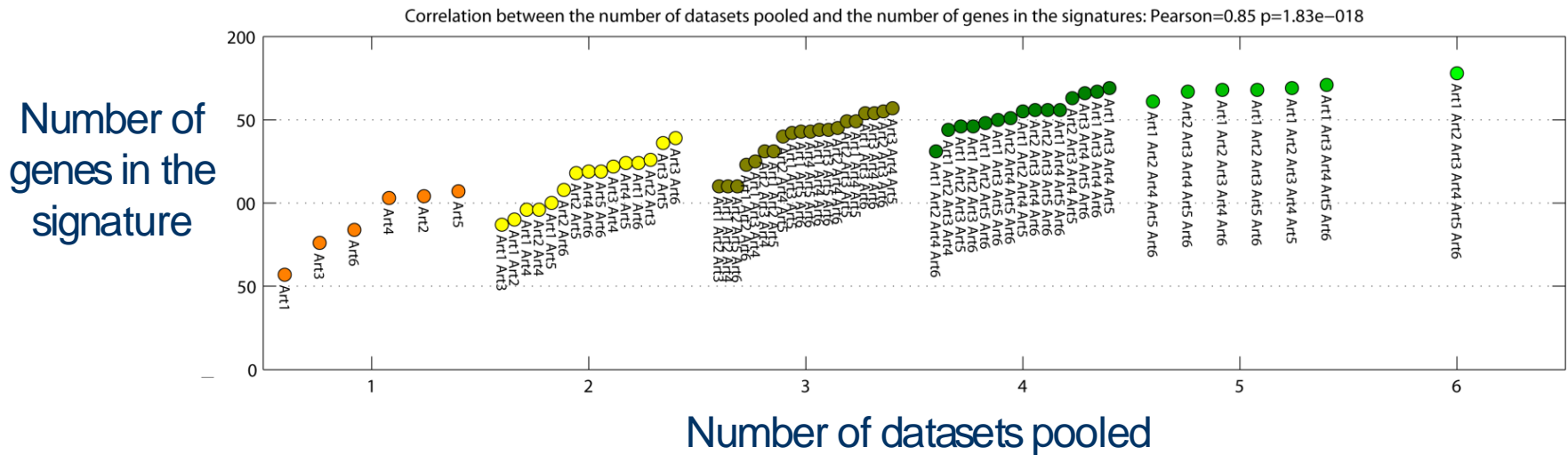


Number of datasets pooled

Pooling 1 to 6 artificial datasets



Pooling 1 to 6 artificial datasets



Significant trends between number of pooled datasets and

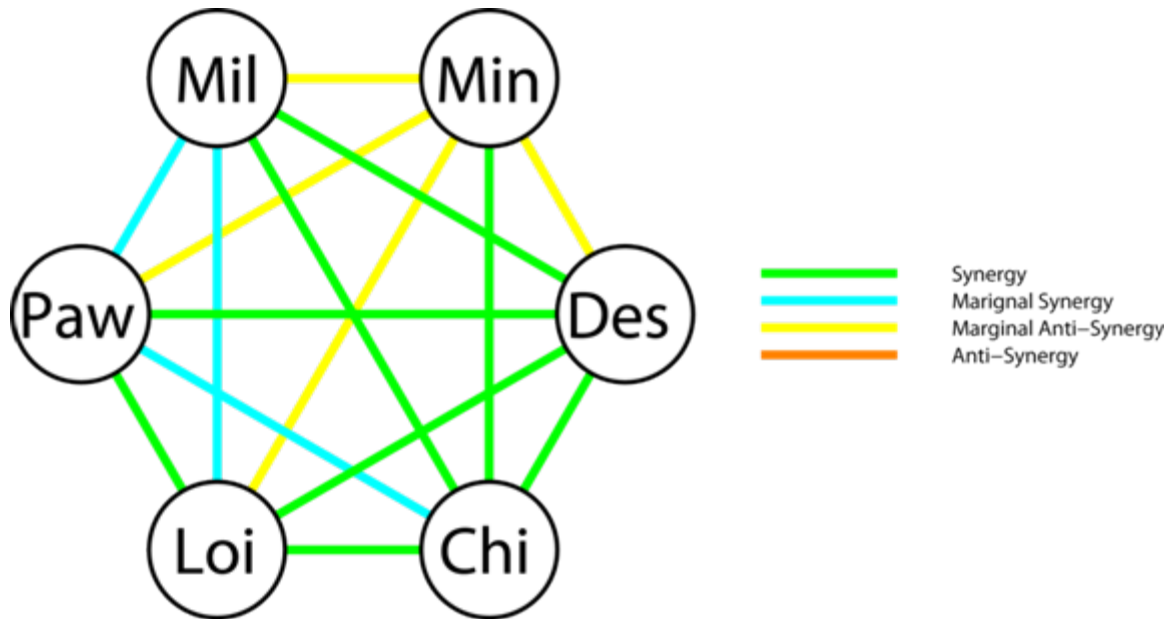
- DLCV error
- large validation set error
- signature size

2008: Seven breast cancer datasets

- Six datasets (all Affymetrix HG U133A)
- Same pre-processing applied
- Overlapping samples discarded
- Seventh dataset used for validation (Vijver et al.)

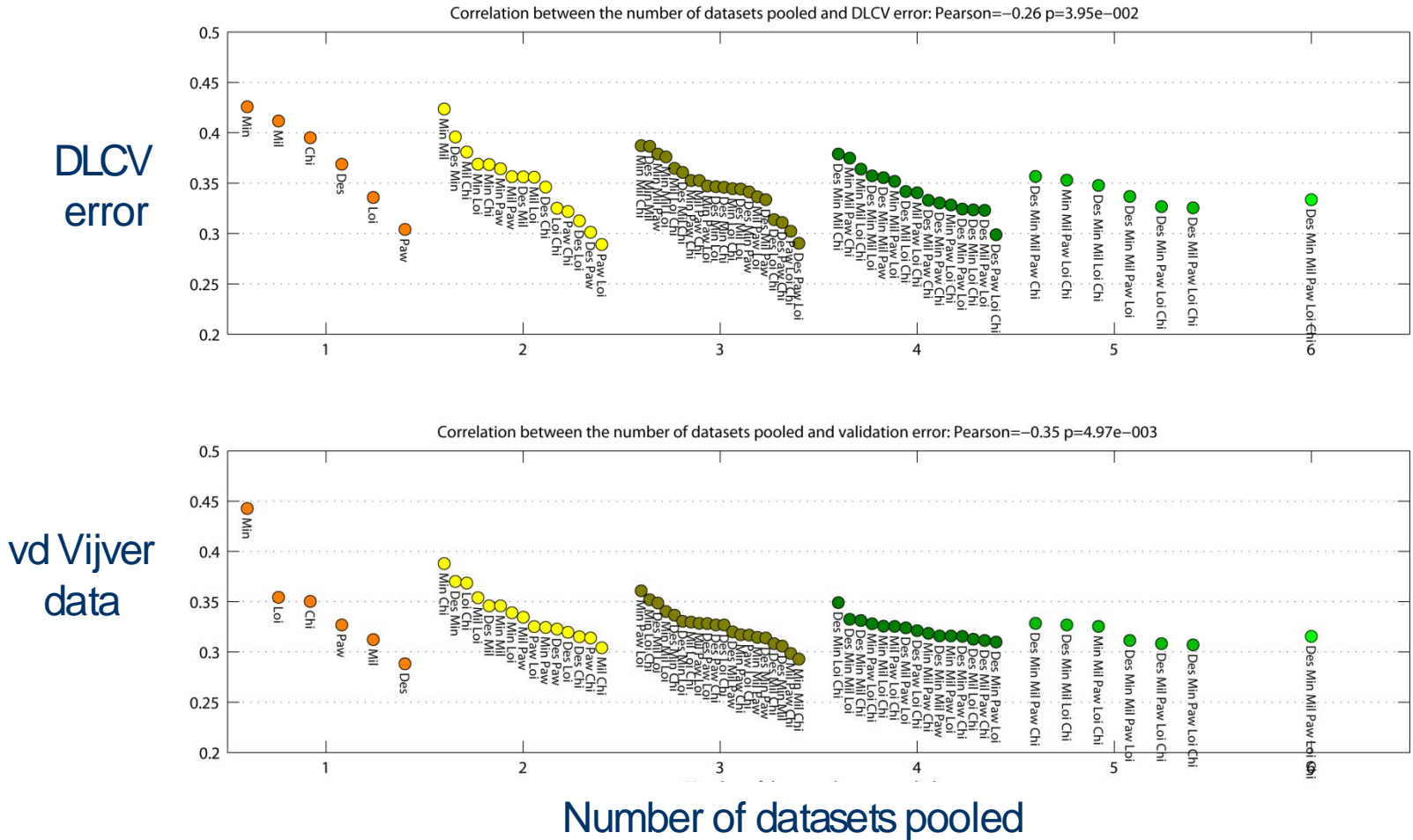
Publication:	Label	Survival	Total
Desmedt et al.	Des	DMFS	147
Minn et al.	Min	DMFS	96
Miller et al.	Mil	SOS	247
Pawitan et al.	Paw	SOS	156
Loi et al.	Loi	DMFS	178
Chin et al.	Chi	DMFS	123
Total number of samples:			947

Pooling 2 breast cancer datasets



Significant synergy: 11/15 improved performance

Pooling 1 to 6 breast cancer datasets



Significant correlation ($p < 0.05$) between error and pooled number of datasets

Choice of classifier (1)

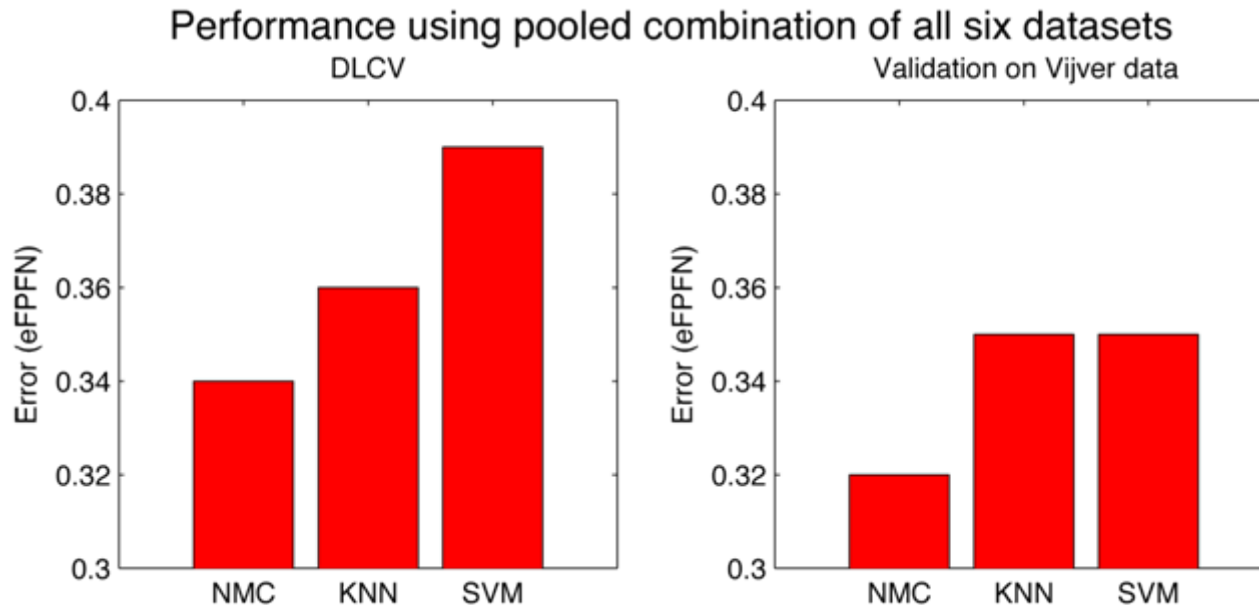
Tested classifiers:

- Nearest mean classifier
- K-Nearest Neighbor (3-NN)
- Support Vector Machine classifier (SVM)

They all show the same trends: pooling leads to a better classifier

Choice of classifier (2)

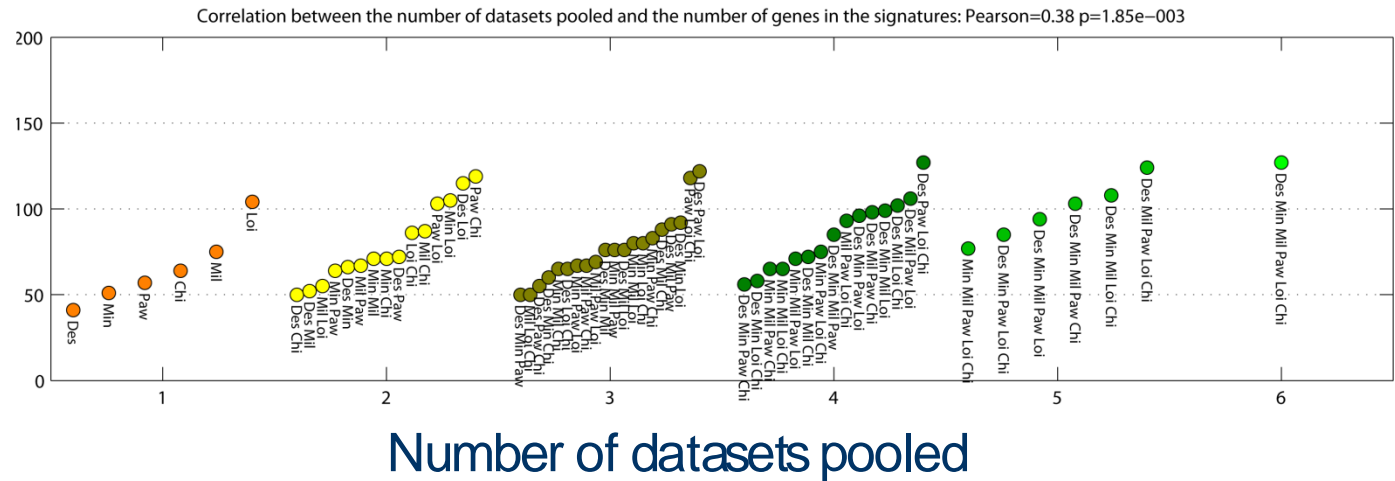
- Previous work: Nearest mean classifier performs best.
- Non-linear classifiers should benefit from more samples...



Nearest mean classifier remains the best option

Pooling 1 to 6 breast cancer datasets

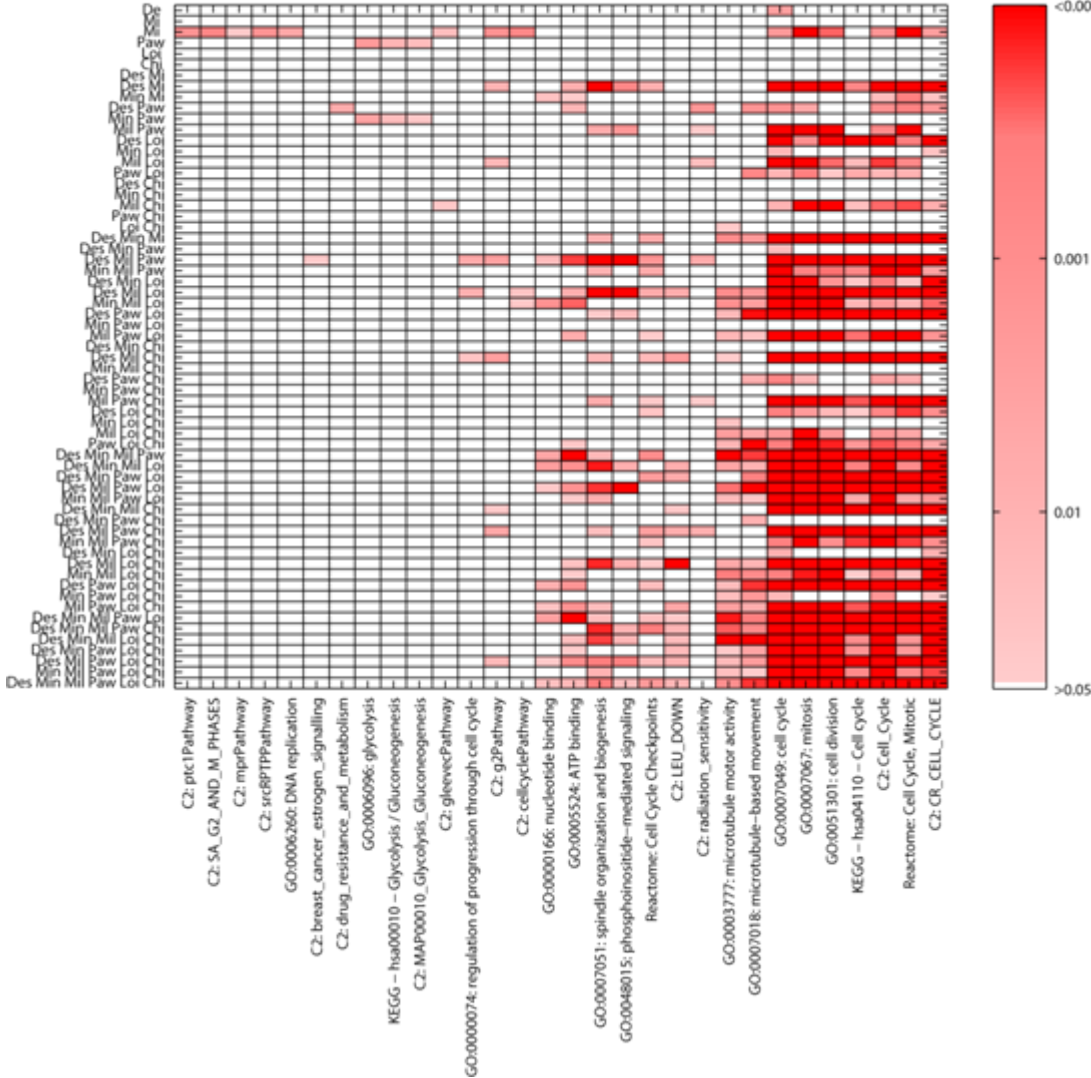
Number of genes in the signature



Same trends as on the artificial data

Functional enrichment

Signatures from each combination of pooled datasets
 Enrichment p-values, Bonferroni corrected per signature, for each gene set at least 1 enrichment p<0.05



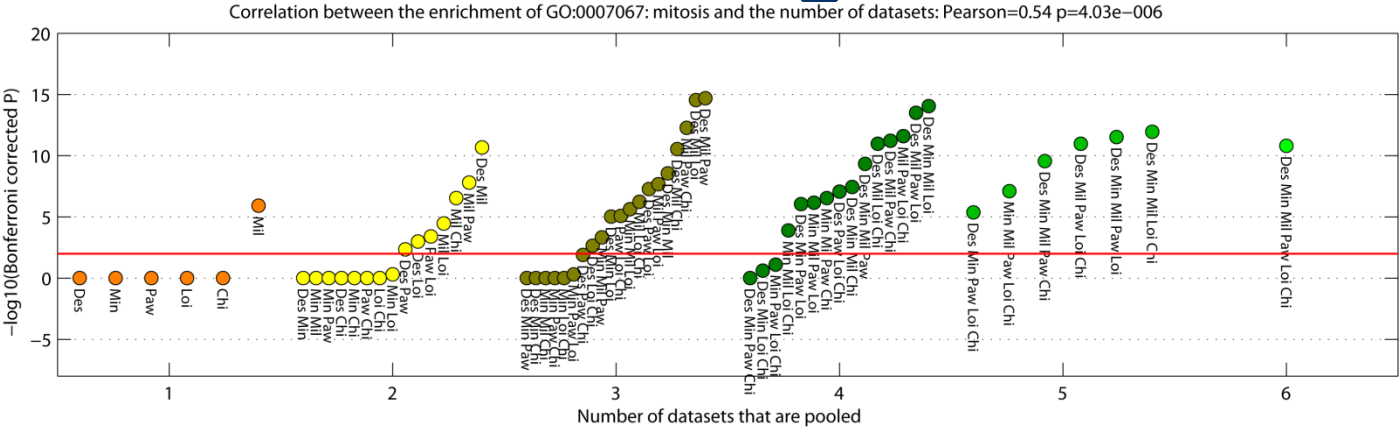
Enrichments:

- Proliferation the main component
- Some gene sets only enriched in 1 dataset

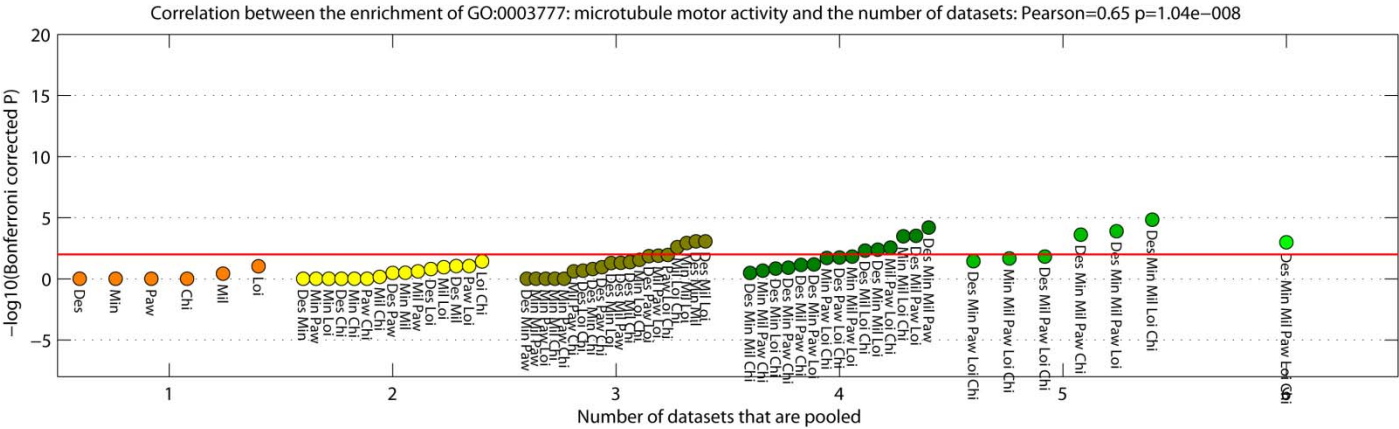
Common signals become stronger

Functional enrichment: 2 gene sets

Mitosis



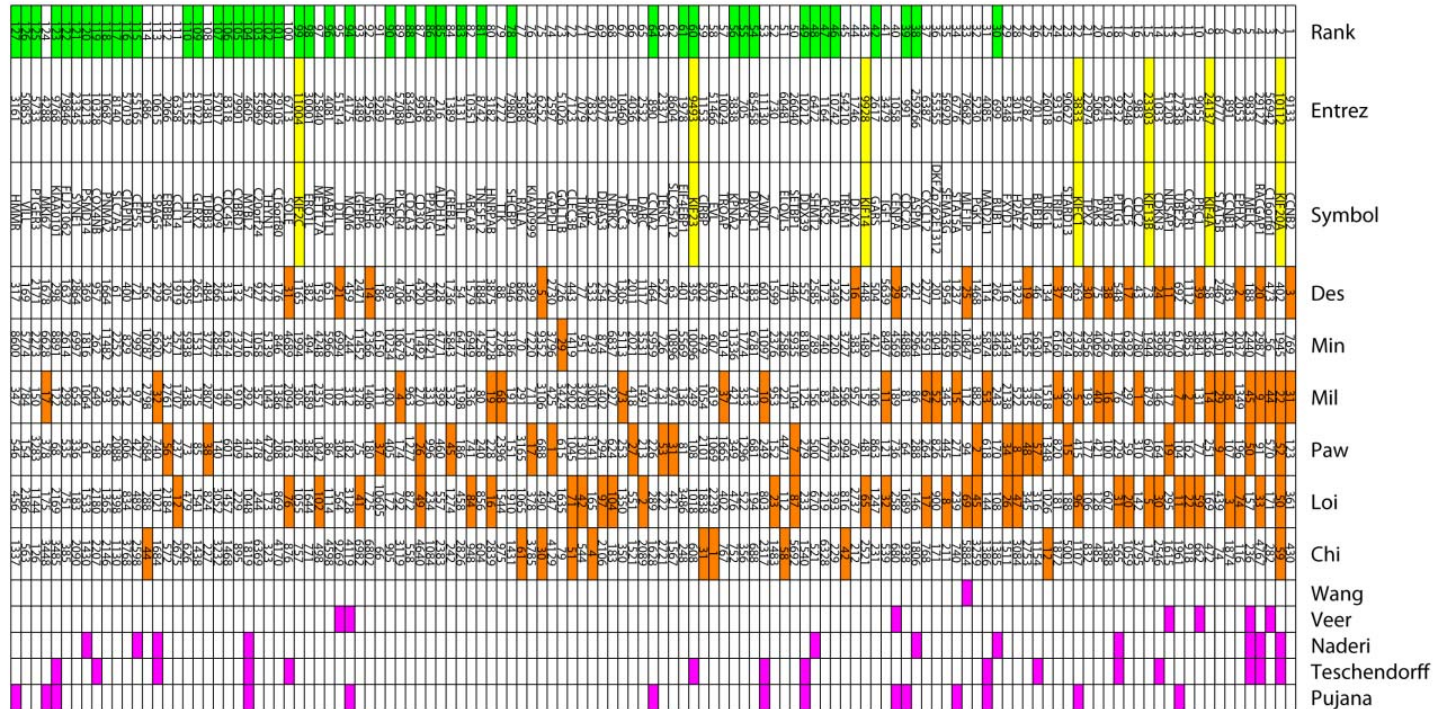
Microtubule motor activity



Number of datasets pooled

Pooling required to detect these enrichments.

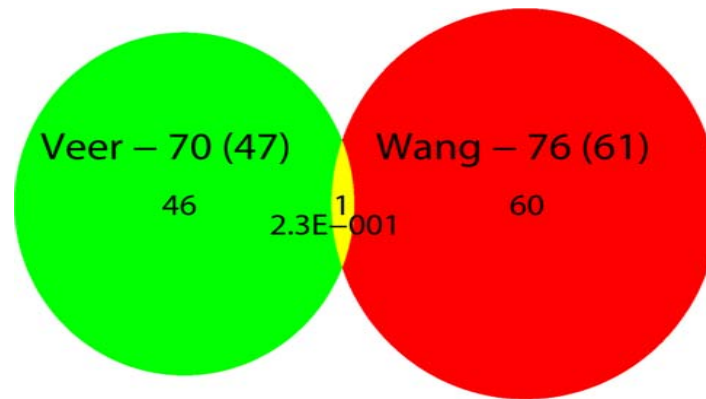
127 gene signature and the rest



	#Datasets	P
Teschendorff	3	2.25E-18
Naderi	1	1.21E-11
Veer	1	7.43E-07
Wang	1	4.68E-01

Limited overlap among signatures (1)

Two existing signatures: 1 gene overlap



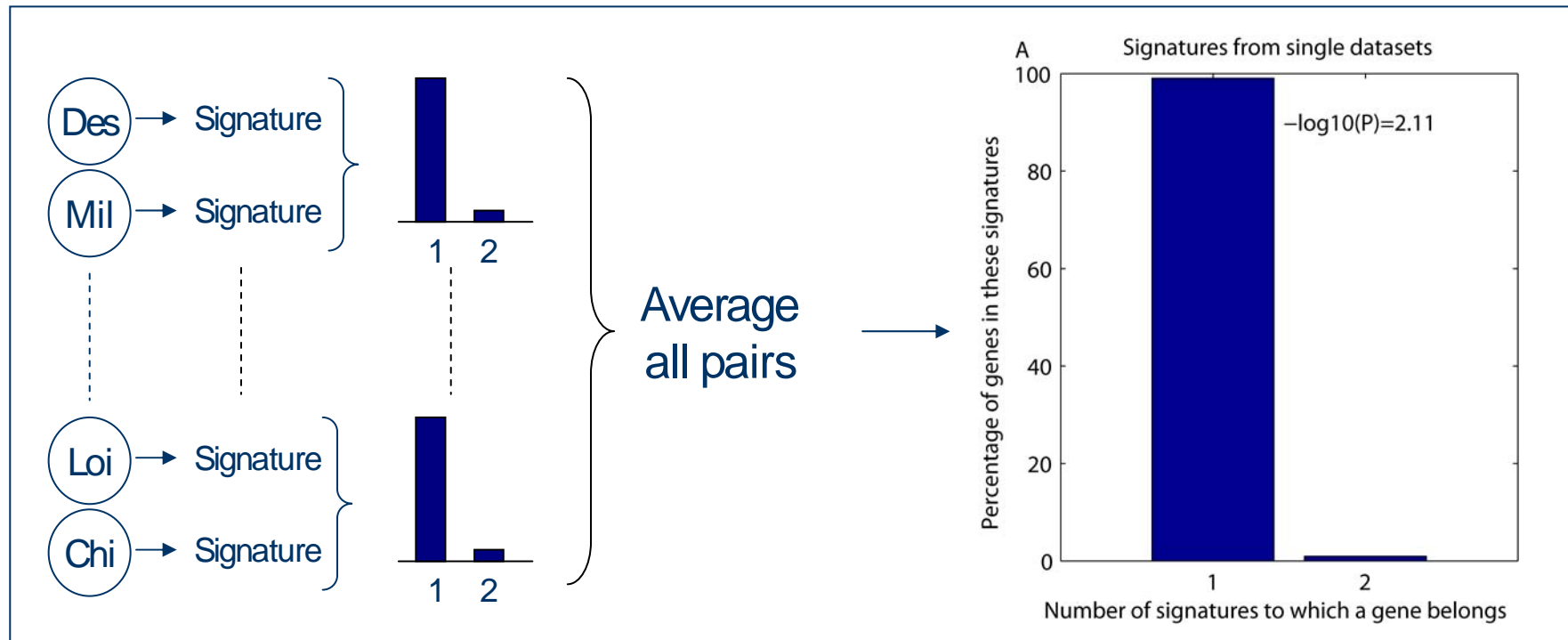
Five explanations:

1. Different platform technologies;
2. Differences in supervised protocols;
3. Dissimilar genes, the same pathways;
4. Clinical composition (i.e. sample heterogeneity);
5. Small sample size problems

Limited overlap among signatures (2)

Five explanations:

1. Different platform technologies;
2. Differences in supervised protocols;

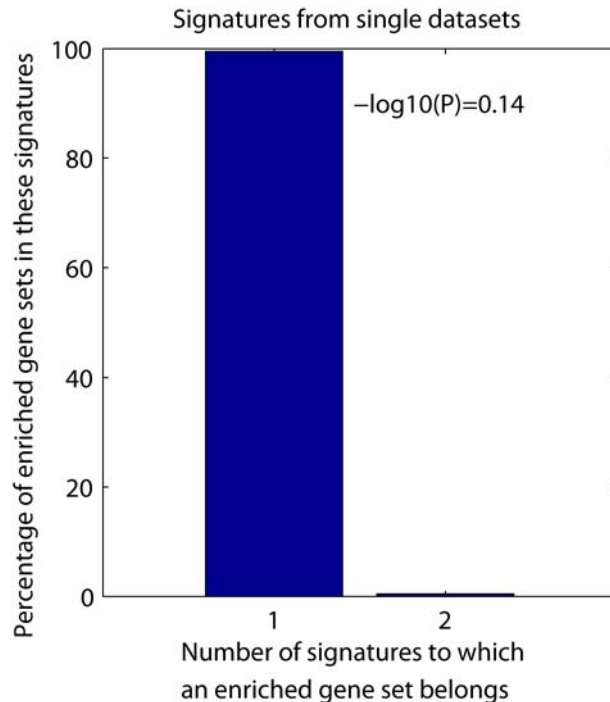


Reject Explanations 1 and 2.

Limited overlap among signatures (3)

Five explanations:

3. Dissimilar genes, the same pathways;



Also limited overlap, reject Explanation 3
(When is a process represented in a signature?)

Limited overlap among signatures (4)

Five explanations:

4. Clinical composition (i.e. sample heterogeneity)

Evaluation:

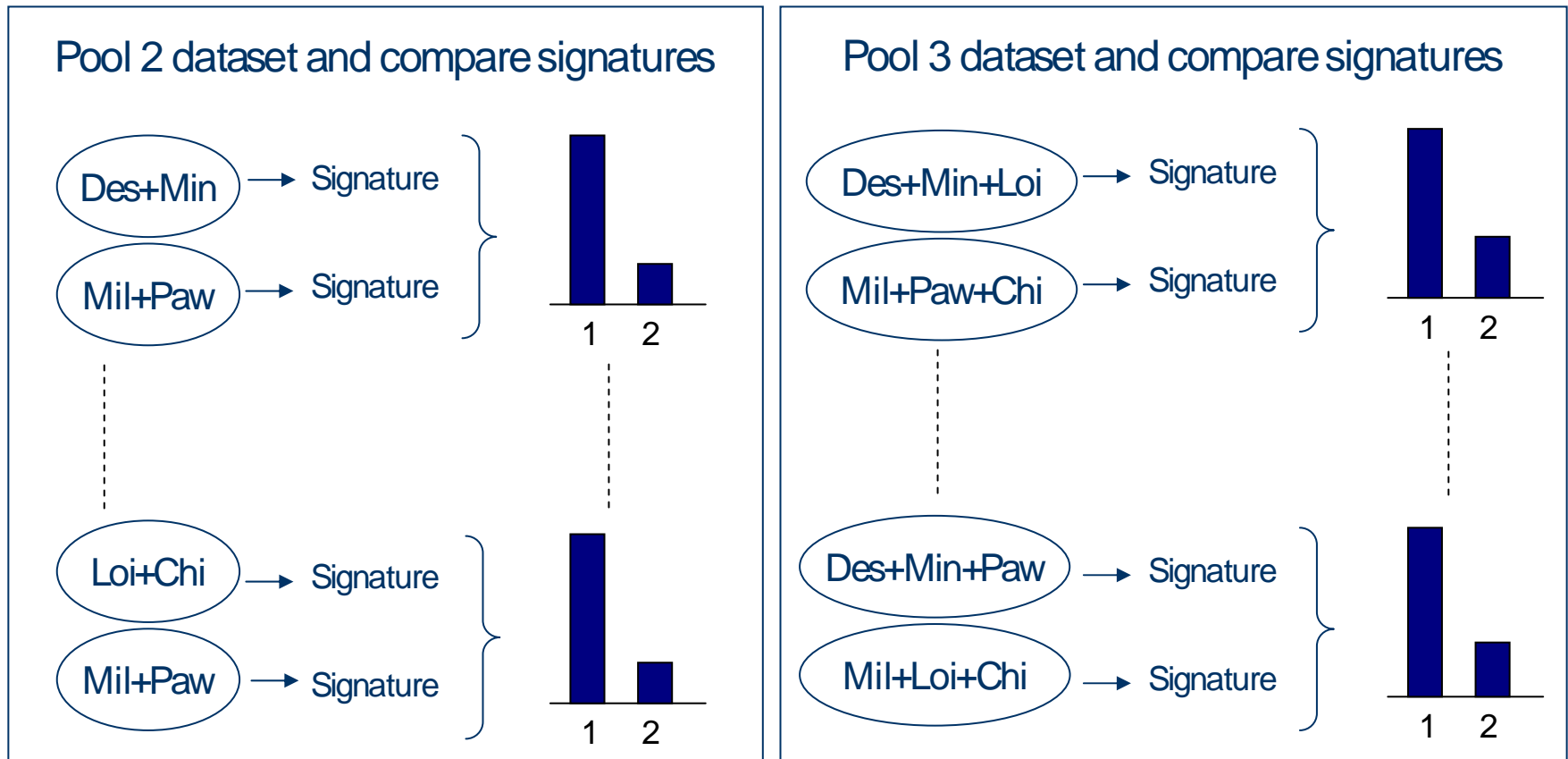
1. Repeat analysis on ER+/ER- subgroups
2. The same trends/limited overlap

Reject Explanation 4

Limited overlap among signatures (5)

Five explanations:

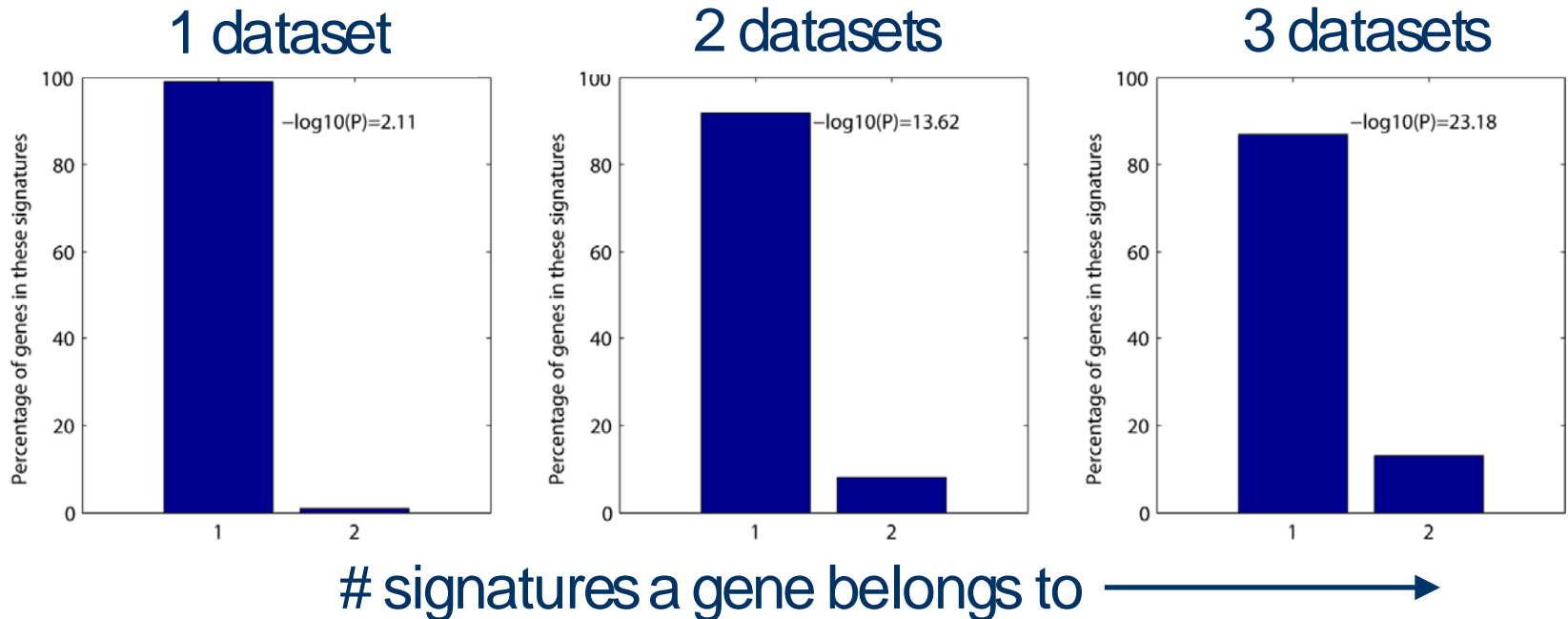
5. Small sample size problems



Limited overlap among signatures (6)

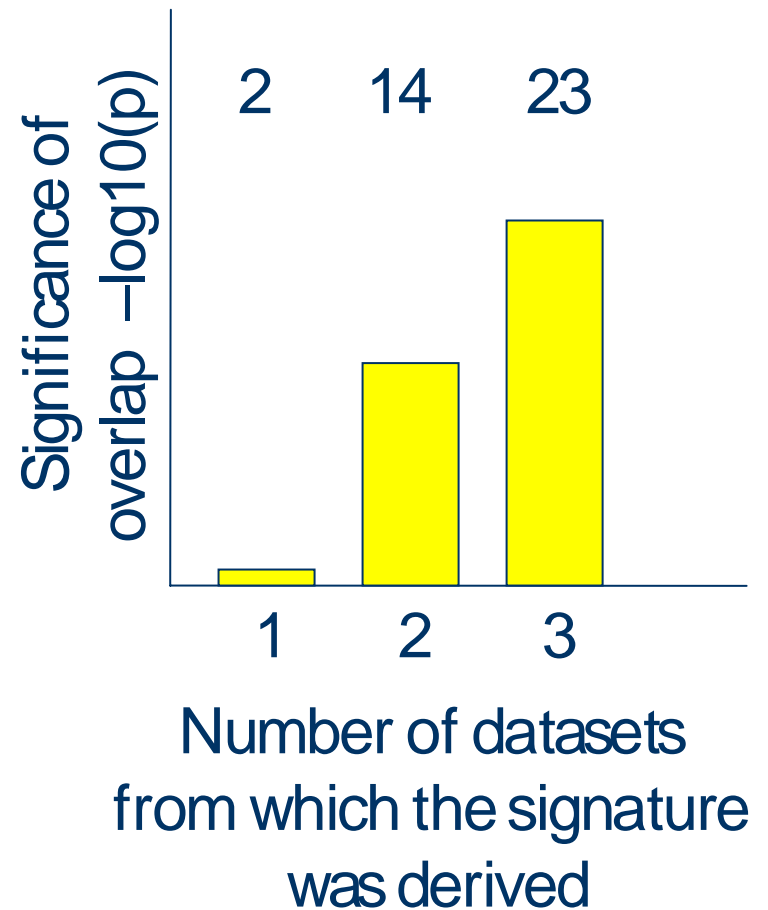
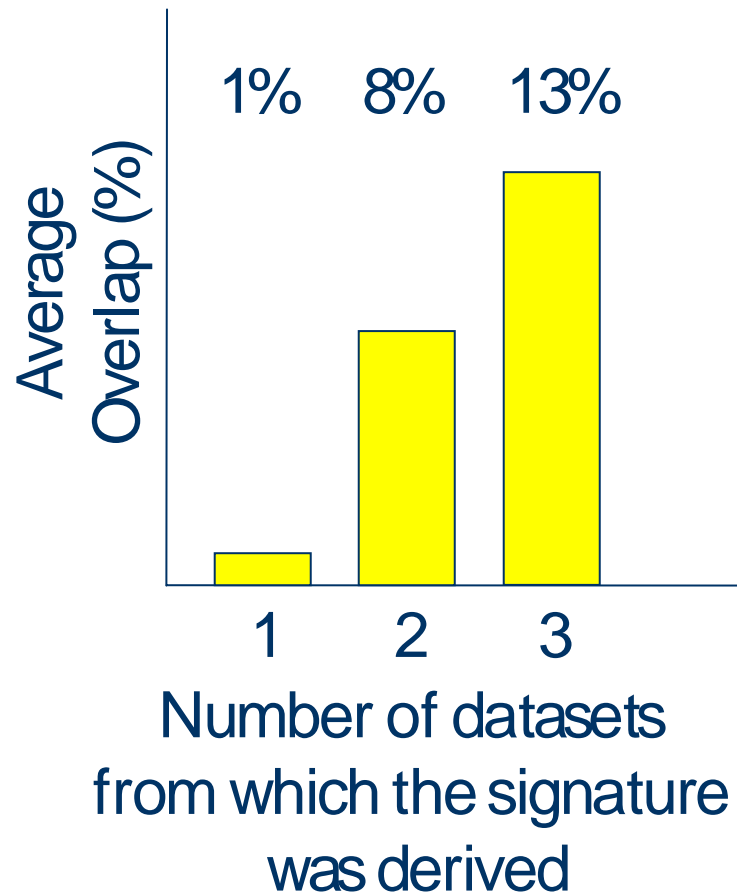
Five explanations:

5. Small sample size problems



Explanation 5 most plausible

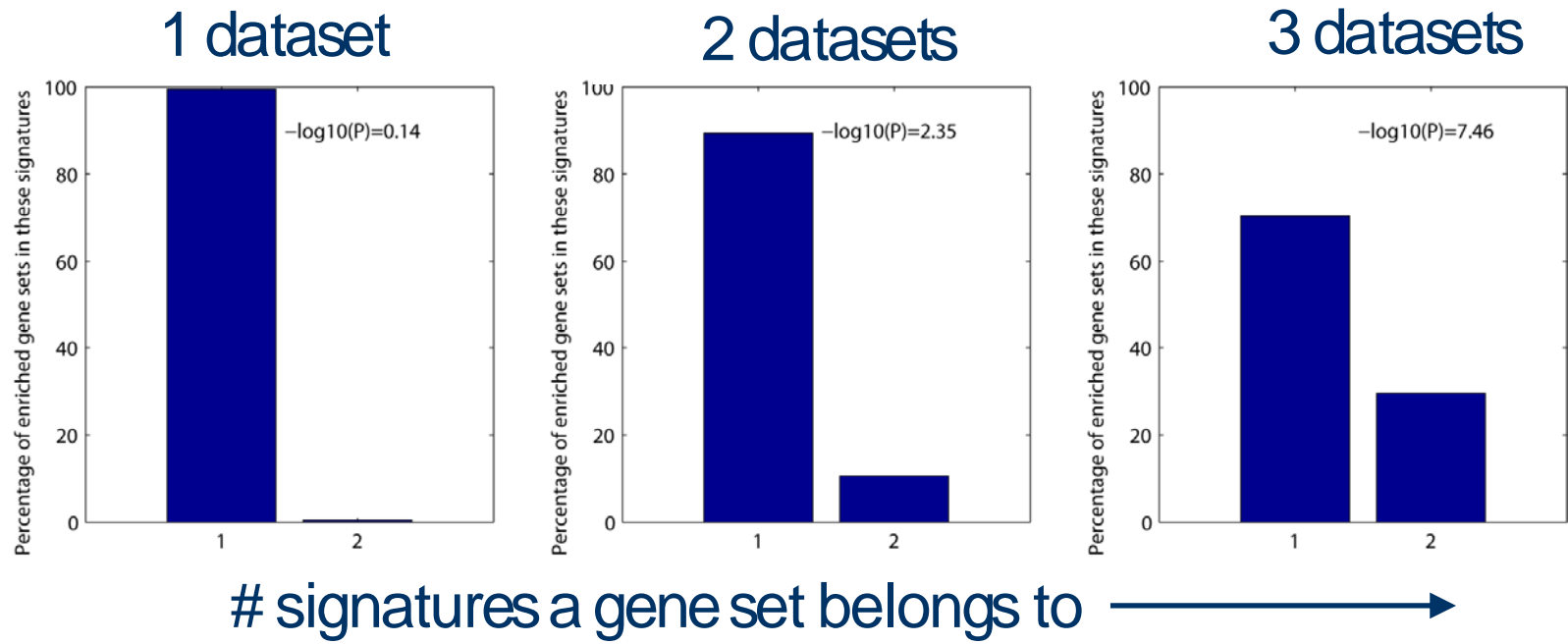
Limited overlap among signatures (3)



Limited overlap among signatures (7)

Five explanation:

5. Small sample size problems

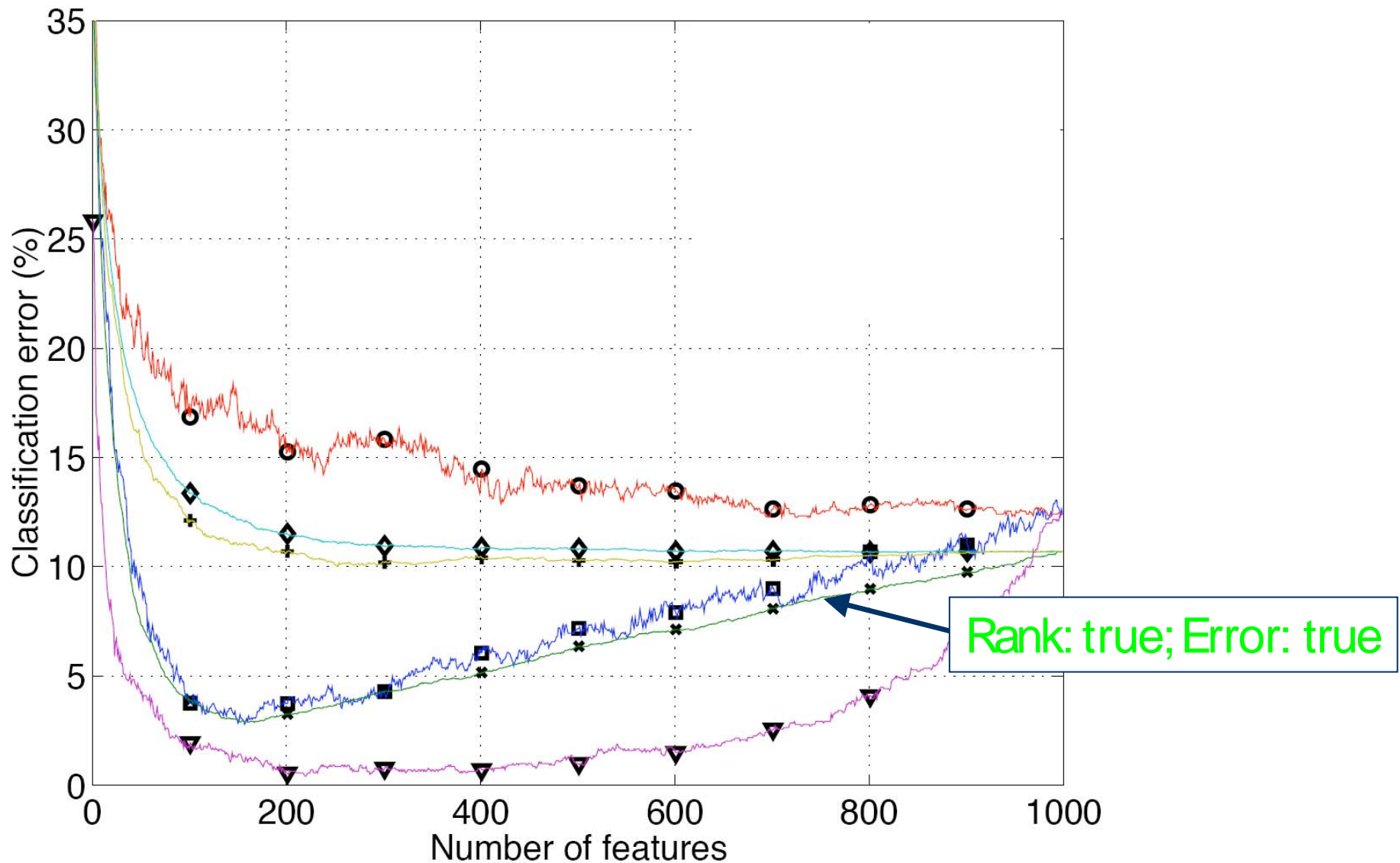


Explanation 5 most plausible (heterogeneity)

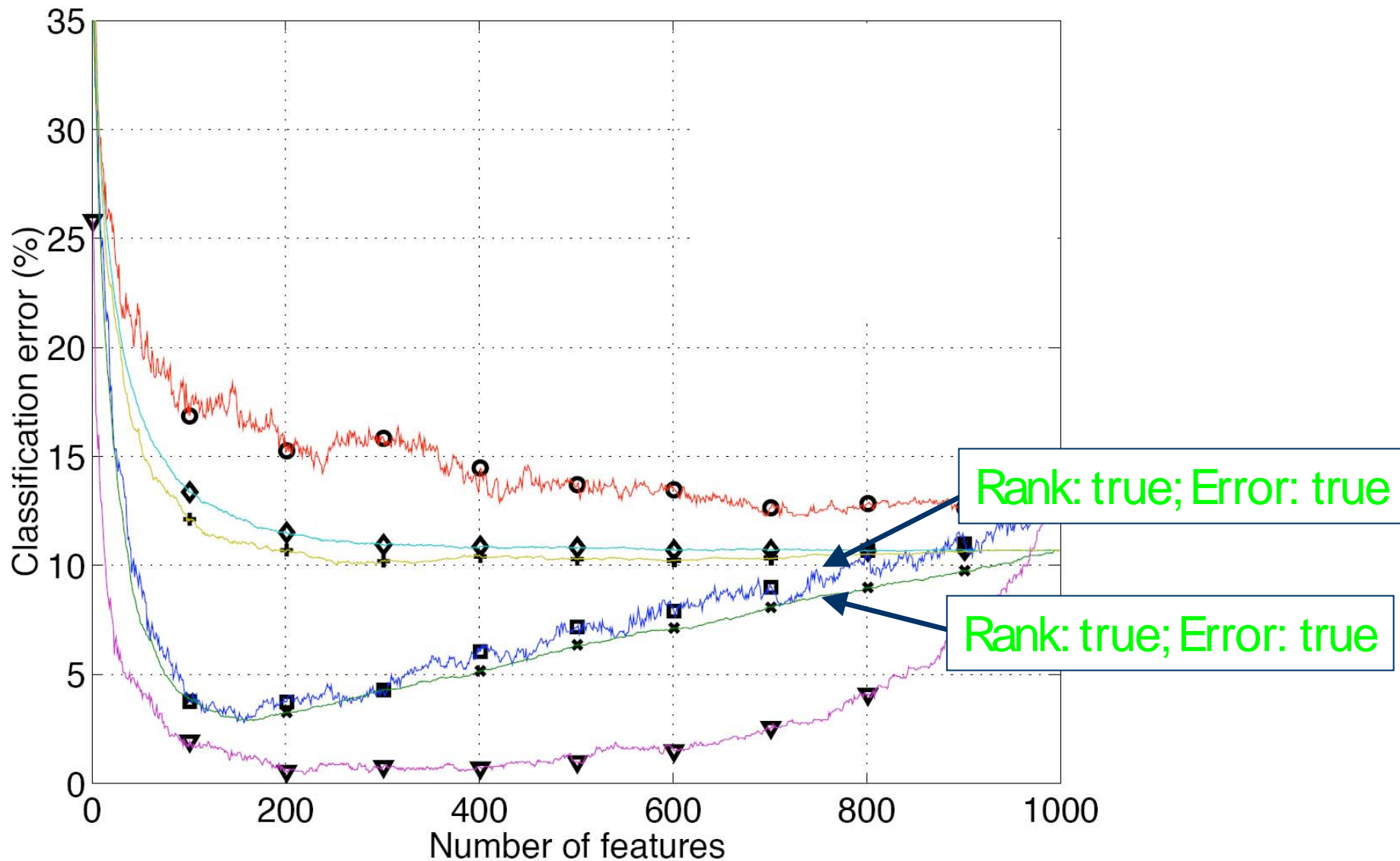
Conclusions

- Repeated Random Resampling
 - Leads to inferior ranking of genes
- Pooling datasets
 - Pooling leads to higher accuracy and a convergence of signature genes
 - Nearest mean classifier remains the best choice
 - Limited signature overlap due to small sample size
 - Note1: heterogeneity
 - Note 2: same processes?
 - To extract signatures, datasets should be pooled, rather than analyzed in isolation

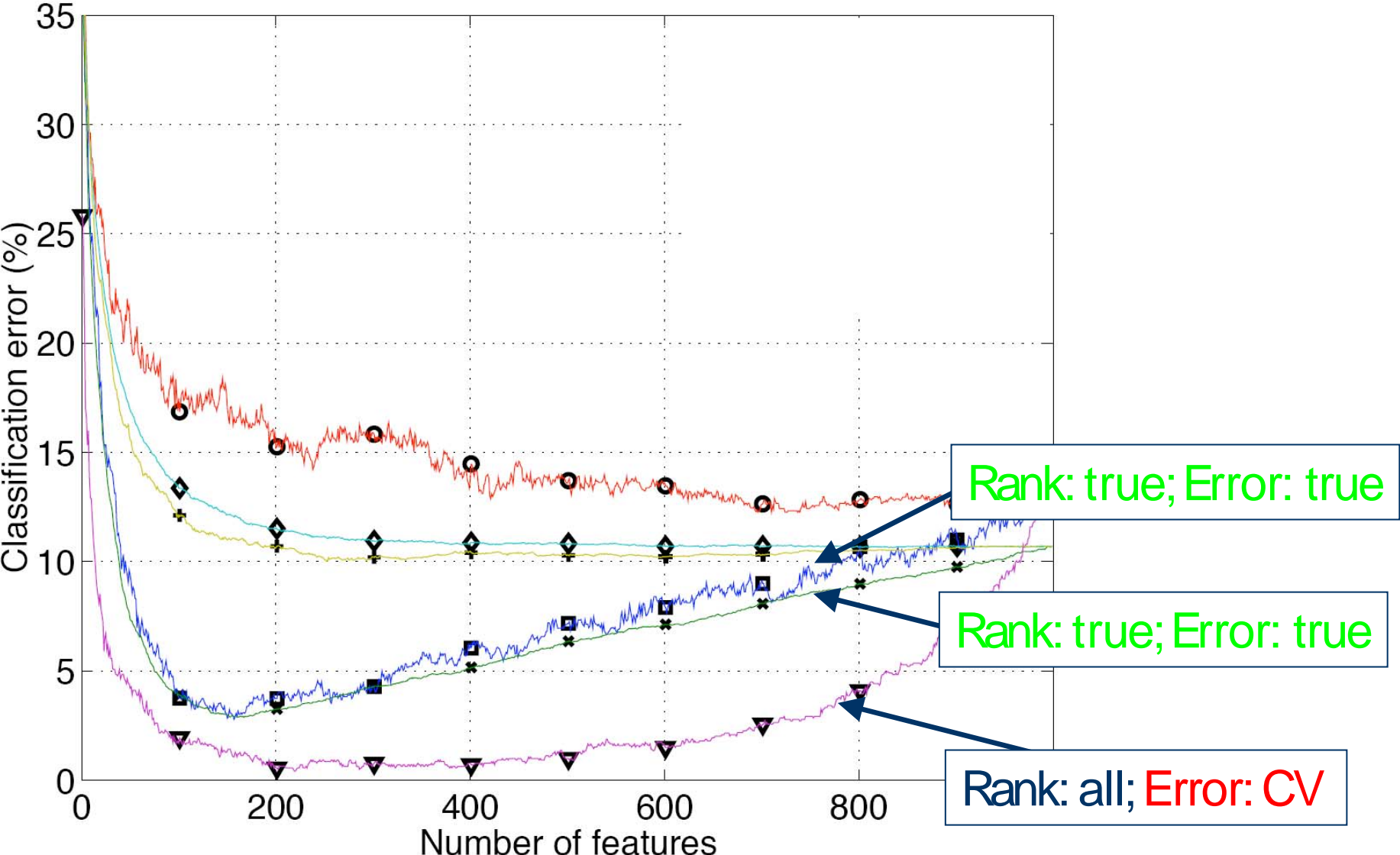
Reflection



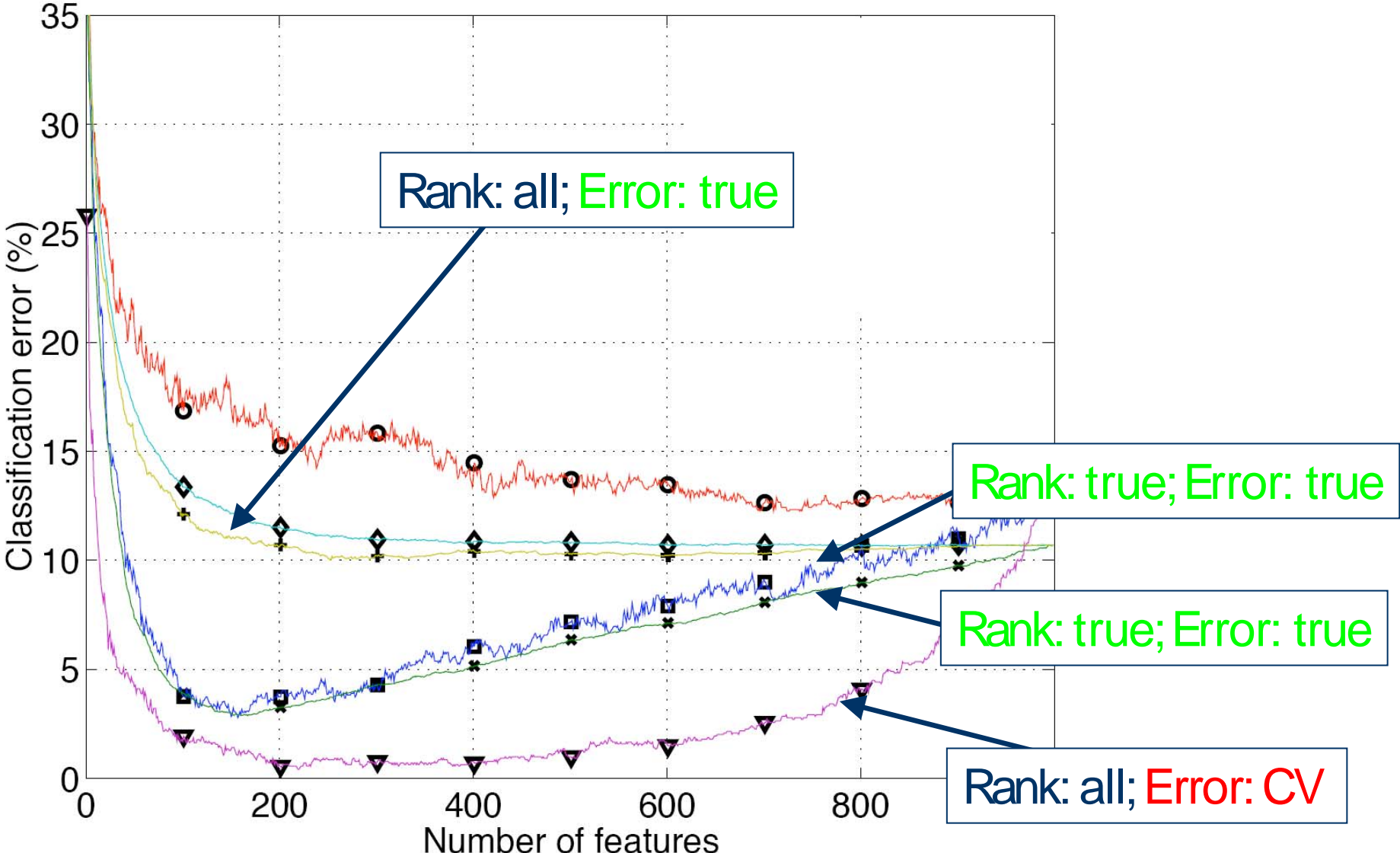
Reflection



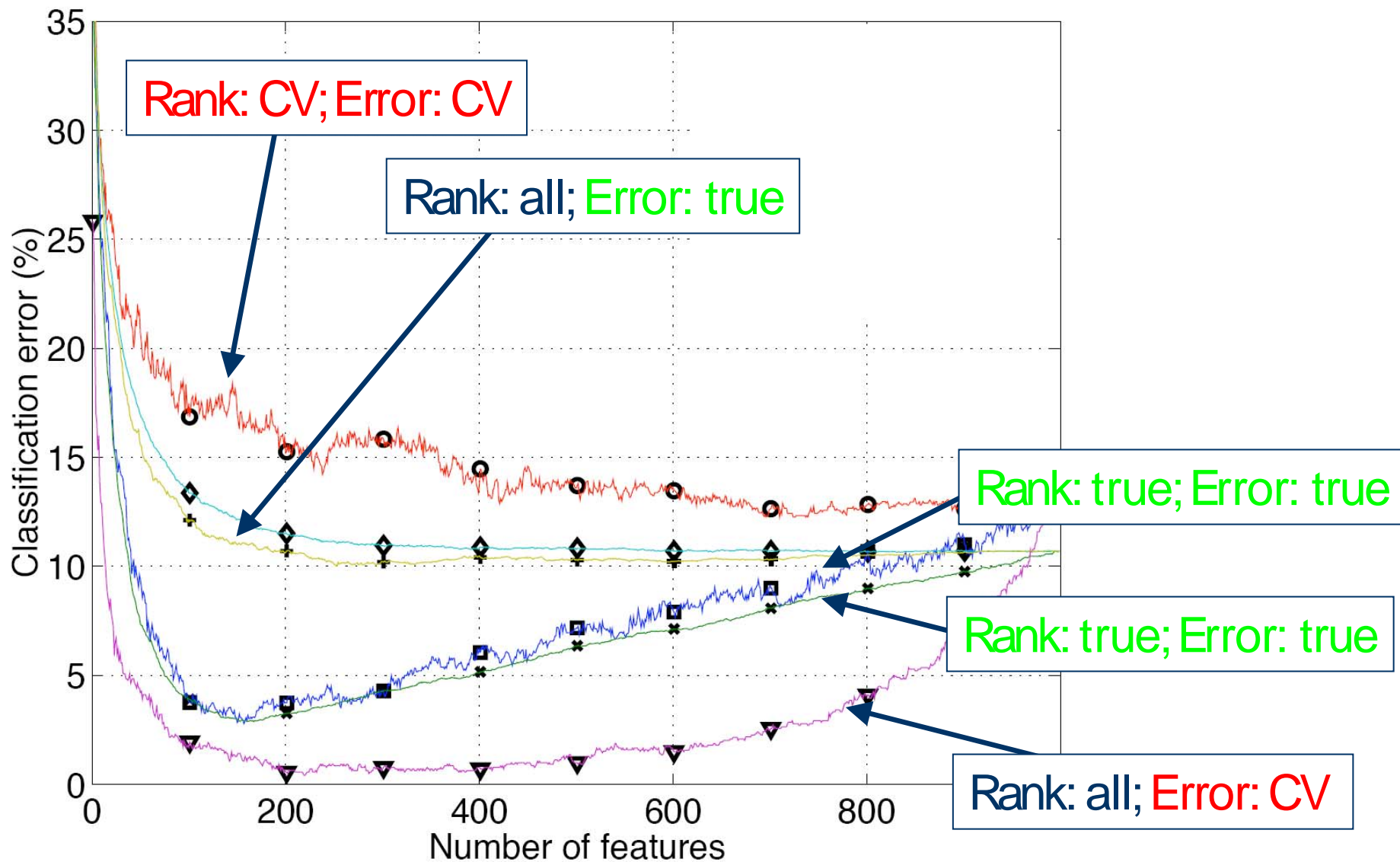
Reflection



Reflection



Reflection



Alternative analysis

- Given large collection of BC samples
- Many BC gene expression signatures (outcome)
- Set out to:
 - Compare existing signatures
 - Derive new (functional) insights from the large collection of data

BC prognostic gene signatures

- Seven studies that published signatures
- Nine signatures (two studies with two variants)
- Signatures applied as described in studies
- All signatures are claimed to be prognostic

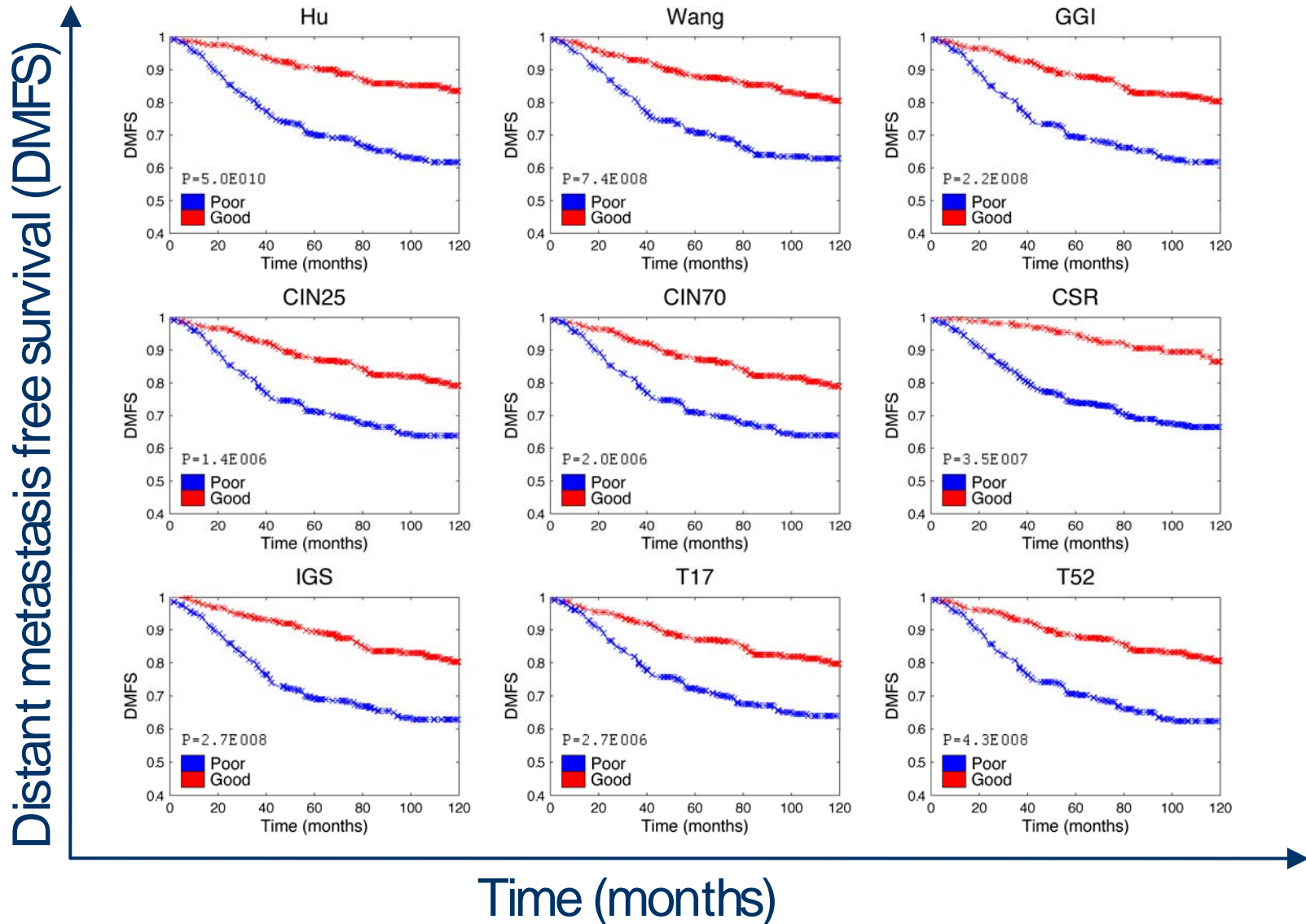
Study	Label	Detail
Wang <i>et al.</i>	Wang	Untreated, LN-
Chang <i>et al.</i>	CSR	Core serum response
Carter <i>et al.</i>	CIN25/70	Chromosomal Instability
Teschendorff <i>et al.</i>	T17/52	ER+
Hu <i>et al.</i>	HU	Intrinsic genes
Liu <i>et al.</i>	IGS	Invasiveness signature
Sotiriou <i>et al.</i>	GGI	Genomic grade index

Seven breast cancer datasets

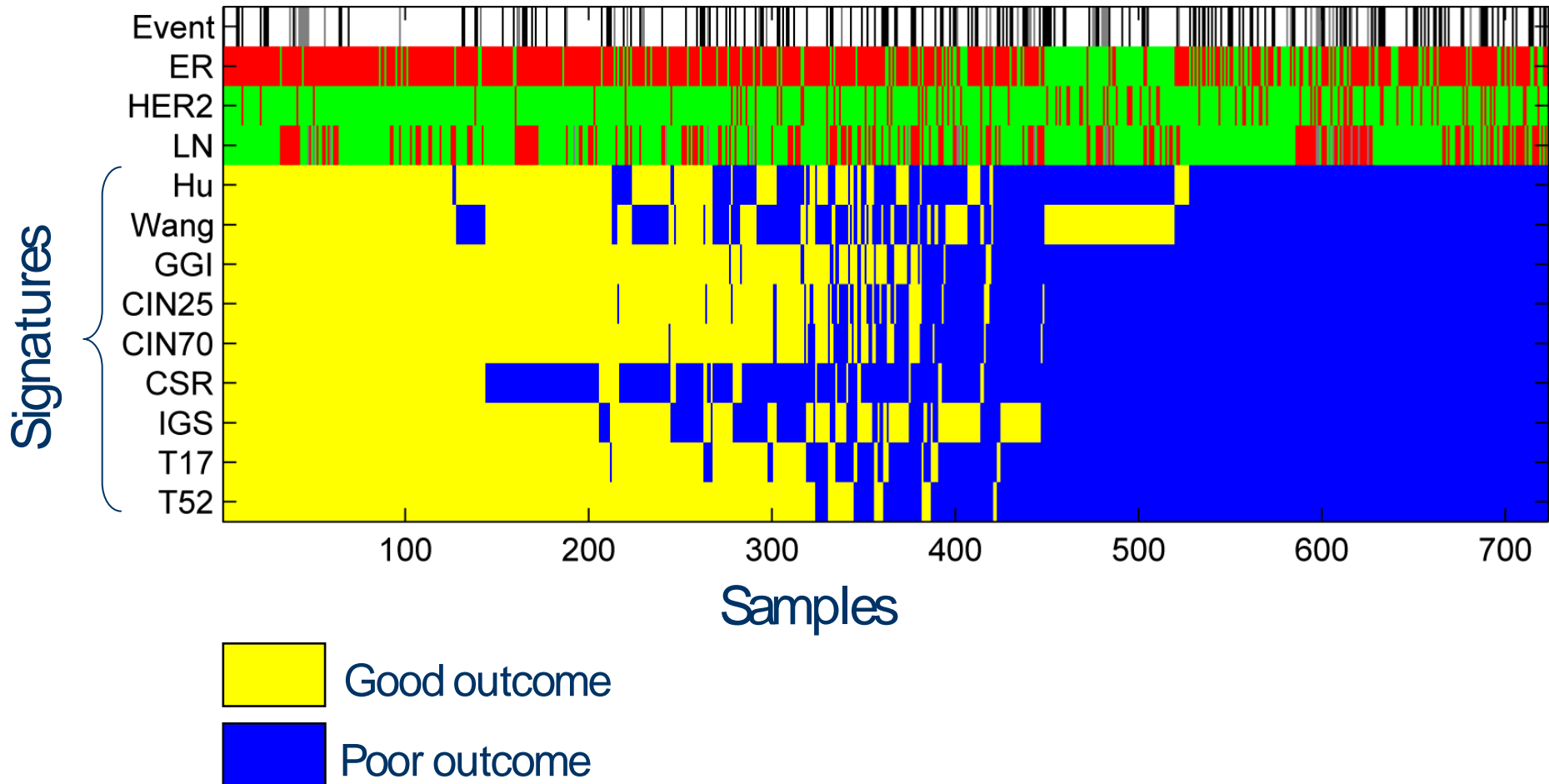
- Six datasets (all Affymetrix HG U133A)
- Same pre-processing applied from raw data
- Seventh dataset used for validation (NKI 295, Vijver et al.)
- Outcome: Distant Metastasis Free Survival (**DMFS**)

Study	Label	Total
<i>Desmedt et al.</i>	Des	147
<i>Minn et al.</i>	Min	96
<i>Miller et al.</i>	Mil	247
<i>Pawitan et al.</i>	Paw	156
<i>Loi et al.</i>	Loi	178
<i>Chin et al.</i>	Chi	123
Total number of samples:		947

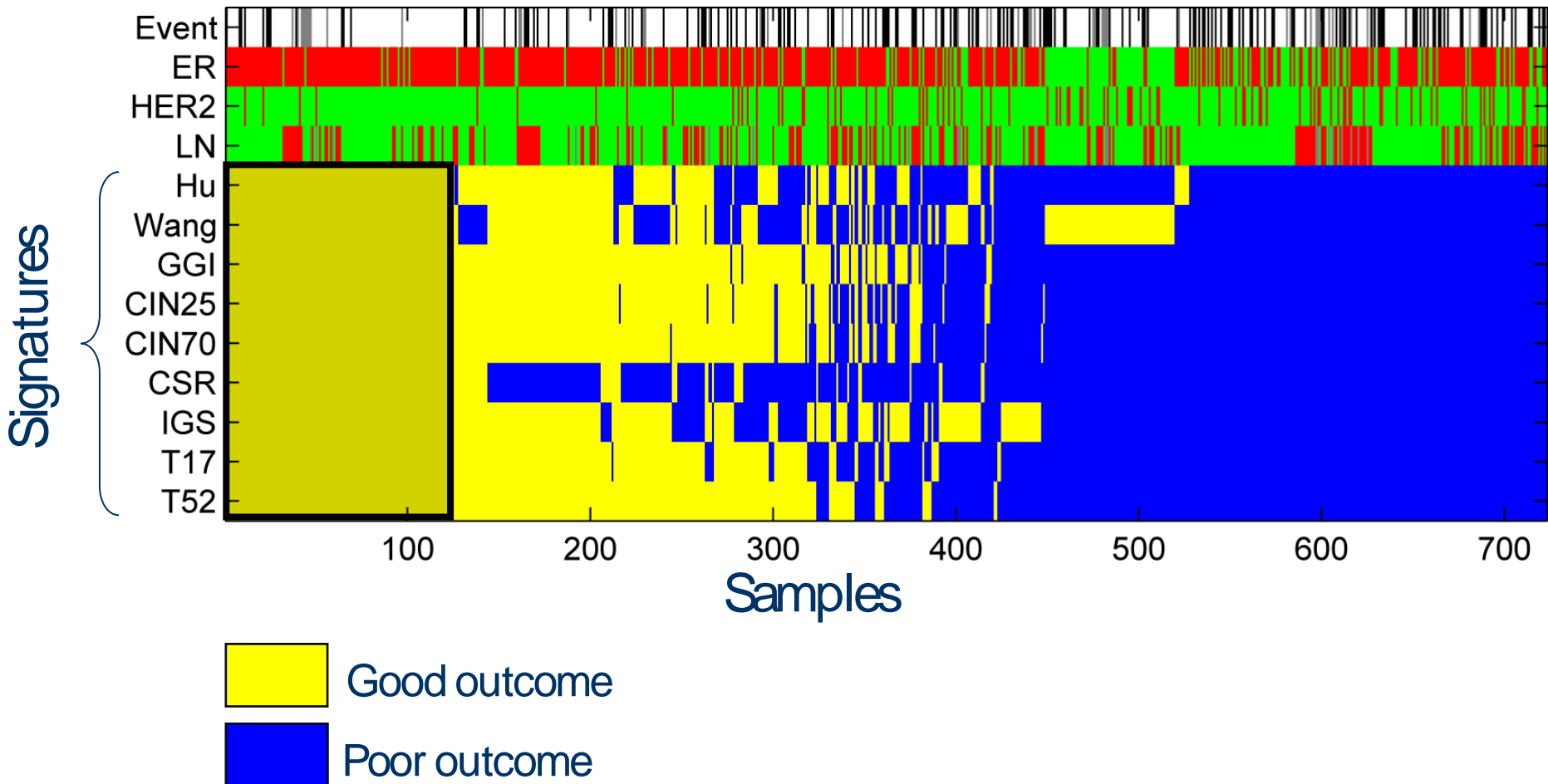
Signature performance on n = 947



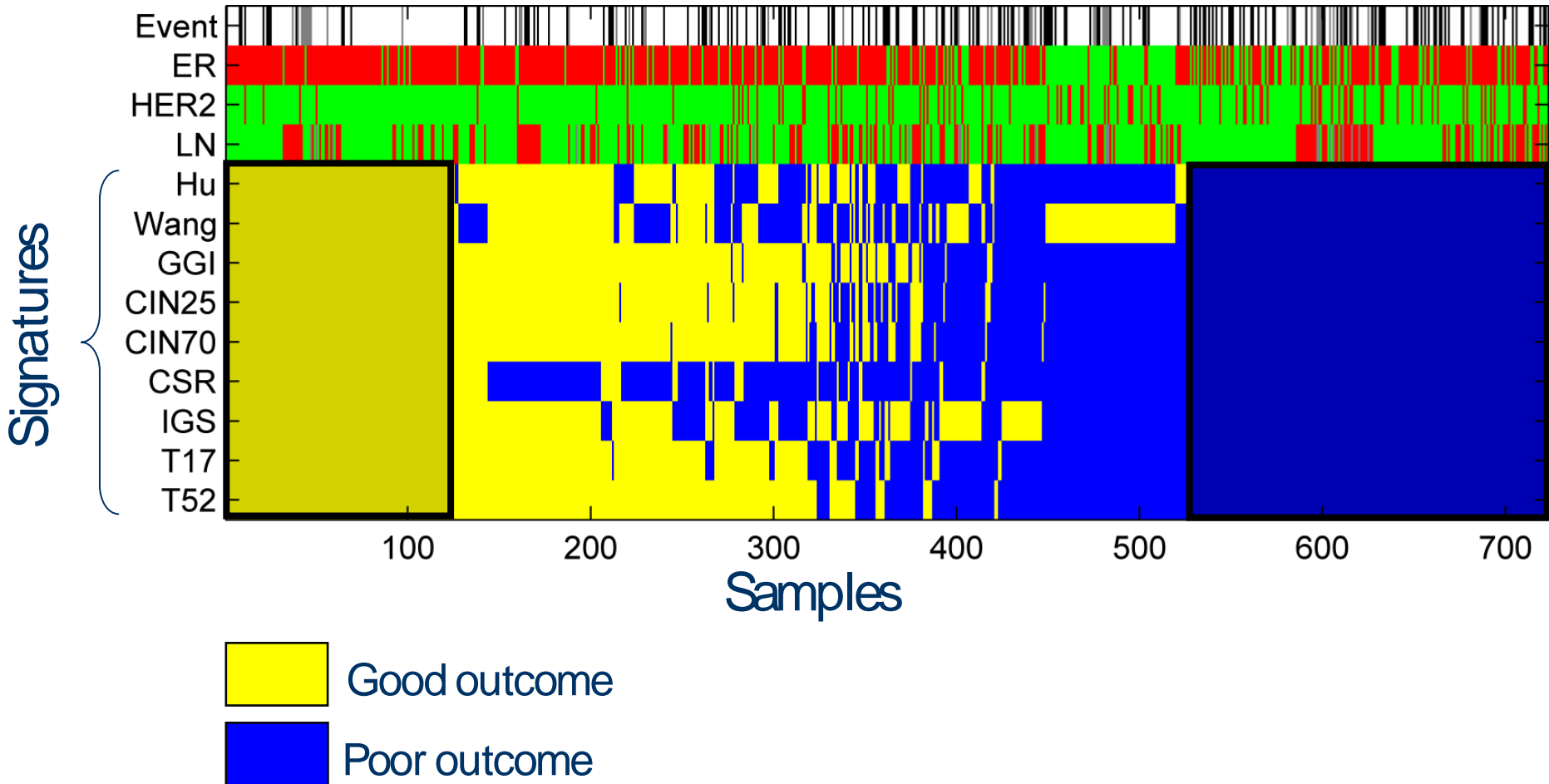
Signature concordance (1)



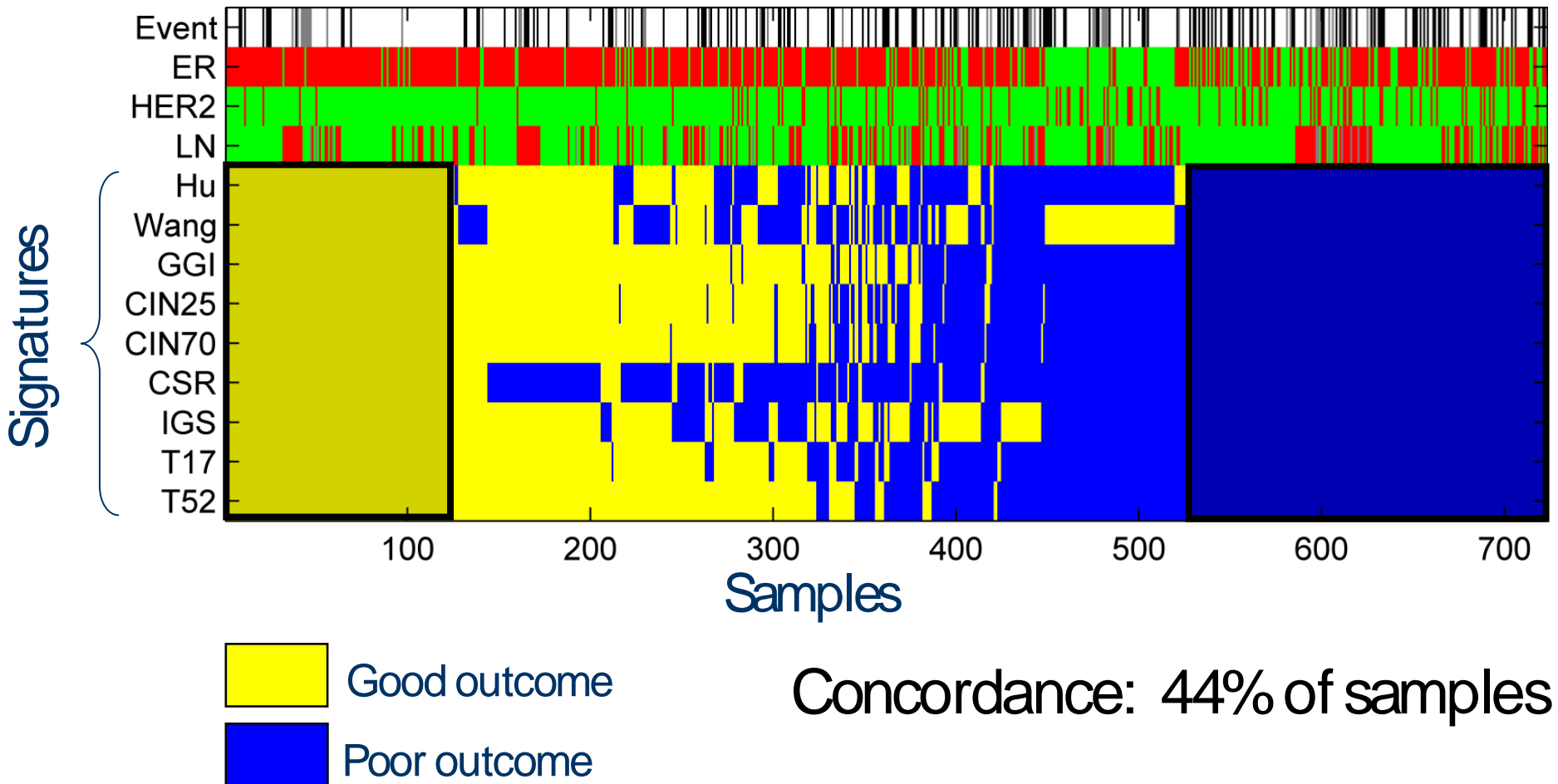
Signature concordance (2)



Signature concordance (3)

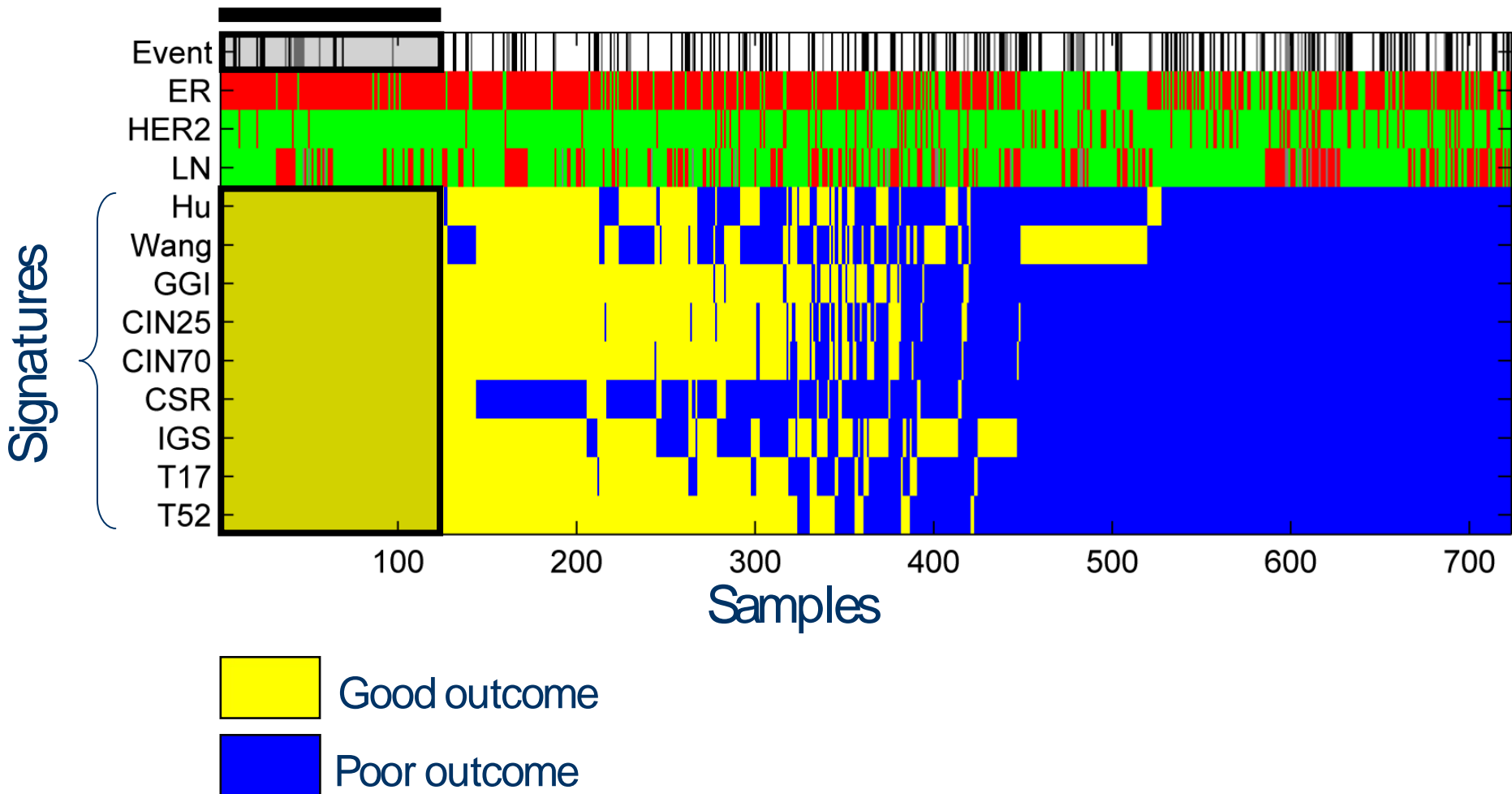


Signature concordance (4)



Signature concordance (5)

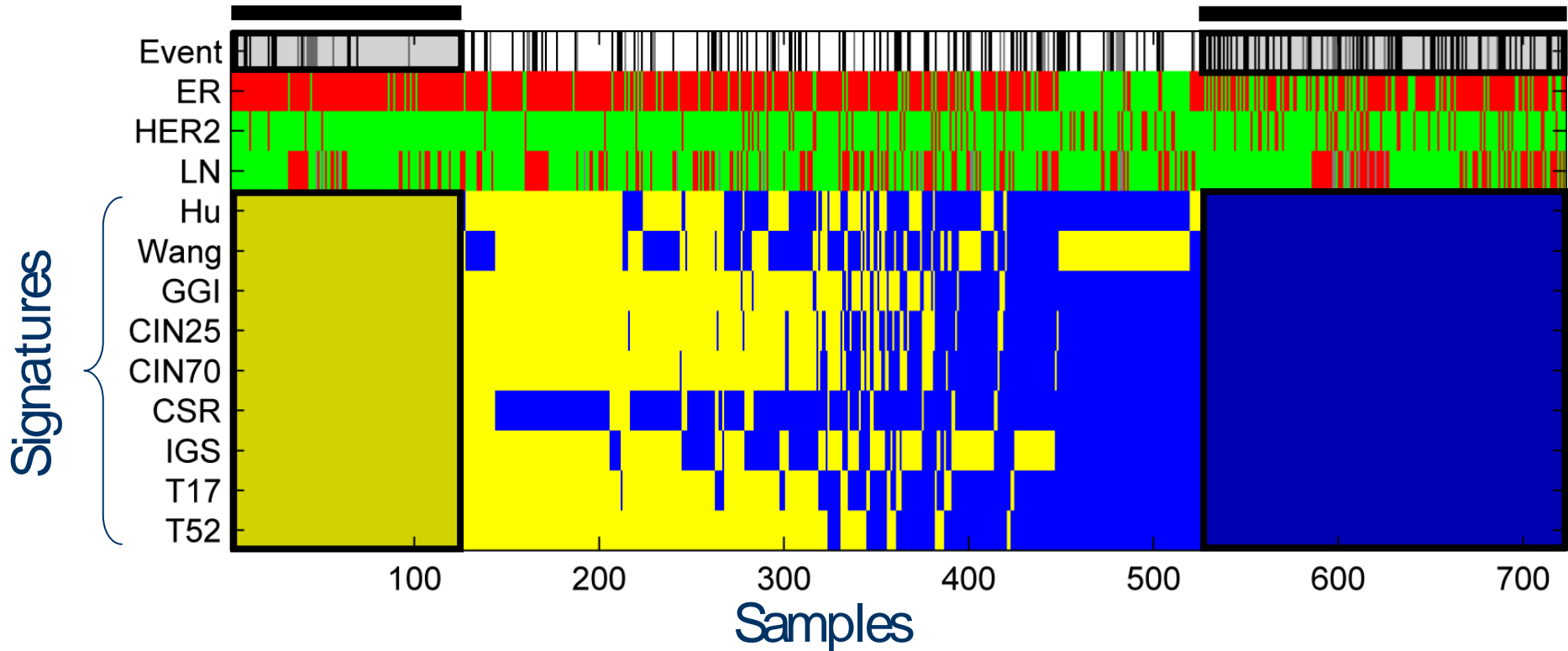
7% events



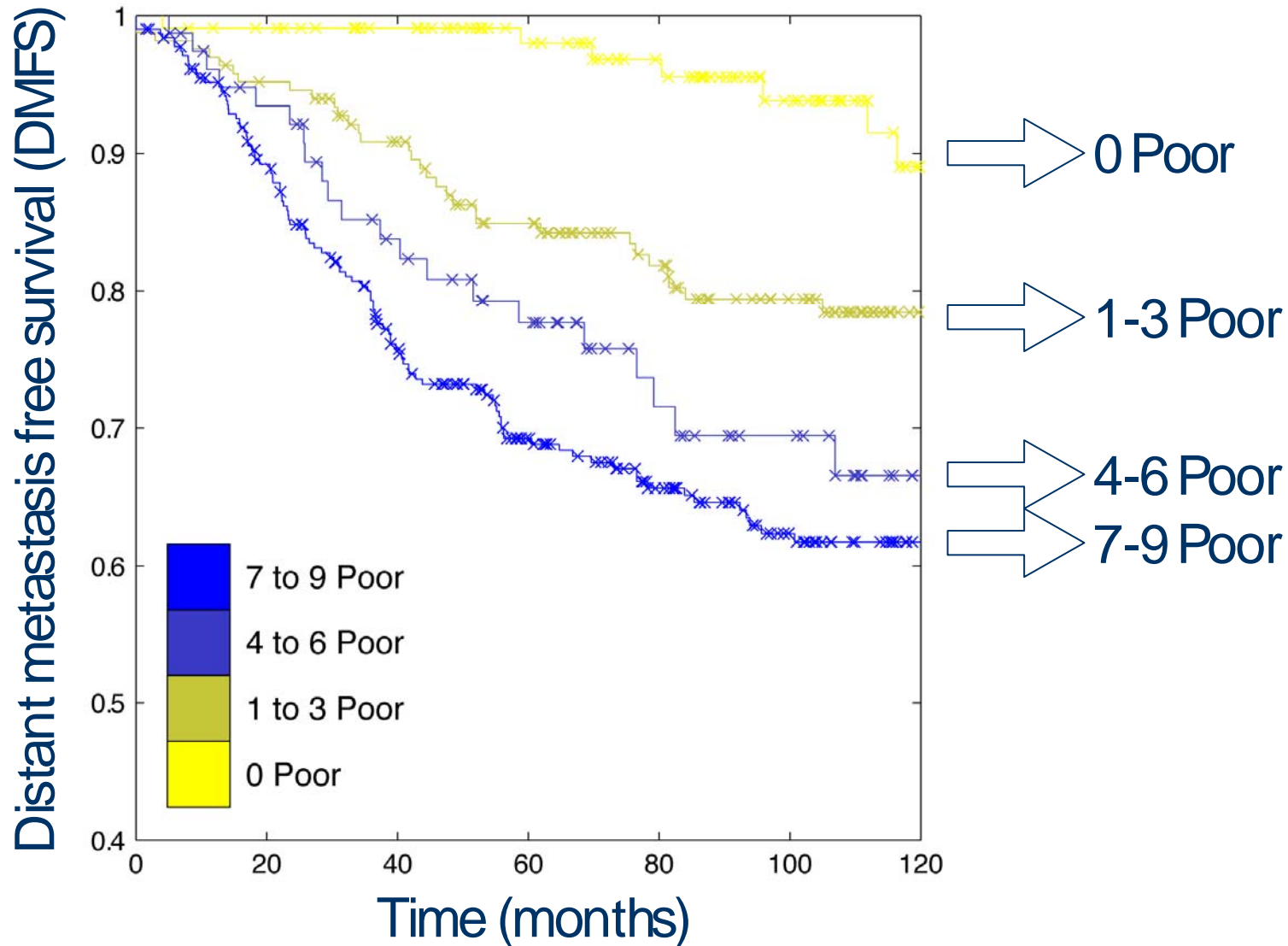
Signature concordance (6)

7% events

35% events



Classification using existing signatures (3)

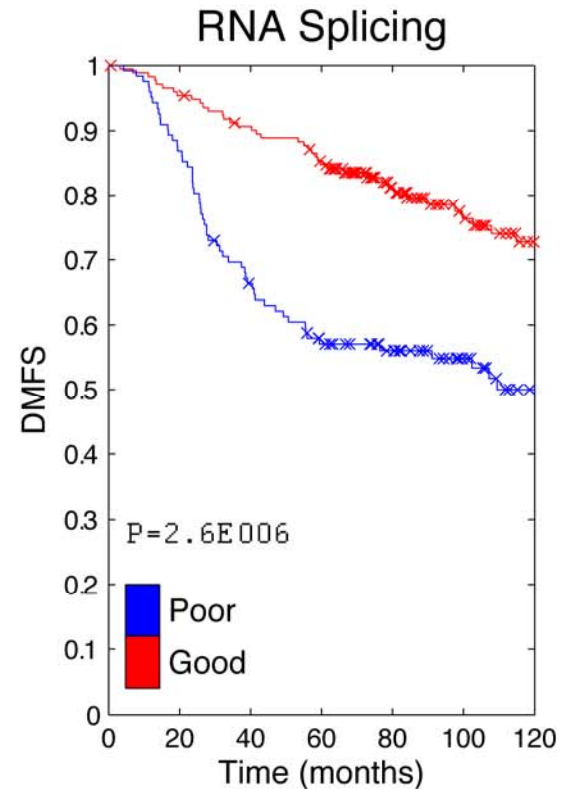
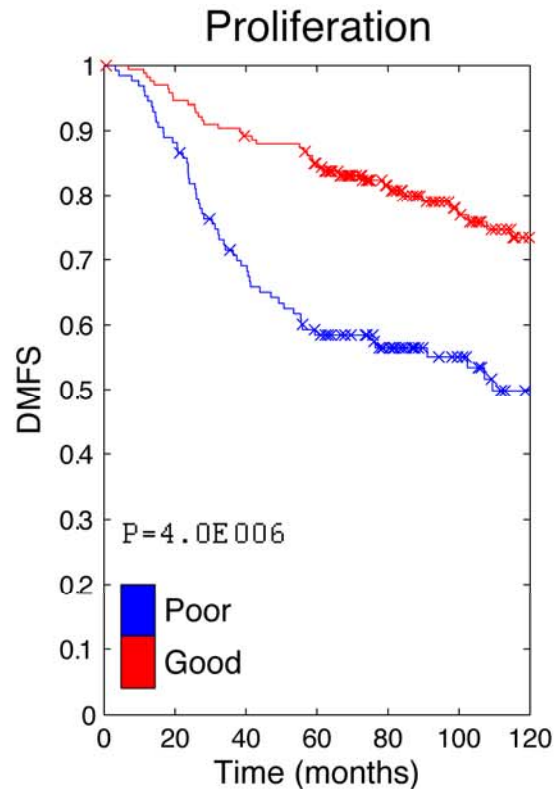
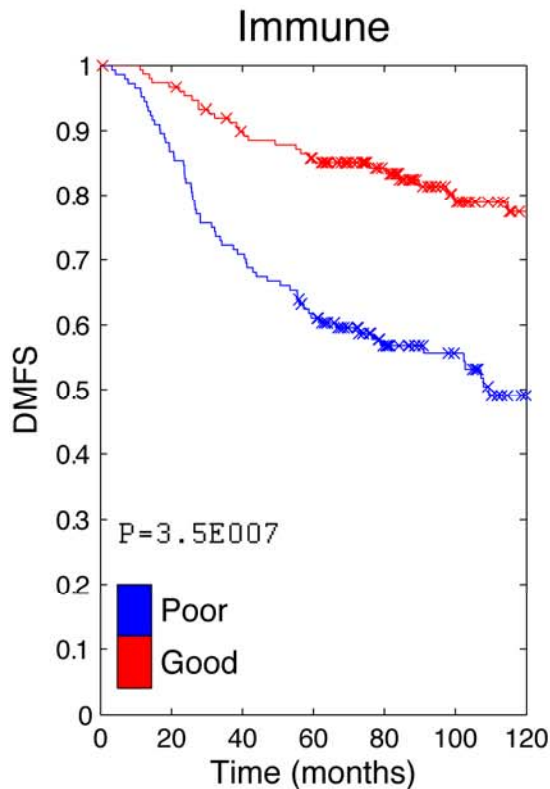


Which processes are represented?

- Limited signature overlap
 - Same processes, but different genes
- Enlarging signatures
 - Single gene from process sufficient for prediction, not to detect enrichment
 - Strengthen this signal by enlarging the signatures with correlated genes (Spearman rank correlation > 0.7)
 - Chin and Loi as training sets, rest as test set
- Enrichment analysis
 - Databases with process information (GO, KEGG etc.)
 - Compute enrichment of processes in enlarged signatures

Do all modules have prognostic power?

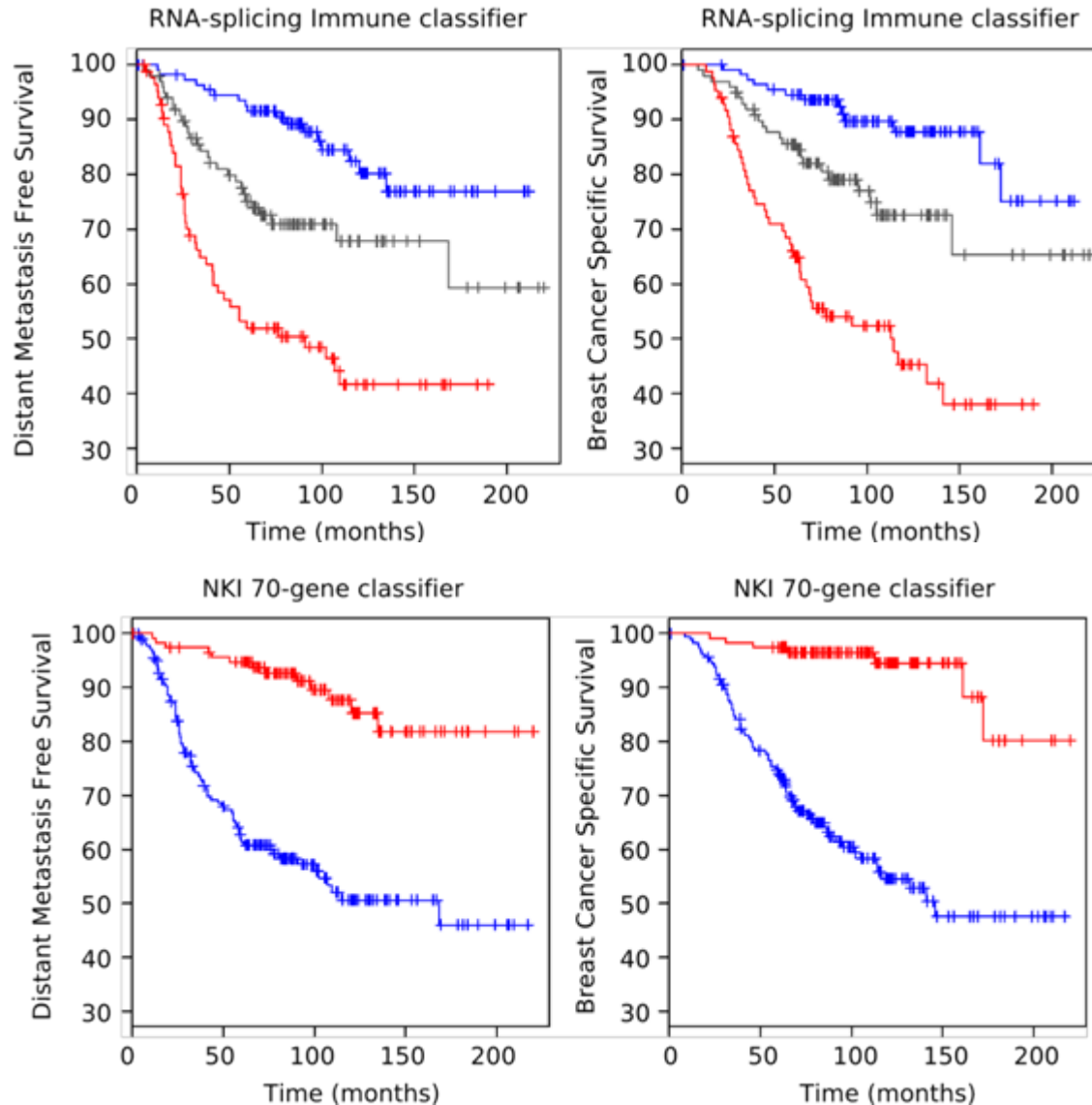
- Train a classifier using genes from modules
- Significant separation: $p=10^{-3}$ - 10^{-7} (logrank test)



Composite classifier

- Evaluated all pair-wise combinations
- 3 groups:
 - concordant good
 - concordant poor
 - Discordant
- RNA splicing - Immune response the best

RNA splicing – Immune classifier



Conclusions (1)

- Signature performance is similar
- 44% concordance amongst signatures
- Identified 10 different functional modules
- Individual modules all have prognostic power
- Proliferation is a strong common denominator
- Immune, Proliferation and RNA splicing show best performance

Acknowledgements



Martin van Vliet

Fabien Reyat

Hugo Horlings

Marc van de Vijver

Marcel Reinders

<http://bioinformatics.nki.nl>

Enrichment analysis (1)

- Databases:
 - Reactome database
 - KEGG
 - Molecular Signatures Database (Broad)
 - Gene ontology database
- Only sets with at least 5 probe sets: 1889 sets
- Benjamini-Hochberg correction