Identification of Breast Cancer Molecular Subtypes

Benjamin Haibe-Kains[1,2]

[1]Functional Genomics Unit, Institut Jules Bordet, Brussels, Belgium
[2]Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium

1. Background

SINCE the advent of array-based technology and the sequencing of the human genome, scientists attempted to bring new insights into breast cancer biology and prognosis. The identification of natural groups of tumors (called subtypes) from gene expression data was the subject of intense research this last decade. Such an identification usually involved two steps:

1. Identification of sets of co-expressed genes (gene clustering).
2. Identification of groups of similar tumors (sample clustering).

Perou and Sorlie et al. were the first to conduct a gene expression profiling study of breast tumors in order to highlight the existence of subtypes, i.e. groups of tumors exhibiting similar "genetic portraits". Their method consisted in (i) identifying the genes with low variance between tumor samples from the same patient but high variance between tumor samples from different patients (referred to as *intrinsic gene list* in the literature); (ii) performing a two-way hierarchical clustering to identify sets of co-expressed genes (Figure 1) and groups of similar tumors (Figure 2). In addition to highlight the importance of ER and HER2 phenotypes and proliferation, they also shown that the molecular subtypes exhibited different clinical outcome as illustrated by the survival curves in Figure 3.

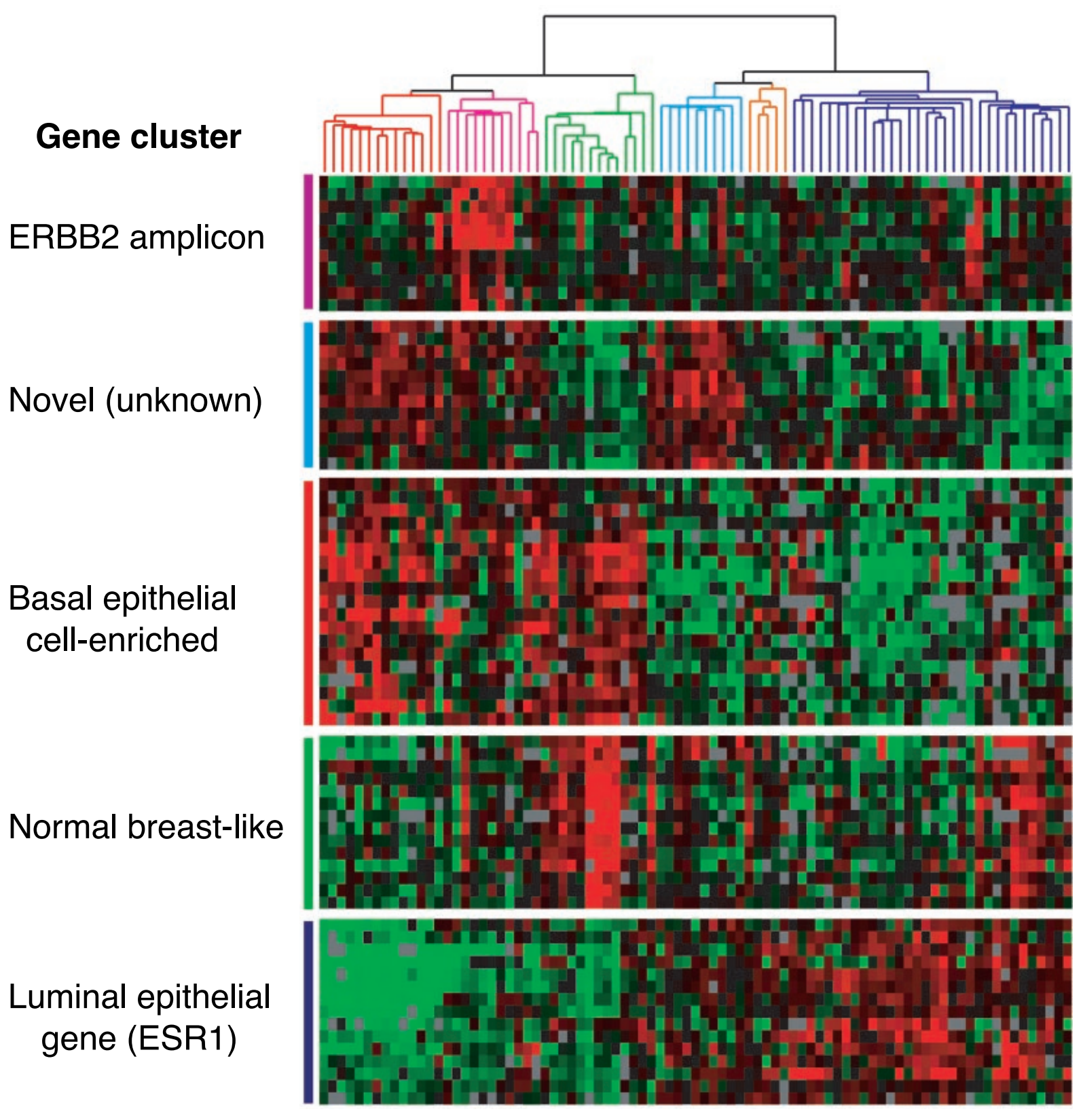

Figure 1: *Gene clustering [5].*

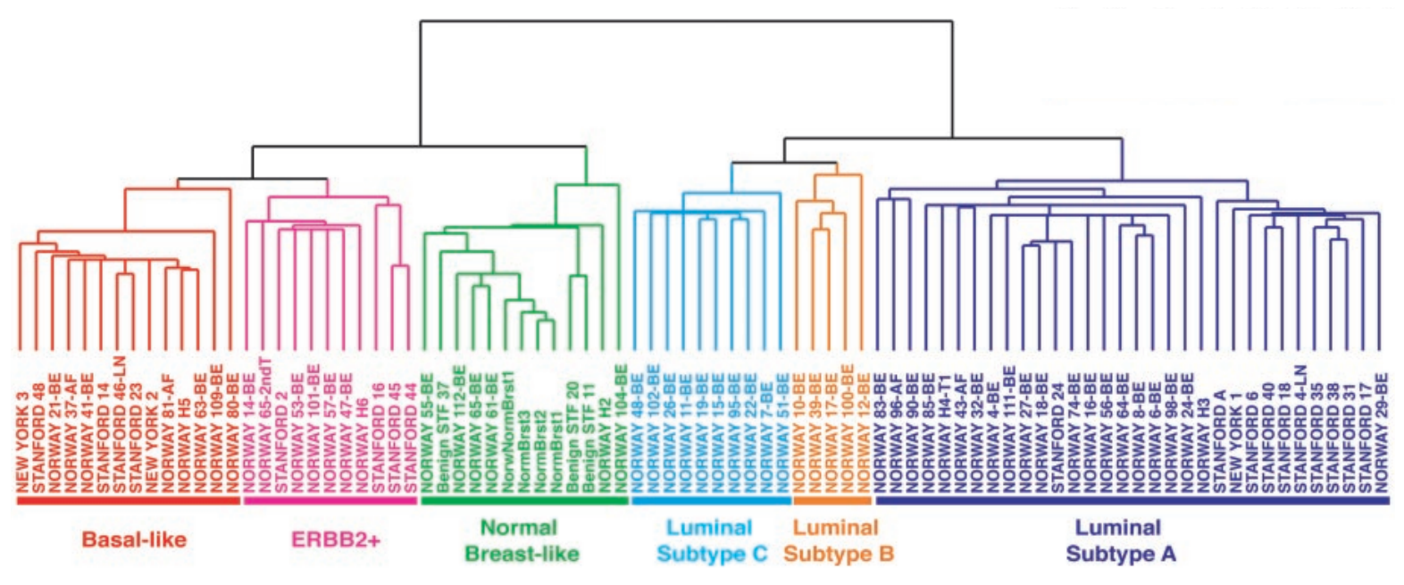

Figure 2: *Sample clustering [5].*

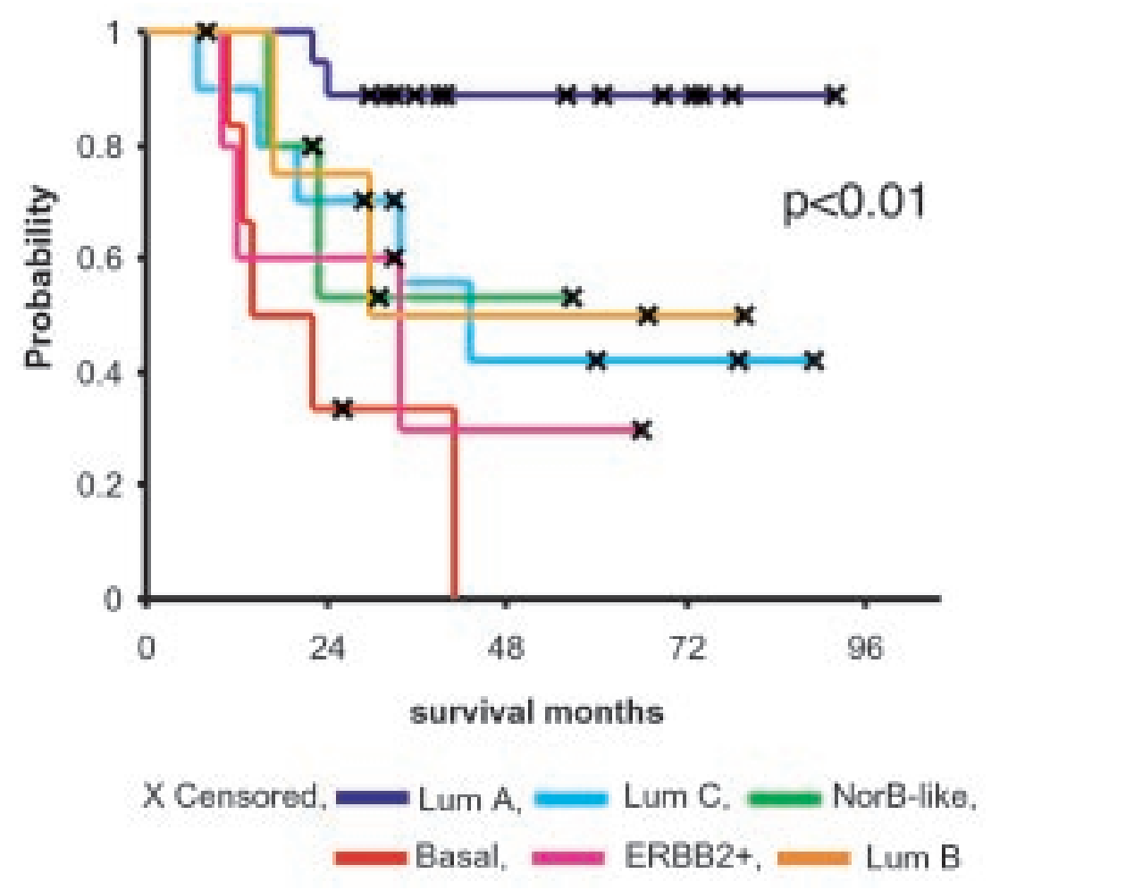

Figure 3: *Survival curves for each subtype [5].*

Although these results brought new insights into breast cancer biology, the method suffers from serious drawbacks [4]:

- The dendrogram was cut subjectively to identify the subtypes, making difficult the implementation of an automatic tool.
- The hierarchical clustering used in combination with a large number of genes is unstable due to the curse of dimensionality, making difficult the reproducibility of the results.
- The model fitted by hierarchical clustering does not allow for an easy application to new data, making difficult the validation of the method and the classification of a new patient. To circumvent this difficulty, the authors developed a nearest centroid classifier, called the *single sample predictor* (SSP).
- The model fitted from hierarchical clustering or the SSP lead to a crisp partition of the dataset with no accurate estimation of the classification uncertainty.

In order to address these issues, we sought to develop a novel method for identifying the molecular subtypes in breast cancer. This method, in addition to exhibit several advantages compared to the hierarchical clustering used in the initial publications, yielded robust classification in several independent microarray datasets.

2. Materials and Methods

WE recently introduced a novel clustering model to robustly identify the breast cancer molecular subtypes. This model consists in: (i) identifying gene modules, i.e. sets of genes that are specifically co-expressed with genes of interest; and (ii) identifying molecular subtypes using a simple model-based clustering in a low dimensional space defined by these gene modules.

2.1 Gene Modules

The aim of gene modules is the identification of co-expressed genes related to a biological process of interest. Contrary to the method of Perou and Sorlie et al., we used *a priori* biological knowledge to find clusters of co-expressed genes.
The method includes the following steps:

1. Choice of the biological processes of interest (ER, HER2 phenotypes, proliferation, . . .).
2. Selection of a *prototype* for each biological process.
   - A prototype is a gene known to be related to the biological process of interest (e.g. ESR1 for ER phenotype or AURKA for proliferation).
3. Identification of the genes specifically co-expressed with each prototype to populate gene modules.
   - A gene j is specifically co-expressed with a prototype q if the co-expression of gene j with prototype q is **statistically** higher than with the other prototypes.
4. Finally, a summary of a gene module, called a *gene module score*, is computed by averaging the expressions of the genes in the module.

2.2 Model-Based Clustering

We know from early microarray studies that breast cancer is a molecularly heterogeneous disease. ER and HER2 phenotypes being to be the main discriminators. Moreover, Kapp et al. showed that robust clusters were only identified using pairs of genes related to ER and HER2 phenotypes [3].

In order to robustly identified the breast cancer subtypes, we introduced a simple model-based clustering (mixture of Gaussians) in a two-dimensional space defined by the ESR1 and ERBB2 module scores. This model allows for estimating for the tumors, the probability to belong to each subtype. We used the Bayesian information criterion to select the most likely number of subtypes [2].

We retrieved 15 public microarray datasets of breast cancer patients to validate our model by estimating the prediction strength [6].

3. Results

FROM two large microarray datasets (≈ 600 patients, VDX & NKI), seven gene modules were built in order to represent key biological processes in breast cancer : ER phnotype (ESR1), HER2 phenotype (ERBB2), proliferation (AURKA), immune response (STAT1), angiogenesis (VEGF), tumor invasion (PLAU) and apoptosis (CASP3). We found gene modules of various size. As expected, the largest gene modules are related to ER phenotype and proliferation. A gene ontology analysis confirmed the coherence of the gene modules with respect to the prototypes or biological processes of interest.

Gene module	Size
ESR1	468
AURKA	228
STAT1	94
PLAU	67
ERBB2	27
VEGF	13
CASP3	8

We used the ESR1 and ERBB2 module scores as input space for the model-based clustering to identify the breast cancer subtypes. We fitted our model on VDX (Figure 4) and validated it on 14 independent datasets (≈ 2700 patients). As sketched by Figure 5, we confirmed that molecular subtypes exhibit different clinical outcome.

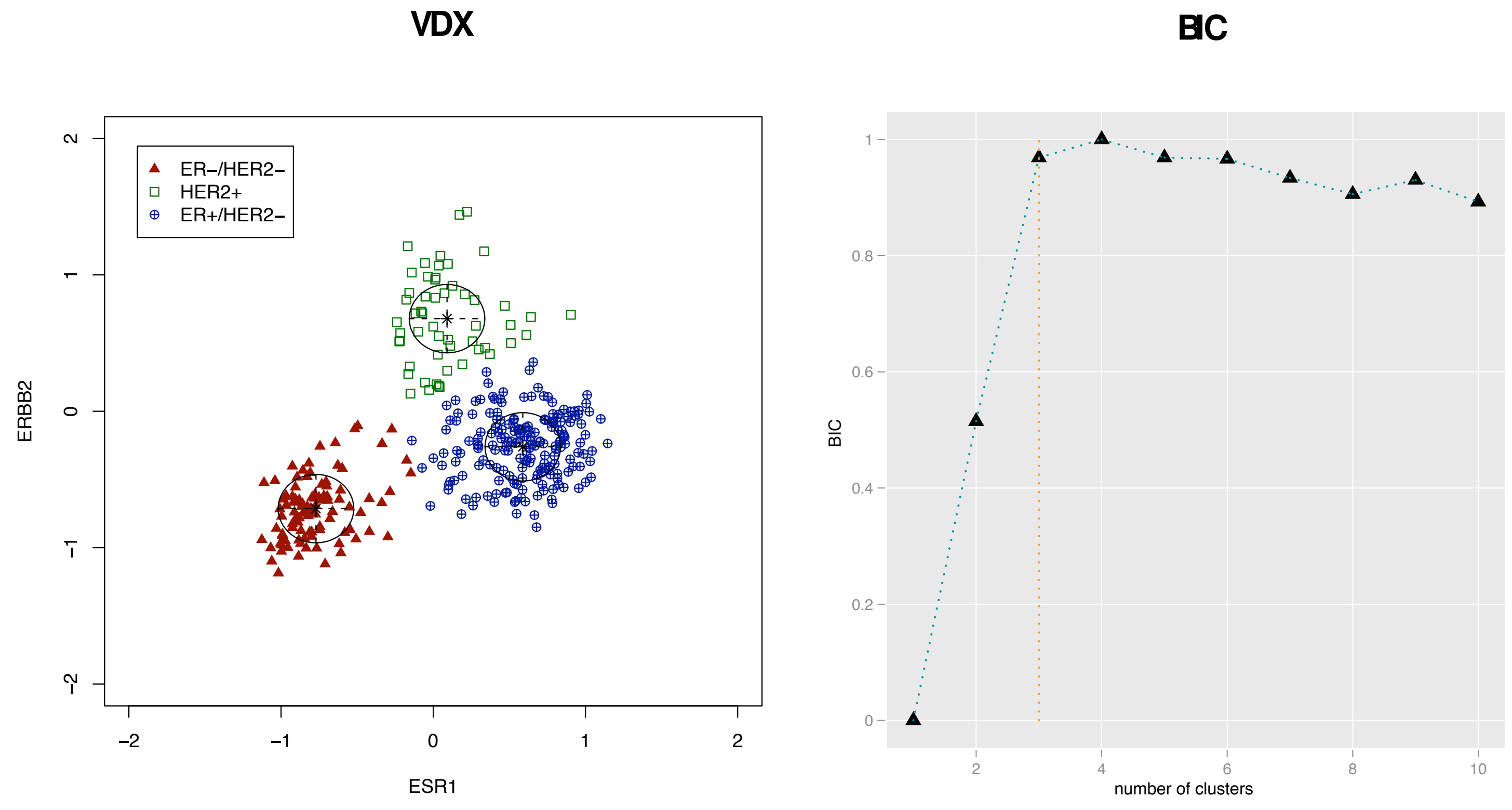

Figure 4: *Model-based clustering fitted on the training set (VDX).*

We computed the prediction strengths of our model in all the datasets and compared them to the estimates obtained for the SSP.

Dataset	ER-/HER2-	HER2+	ER+/HER2-
VDX	1.00	1.00	1.00
NKI	1.00	1.00	0.99
TBG	1.00	1.00	0.83
UPP	1.00	1.00	0.84
UNT	1.00	0.89	0.94
MAINZ	1.00	1.00	0.91
STNO2	1.00	0.69	0.97
NCI	0.85	0.83	0.93
MSK	1.00	1.00	0.96
STK	1.00	0.91	0.88
DUKE	1.00	0.82	0.92
UNC2	1.00	0.87	0.96
CAL	1.00	1.00	0.95
DUKE2	1.00	0.64	0.95
NCH	1.00	0.82	0.98

Dataset	Basal	ERBB2	LuminalA	LuminalB	Normal
VDX	0.86	0.00	0.50	1.00	0.00
NKI	0.97	0.33	0.64	0.54	0.35
TBG	0.82	0.80	0.50	0.54	0.54
UPP	0.51	0.56	0.94	0.52	0.67
UNT	0.39	0.18	0.88	0.45	0.62
MAINZ	0.53	0.00	0.90	0.83	0.00
STNO2	0.71	0.32	0.90	0.42	0.51
NCI	0.85	0.46	1.00	0.00	0.65
MSK	0.72	0.67	0.53	0.83	0.27
STK	1.00	0.00	0.49	0.47	0.46
DUKE	0.55	0.00	0.47	1.00	1.00
UNC2	0.40	0.37	0.98	0.60	0.50
CAL	0.66	0.59	0.53	1.00	0.60
DUKE2	0.97	1.00	0.46	1.00	0.90
NCH	0.19	0.36	0.88	0.32	0.42

Table 1: *Prediction strength for the model-based clustering (left) and the SSP (right).*

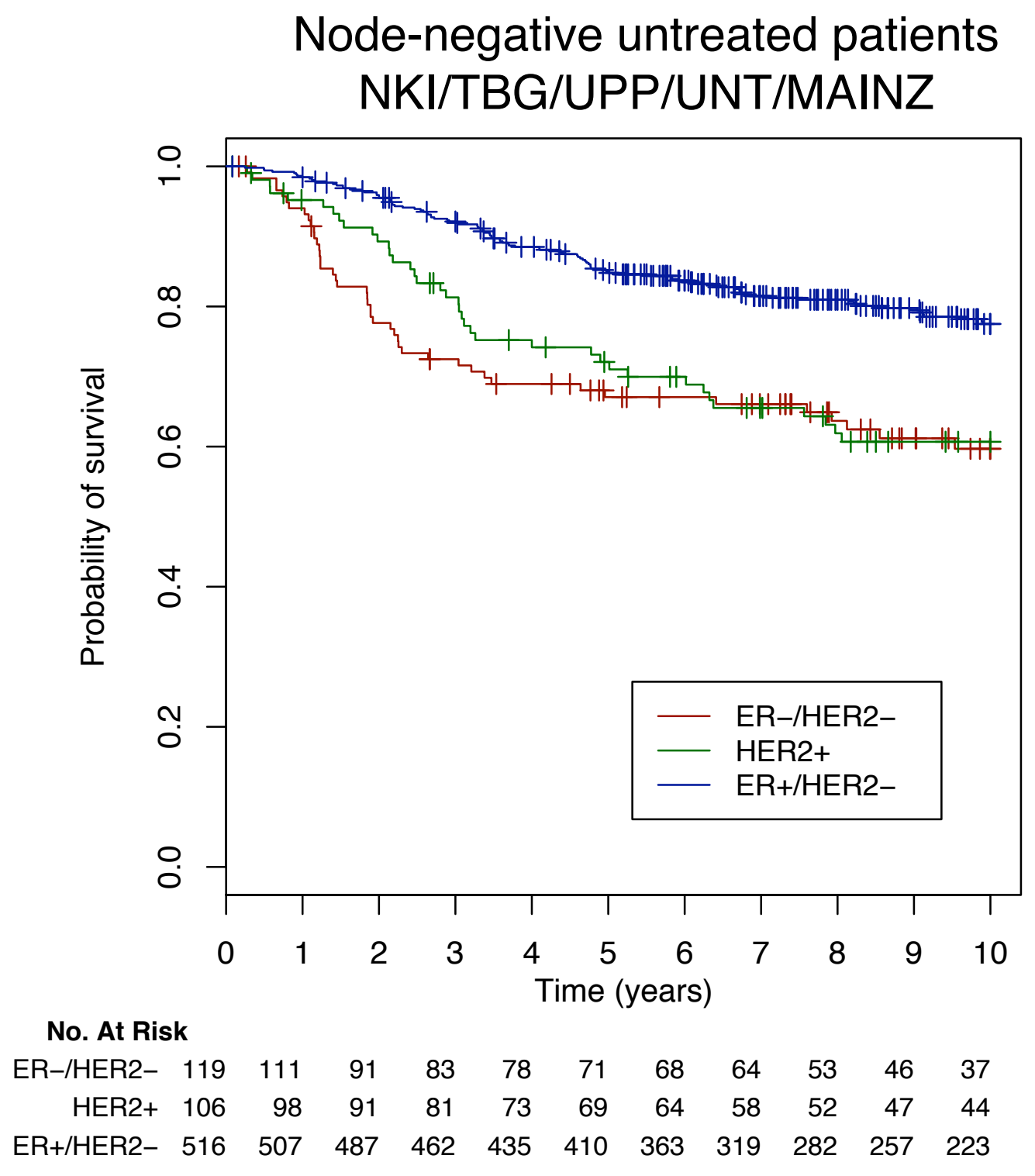

Figure 5: *Survival curves for each subtype.*

4. Conclusions

OUR novel method has several advantages compared to the previously published hierarchical clustering model (SSP): (i) the low-dimensionality of the input space (two dimensions) increases the stability of the clustering and facilitates the visualization of the clustering results; (ii) the model is easily applicable to new data; (iii) the model returns probabilities for a patient to belong to each subtype, facilitating the interpretation of the results (classification uncertainty). Moreover, this novel clustering model yields robust classifications in numerous microarray datasets. Given its easy applicability and its good performance, this new model could be used by doctors in order to study the prognosis and the effect of treatments with respect to the molecular subtypes of breast cancer.

This work was done in collaboration with the Swiss Institute for Experimental Cancer Research (Bioinformatics Core Facility headed by Mauro Delorenzi) and published in [1].

References

[1] C. Desmedt, B. Haibe-Kains, P. Wirapati, et al. Biological Processes Associated with Breast Cancer Clinical Outcome Depend on the Molecular Subtypes. *Clin Cancer Res*, 14(16):5158–5165, 2008. doi:10.1158/1078-0432.CCR-07-4756.

[2] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of American Statistical Asscoiation*, 97(458):611–631, 2002.

[3] A. Kapp, S. Jeffrey, A. Langerod, et al. Discovery and validation of breast cancer subtypes. *BMC Genomics*, 7(1):231, 2006. ISSN 1471-2164. doi:10.1186/1471-2164-7-231.

[4] L. Pusztai, C. Mazouni, K. Anderson, et al. Molecular Classification of Breast Cancer: Limitations and Potential. *Oncologist*, 11(8):868–877, 2006. doi:10.1634/theoncologist.11-8-868.

[5] T. Sorlie, C. M. Perou, R. Tibshirani, et al. Gene expression patterns breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Matl. Acad. Sci. USA*, 98(19):10869–10874, 2001.

[6] R. Tibshirani and G. Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005.