Gene Modules, Breast Cancer Subtypes and Prognosis

Benjamin Haibe-Kains[1,2]

[1]Functional Genomics Unit, Institut Jules Bordet

[2] Machine Learning Group, Université Libre de Bruxelles

November 19, 2008

Research Groups

Functional Genomics Unit (Christos Sotiriou)

- 9 researchers (1 Prof, 5 postDocs, 3 PhD students), 5 technicians.
- Research topics : Genomic analyses, clinical studies and translational research.
- Website : http://www.bordet.be/en/services/medical/array/practical.htm.
- National scientific collaborations : ULB, Erasme, ULg, Gembloux, IDDI.
- International scientific collaborations : Genome Institute of Singapore, John Radcliffe Hospital, Karolinska Institute and Hospital, MD Anderson Cancer Center, Netherlands Cancer Institute, Swiss Institute for Experimental Cancer Research, NCI/NIH, Gustave-Roussy Institute.

Research Groups

Machine Learning Group (Gianluca Bontempi)

- 10 researchers (2 Profs, 1 postDoc, 7 PhD students), 2 graduate students).
- Research topics : Bioinformatics, Classification, Regression, Time series prediction, Sensor networks.
- Website : http://www.ulb.ac.be/di/mlg.
- Scientific collaborations in ULB : IRIDIA, Physiologie Molculaire de la Cellule (IBMM), Conformation des Macromolcules Biologiques et Bioinformatique (IBMM), CENOLI (Sciences), Functional Genomics Unit (Institut Jules Bordet), Service d'Anesthesie (Erasme).
- Scientific collaborations outside ULB : UCL Machine Learning Group (B), Politecnico di Milano (I), Universitá del Sannio (I), George Mason University (US).
- The MLG is part to the "Groupe de Contact FNRS" on Machine Learning and to CINBIOS: http://babylone.ulb.ac.be/Joomla/.

Summary

- Introduction
- Breast Cancer Subtypes
  - New Clustering Model
    - Gene Modules
    - Model-Based Clustering
- Prognostic Gene Signatures
- Subtypes and Prognosis
- Conclusions

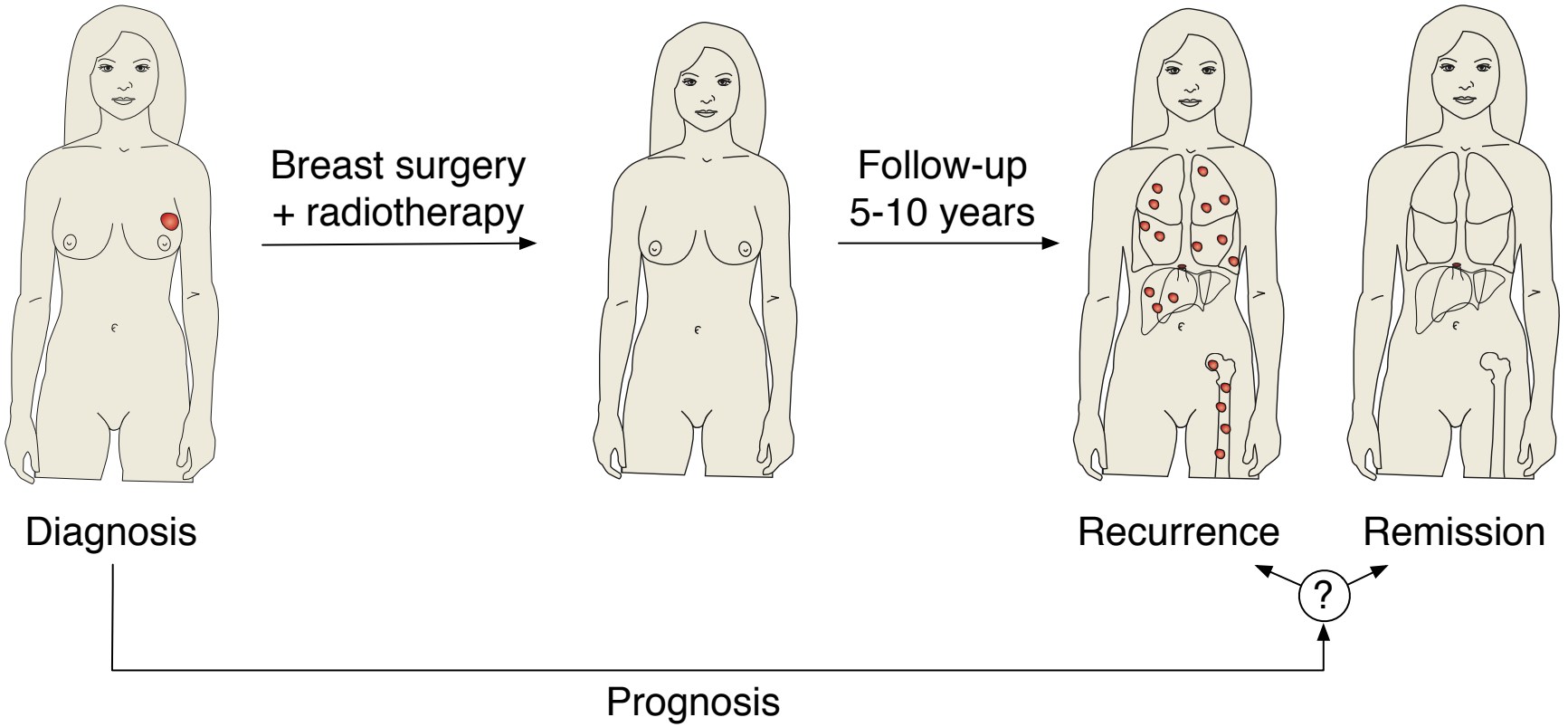
Part I

Introduction

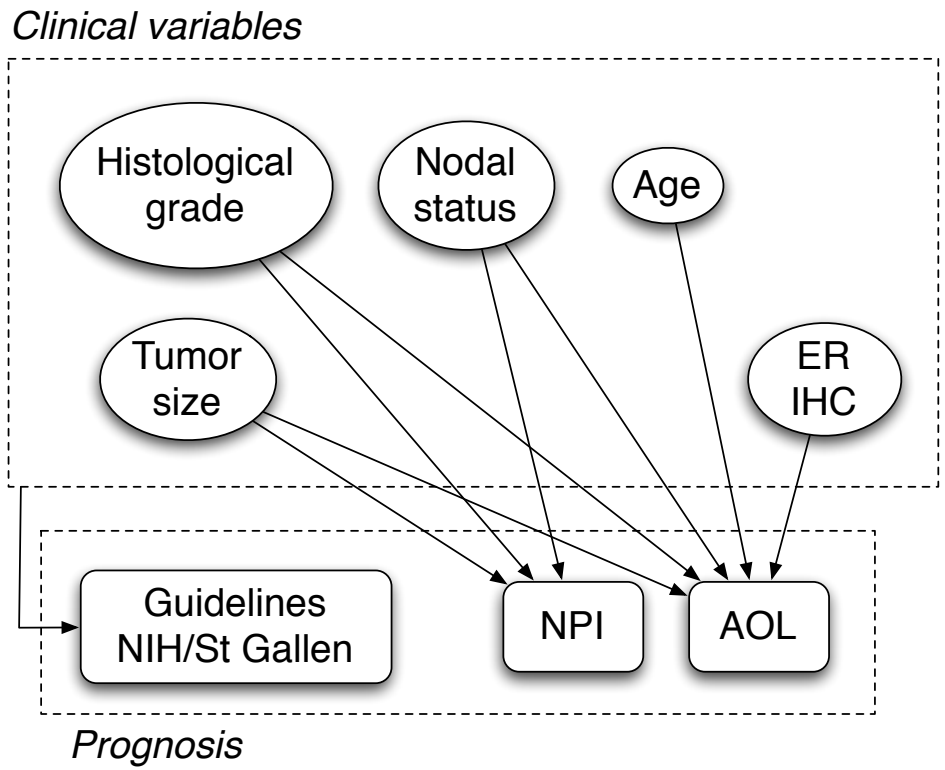
Breast Cancer

- Breast cancer is a global public health issue.
- It is the most frequently diagnosed malignancy in women in the western world and the commonest cause of cancer death for European and American women.
- In Europe, one out of eight to ten women, depending on the country, will develop breast cancer during their lifetime.

Breast Cancer Prognosis



Current Clinical Tools for Prognosis

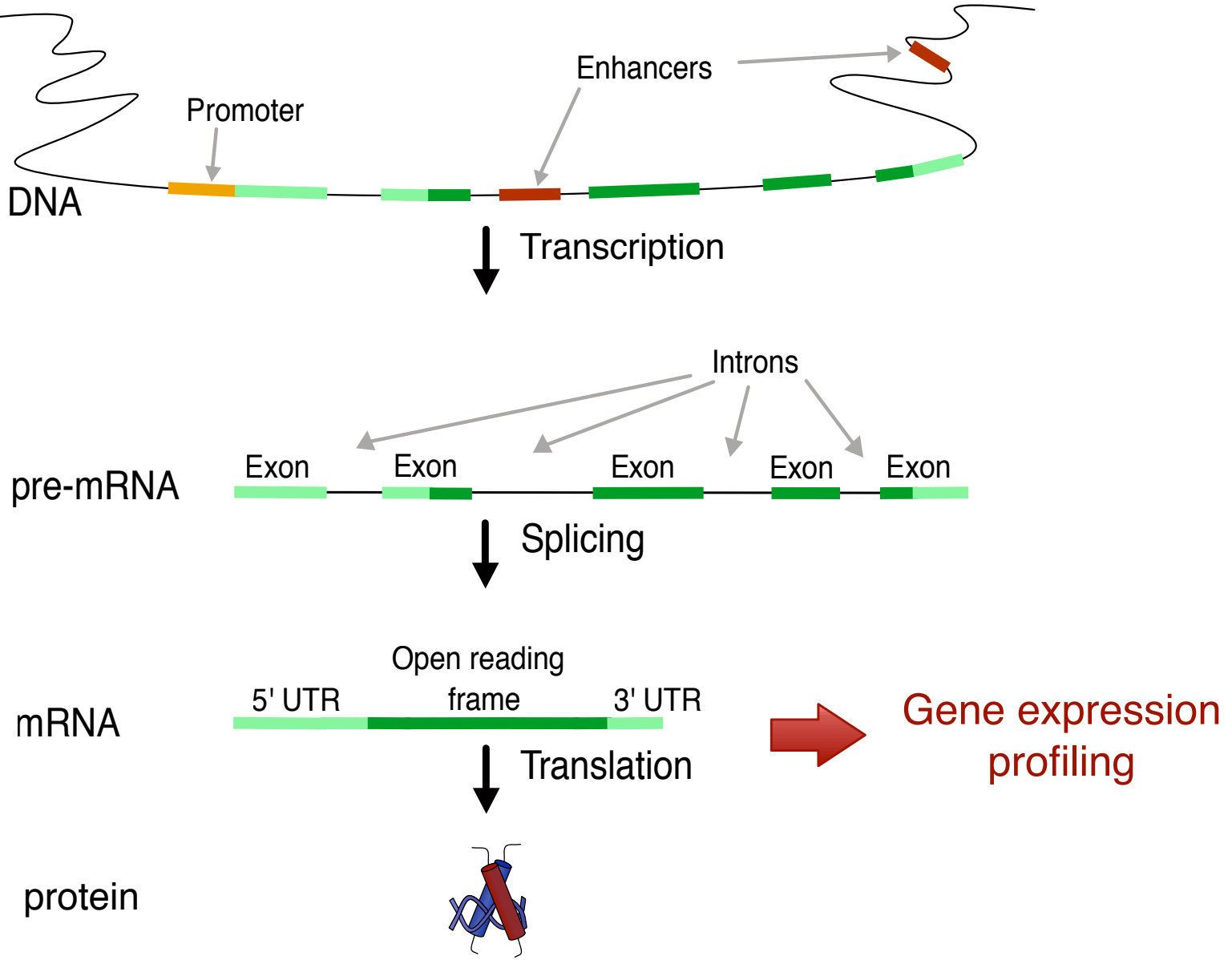


- Need to improve current clinical tools to detect patients who really need adjuvant systemic therapy.

Potential of Genomic Technologies for Prognosis

- In the nineties, new biotechnologies emerged:
  - Human genome sequencing.
  - Gene expression profiling (low to high-throughput).
- Genomic data could be used to better understand cancer biology
- . . . and to build efficient prognostic models.

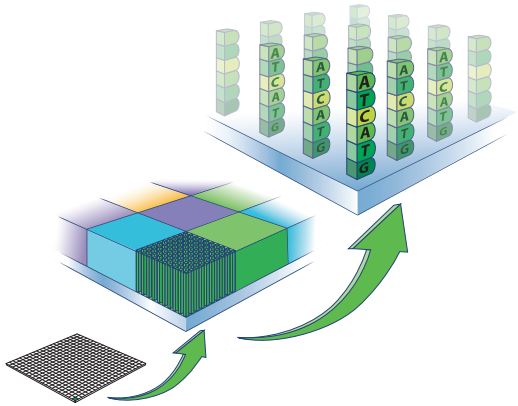
Biology Paradigm



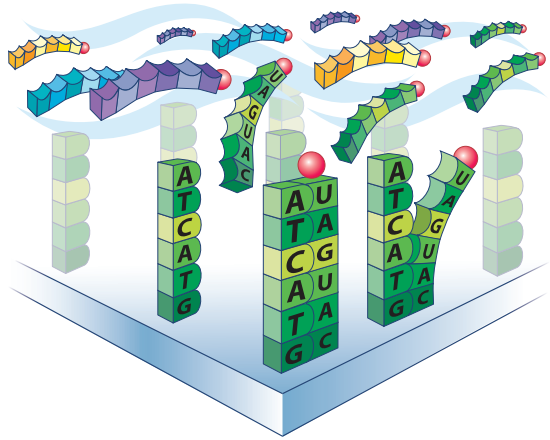
Gene Expression Profiling

- Gene expression profiling using microarray chip:

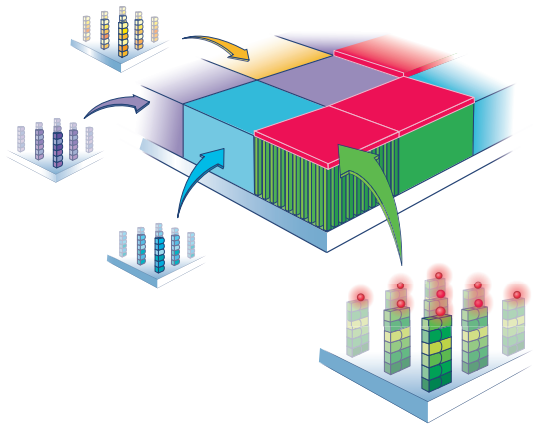
Microarray chip



Hybridization



Detection



Microarray Data

- Few samples (dozens to hundreds).
  - Microarray technology is expensive.
  - Frozen tumor samples are rare (biobank).
- On the other hand, numerous gene expressions are measured.
  - The recent microarray chips cover the whole genome ($\approx$ 50,000 probes representing 30,000 "known genes").

➠ High feature-to-sample ratio (curse of dimensionality).

- Microarray is a complex technology.

➠ High level of noise in the measurements.

- Biology is complex.

➠ Variables are highly correlated (gene co-expressions due to biological pathways).

Part II

Breast Cancer Molecular Subtypes

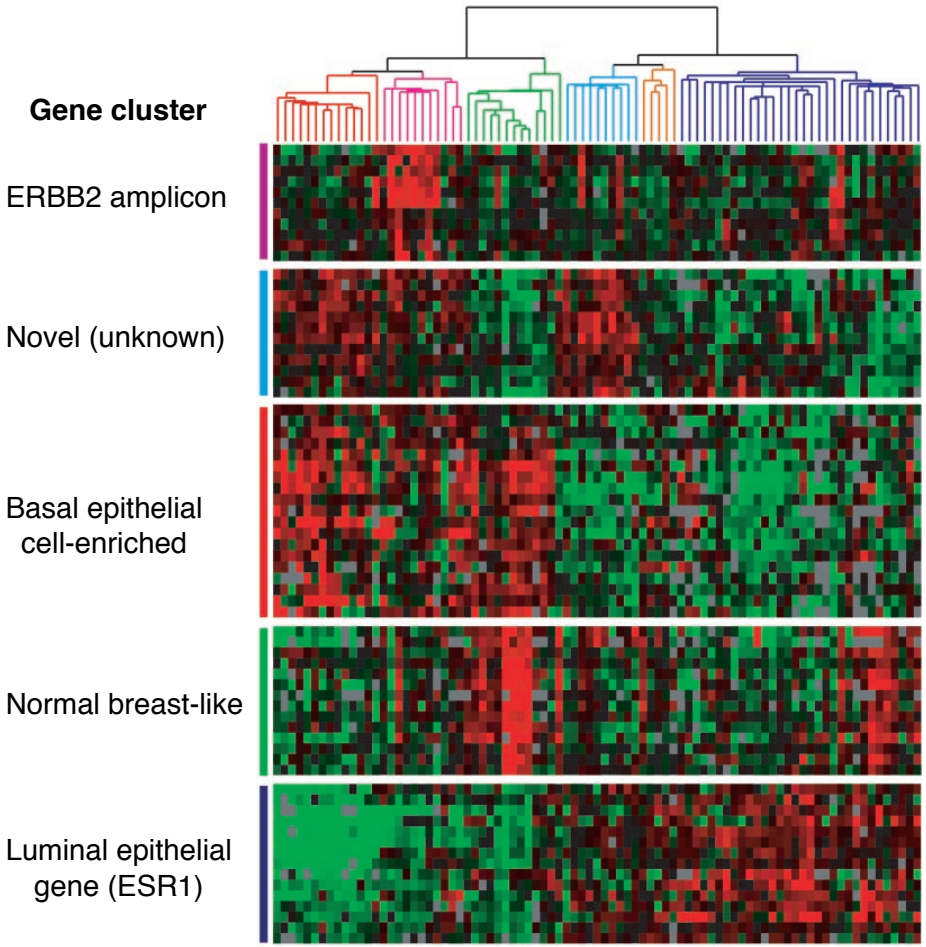
Molecular Heterogeneity

- Early microarray studies showed that breast cancer is a molecularly heterogeneous disease [Perou et al., 2000, Sorlie et al., 2001, Sorlie et al., 2003, Sotiriou et al., 2003].
  - Example: hierarchical clustering on microarray data [Sorlie et al., 2001].

- ➠ Identification of sets of co-expressed genes.
- ➠ Identification of groups of similar tumors.

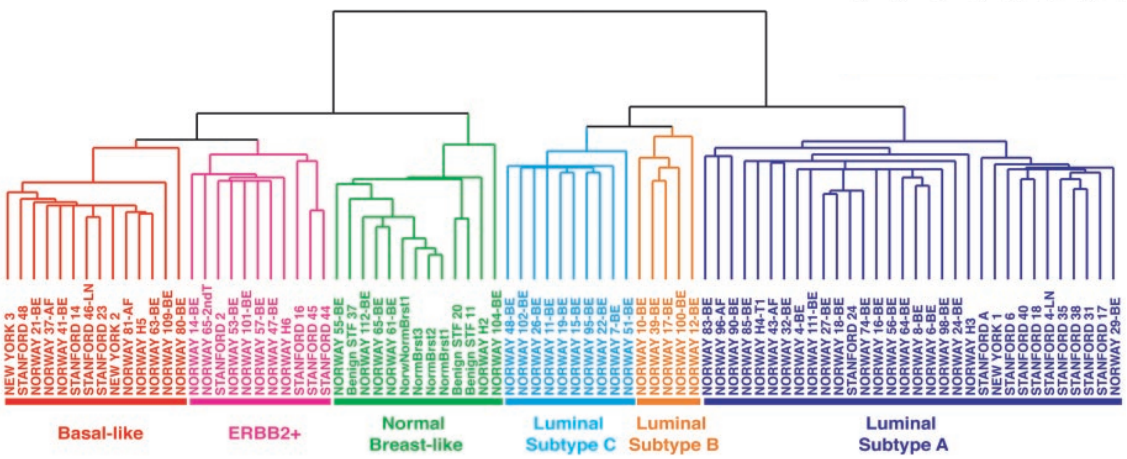
Gene Clusters

- Several gene clusters were identified to be the main discriminators of breast cancer molecular subtypes.
  - Example from [Sorlie et al., 2001]:

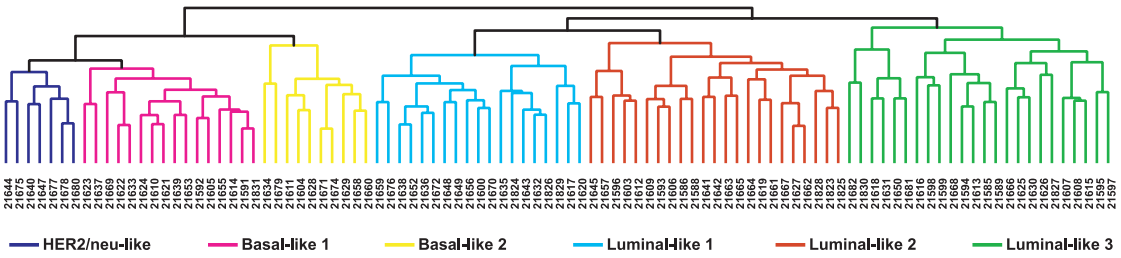


Tumor Clusters

- Perou et al. and Sotiriou et al. identified at least three breast cancer molecular subtypes:
  - Basal-like, mainly ER- and HER2- tumors.
  - ERBB2+ or HER2+ tumors.
  - Luminal-like, which could be further separated in low and high proliferative tumors [Loi et al., 2007].



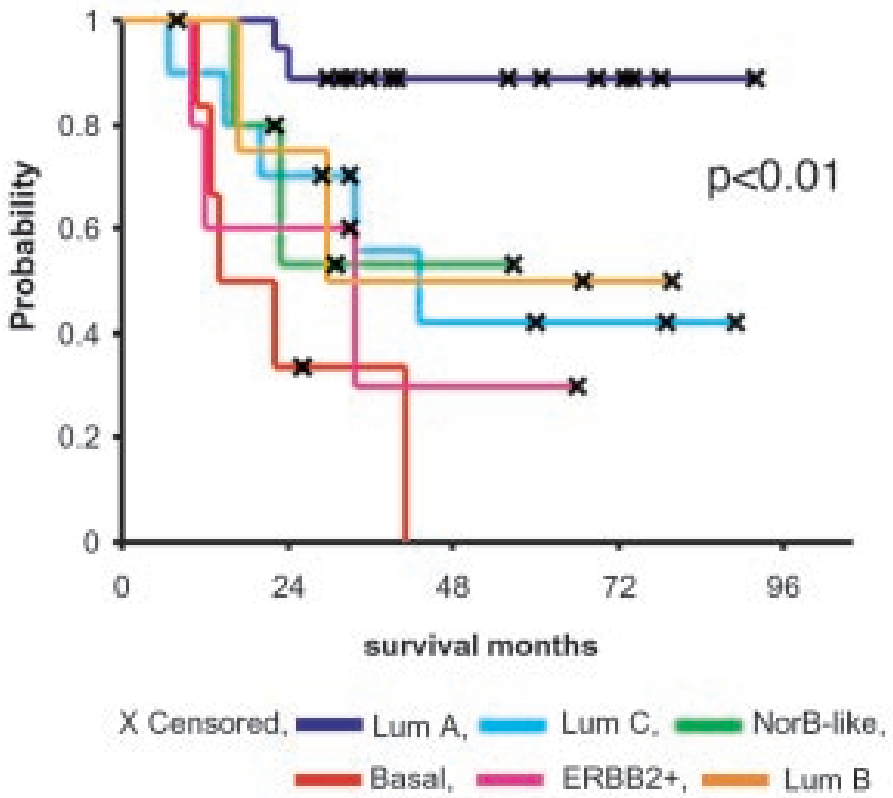
Perou et al.



Sotiriou et al.

Clinical Outcome

- The molecular subtypes exhibited different clinical outcomes, suggesting that the biological processes involved in patients' survival might be different.
  - Example from [Sorlie et al., 2001]:



Breast Cancer Subtypes

Early Results

- These early studies showed similar results, i.e. ER and HER2 phenotypes are the main discriminators in breast cancer (confirmed by [Kapp et al., 2006]).

- However, this classification has strong limitations [Pusztai et al., 2006]:
  - Instability: the results are hardly reproducible due to the instability of the hierarchical clustering method in combination with microarray data (high feature-to-sample ratio).
  - Crispness: hierarchical clustering produces crisp partition of the dataset (*hard partitioning*) without estimation of the classification uncertainty.
  - Validation: the hierarchical clustering is hardly applicable to new data.

New Clustering Model

- Because of these limitations we sought to develop a robust method to identify the breast cancer subtypes.

- This method consists in:
  1. A prototype-based clustering method to identify sets of co-expressed genes (gene modules).
  2. A model-based clustering in a low dimensional space to identify groups of similar tumors (subtypes).

Gene Modules

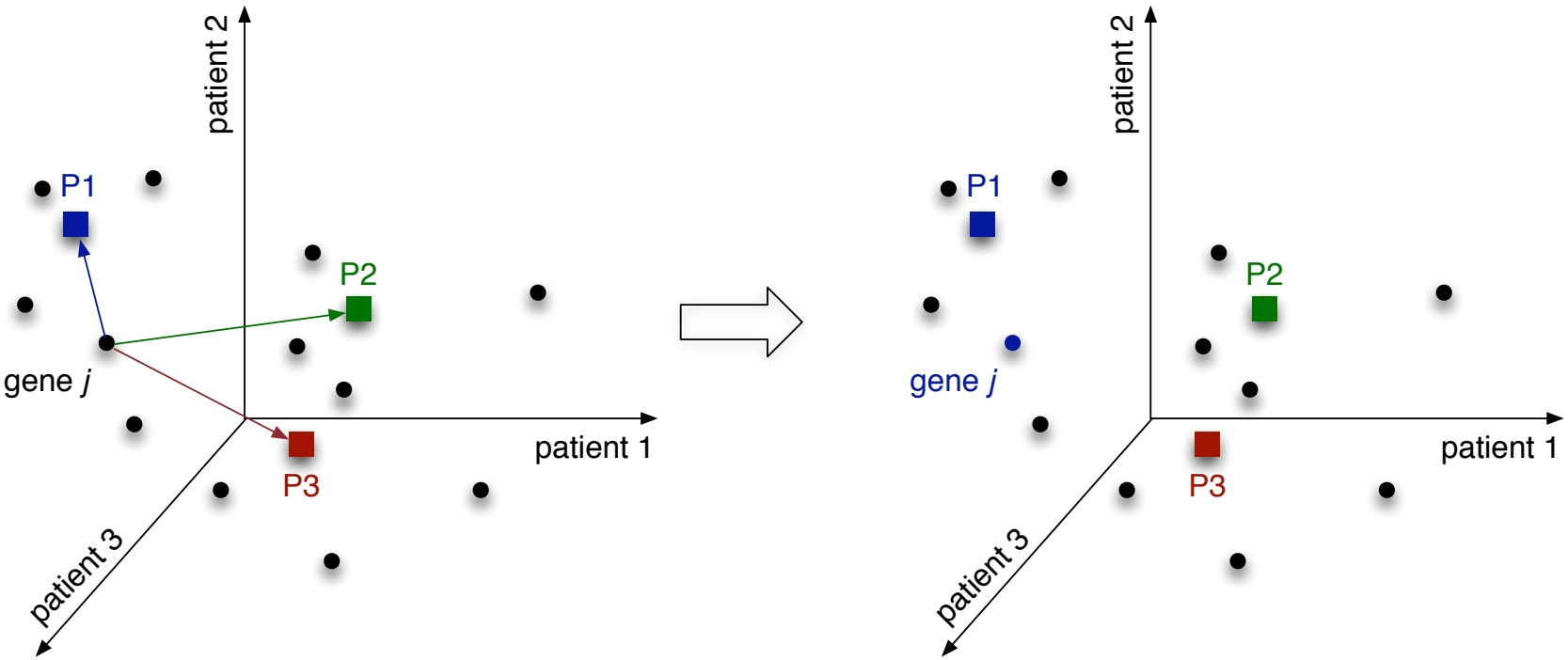
- Aim: identification of co-expressed genes related to a biological process of interest.
- Method:
  1. Choice of the biological processes of interest.
  2. Selection of a *prototype* for each biological process.
     - ⋆ A prototype is a gene known to be related to the biological process of interest (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.

➠ Computation of *gene module scores* by averaging the expressions of the genes in the modules.

Gene Modules

Example

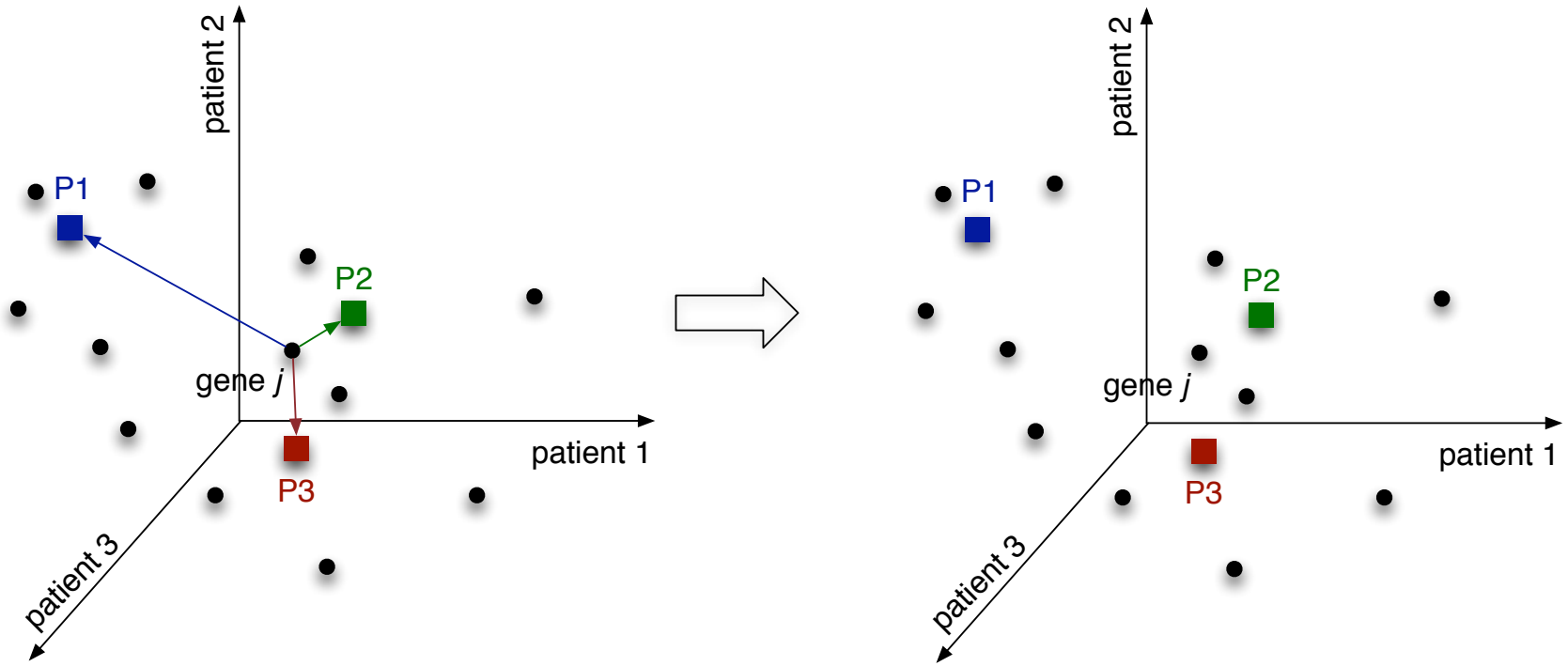
- Choice of three prototypes: P1, P2 and P3.
- Gene *j* assigned to the gene module 1 (prototype P1):



Gene Modules

Example (cont.)

- Gene j not assigned to any gene module:



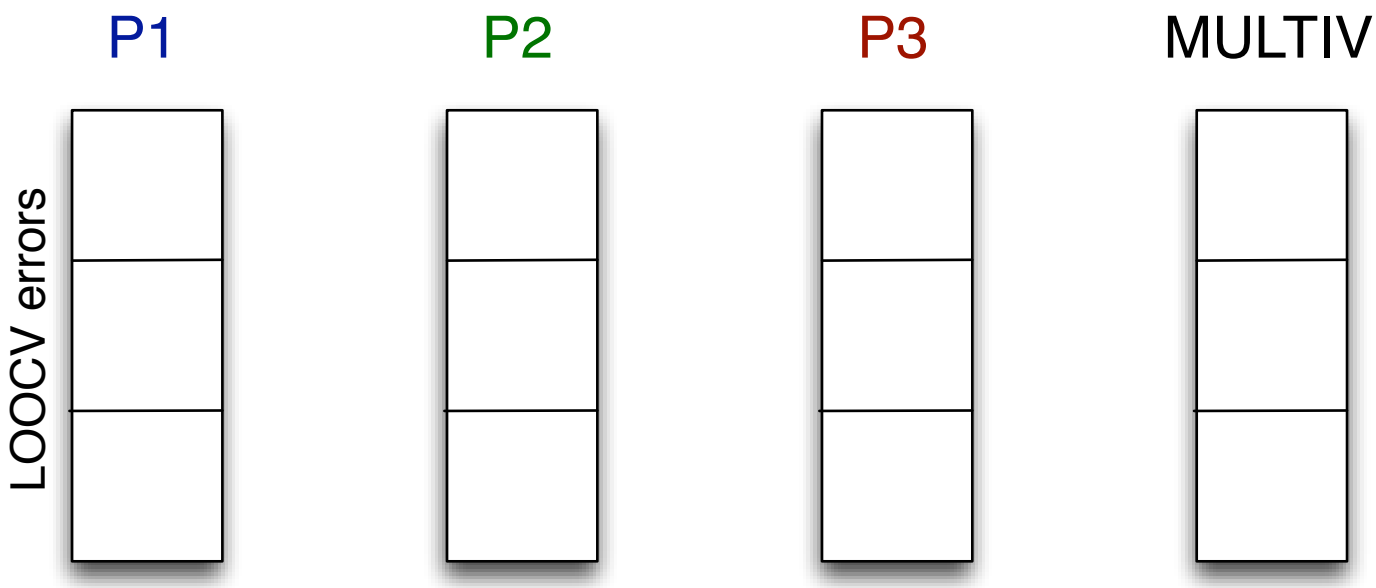
Gene Modules

Method

- We tackled the problem from a prediction point of view.
  - Basic idea: If a gene j is statistically better predicted by prototype q than by all the other prototypes, then gene j is specifically co-expressed with prototype q and is assigned to gene module q.
- For each gene j, we fit a set of linear models:
  - The univariate models using each prototype as explanatory variable and the gene j as response variable.
  - The "best" multivariate model.
- We compute the leave-one-out cross-validation (LOOCV) errors of these models.

Gene Modules

Method: LOOCV Errors



Gene Modules

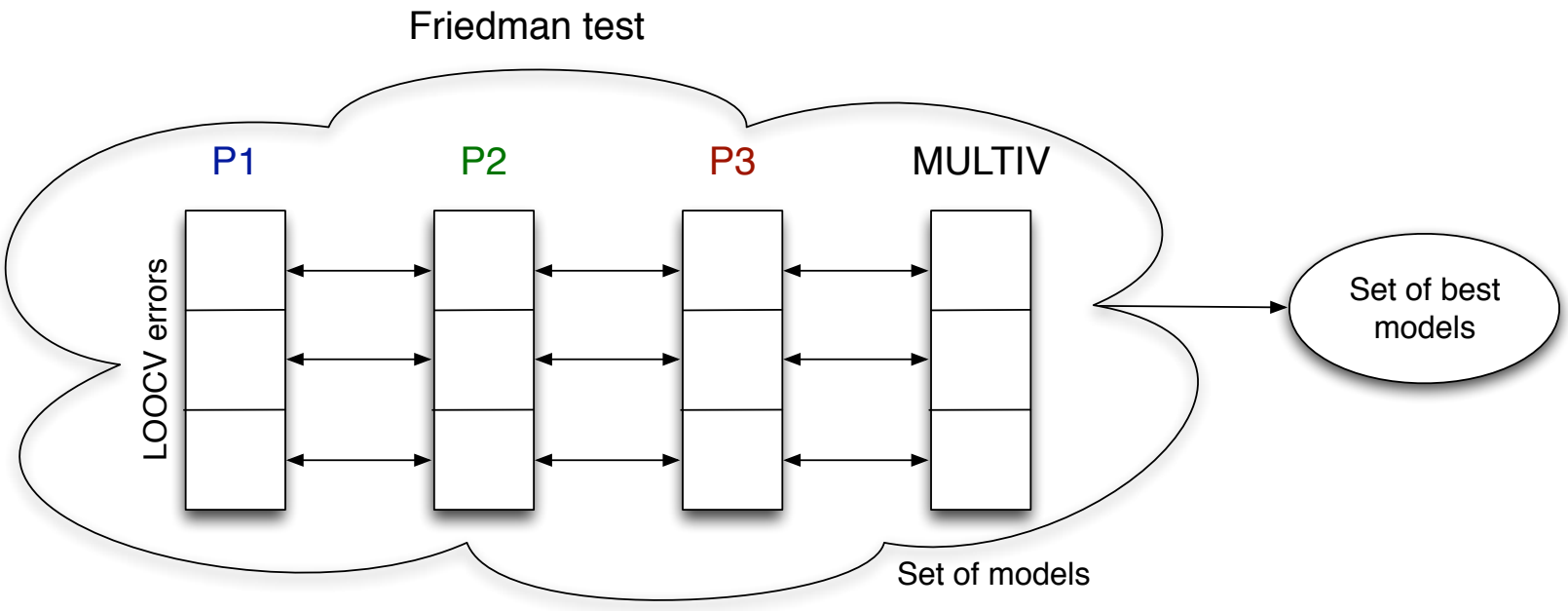
Method: Statistical Model Selection

- We statistically compare the models on the basis of their LOOCV errors (Friedman test).
- ➠ Identification of the set of best models.
  - The models present in the set are statistically better than the absent models.
  - The models present in the set have similar LOOCV errors.

- If there is only one univariate model (using prototype q) in the set of best models, gene j is assigned to the gene module q.

Gene Modules

Method: Statistical Model Selection (cont.)



Gene Modules

Method: Meta-Analysis

- If the sample size is small, many genes will not be assigned to any gene module (not enough evidence for the statistical model selection step).
- We can increase the size of gene modules by integrating several datasets (larger sample size).
- However, merging datasets from different laboratories, cohorts of patients and microarray platforms is a difficult task.

➠ Meta-analysis framework (each dataset is analyzed separately and results are combined).

Gene Modules

Method: Meta-Analysis (cont.)

- Compute the LOOCV errors for the univariate and multivariate models in each dataset separately.
- Check the homogeneity of the standardized coefficients of the univariate models over datasets (heterogeneity test).
  - The relation between gene j and the prototypes should be similar in each dataset.
  - If the coefficients are heterogeneous, discard gene j from the analysis (conservative way).
- Perform a "meta" Friedman test:
  - Combine the p-values returned by the pairwise tests applied in each dataset separately.
  - Consider these meta p-values in the traditional version of Friedman test.

Gene Modules

Method (dis)Advantages

- Advantages:
  - Robust to overfitting (linear models + LOOCV).
  - Control for other biological processes, i.e. prevent the use of highly correlated prototypes (gene modules are then small).
  - Integration of several datasets using different microarray platforms.
    - ★ Insensitive to "batch" effect (meta-analysis framework).
    - ★ Check for heterogeneity between datasets.
  - The gene module scores (signed average) are easily computable whatever the microarray technology (except for very small platforms).
  - Very conservative (control of false positive rate).

- Disadvantages:
  - Limited to linear relation between gene and prototypes.
  - Computationally intensive (statistical model selection step).
  - Very conservative (many false negatives).

Gene Modules

Results

In [Desmedt et al., 2008],

- We selected 7 prototypes to be representative of key biological processes involved in breast cancer:
  - ESR1 gene for ER phenotype.
  - ERBB2 gene for HER2 phenotype.
  - AURKA gene for proliferation
  - STAT1 gene for immune response.
  - VEGF gene for angiogenesis.
  - PLAU gene for tumor invasion.
  - CASP3 gene for apoptosis.
- We used 2 large breast cancer microarray datasets:
  - Wang *et al.* series: 286 node-negative patients on Affymetrix platform (22283 probes).
  - van de Vijver *et al.* series: 295 patients on Agilent microarray platform (24496 probes).

  ➟ $\approx$ 10,000 probes in common (mapping through EntrezGene IDs).

Gene Modules

Results (cont.)

- We found gene modules of various size:
  - ESR1 is the largest gene module as expected.
  - AURKA is the second one, highlighting the importance of proliferation.

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

- Gene ontology analysis confirmed the coherence of the gene modules with respect to the prototypes or biological processes of interest.

Model-Based Clustering

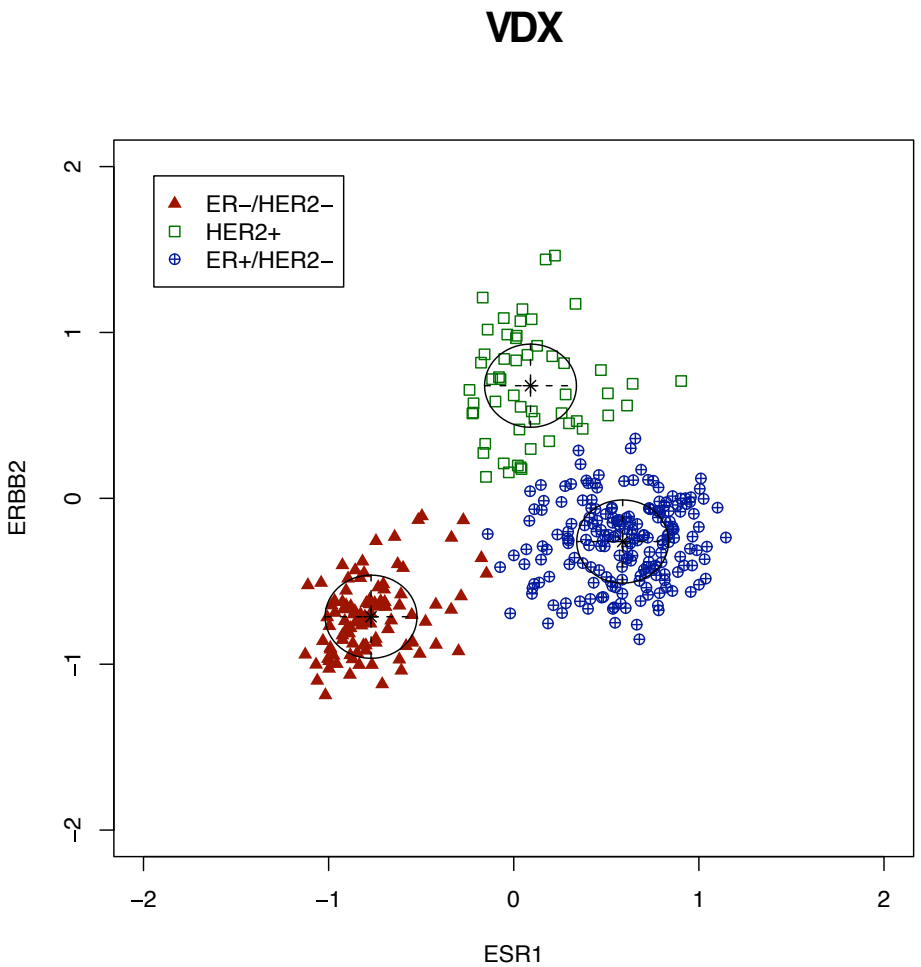
- We know from early microarray studies that breast cancer is a molecularly heterogeneous disease.
- ER and HER2 phenotypes seem to be the main (only?) discriminators.
- However the first classification models, based on hierarchical clustering, are hardly reproducible/applicable to new data.

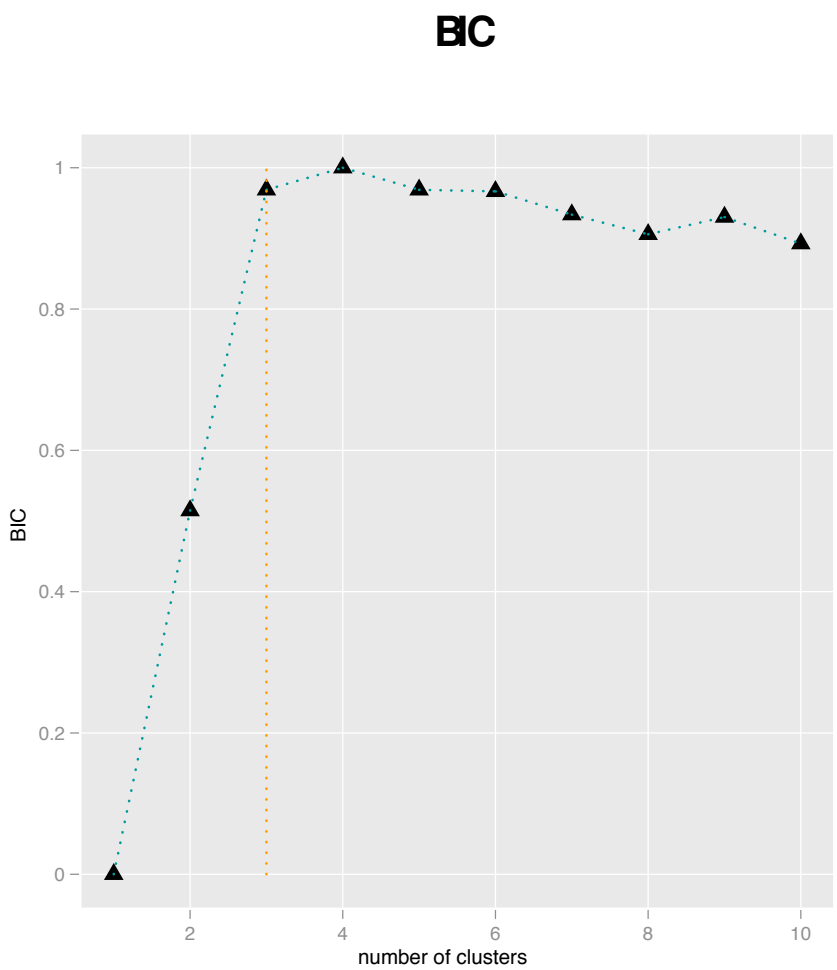
➠ We introduced a simple model-based clustering (mixture of Gaussians) in a two-dimensional space defined by the ESR1 and ERBB2 module scores.

- We used the Bayesian Information criterion (BIC) to select the most likely number of subtypes.
- We validated our model (fitted on Wang's series, VDX) on 14 independent datasets by estimating the prediction strength [Tibshirani and Walther, 2005].

Model-Based Clustering

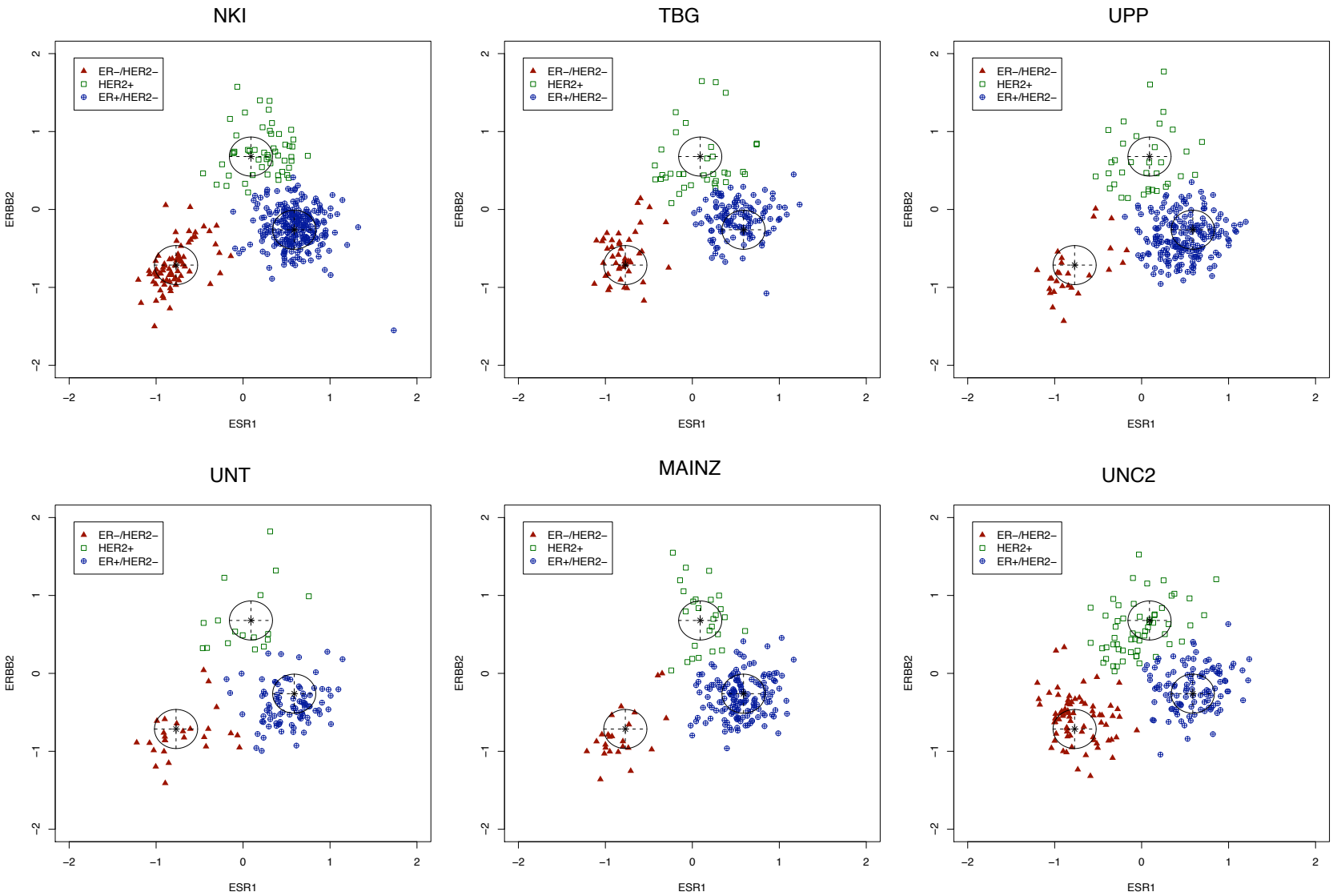
Training





Model-Based Clustering

Validation



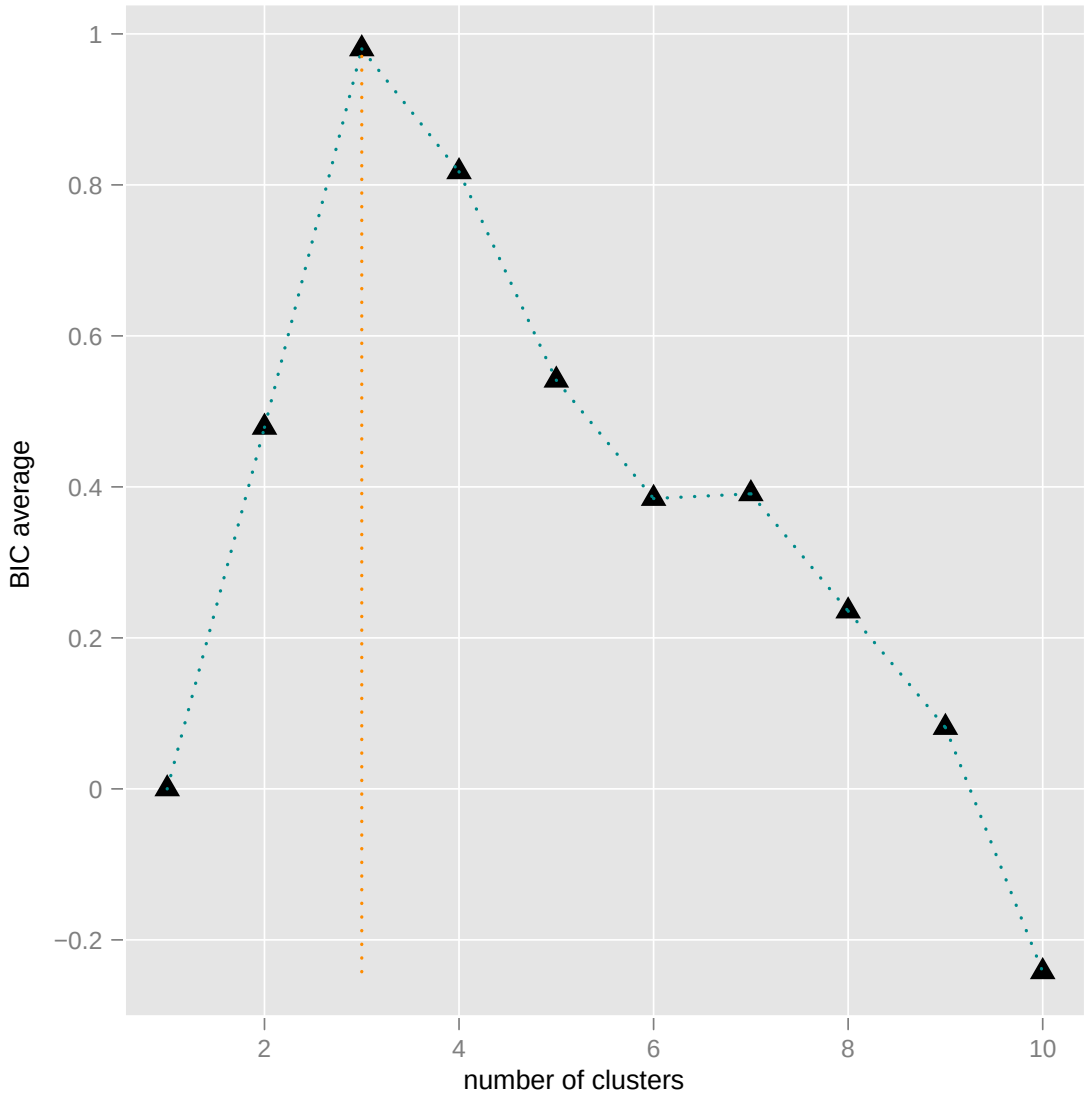
Model-Based Clustering

Validation: Prediction Strength

Reference	Dataset	ER-/HER2-	HER2+	ER+/HER2-
[van de Vijver et al., 2002]	NKI	1.00	1.00	0.99
[Desmedt et al., 2007]	TBG	1.00	1.00	0.83
[Miller et al., 2005]	UPP	1.00	0.93	0.87
[Sotiriou et al., 2006]	UNT	1.00	0.89	0.92
[Schmidt et al., 2008]	MAINZ	1.00	1.00	0.90
[Sorlie et al., 2003]	STNO2	1.00	0.69	0.97
[Sotiriou et al., 2003]	NCI	0.85	0.83	0.93
[Minn et al., 2005]	MSK	1.00	1.00	0.96
[Pawitan et al., 2005]	STK	1.00	0.91	0.87
[Bild et al., 2006]	DUKE	1.00	0.82	0.92
[Hoadley et al., 2007]	UNC2	1.00	0.87	0.96
[Chin et al., 2006]	CAL	1.00	1.00	0.95
[Bonnefoi et al., 2007]	DUKE2	1.00	0.64	0.95
[Naderi et al., 2007]	NCH	1.00	0.82	0.98

Model-Based Clustering

Validation: Number of Clusters

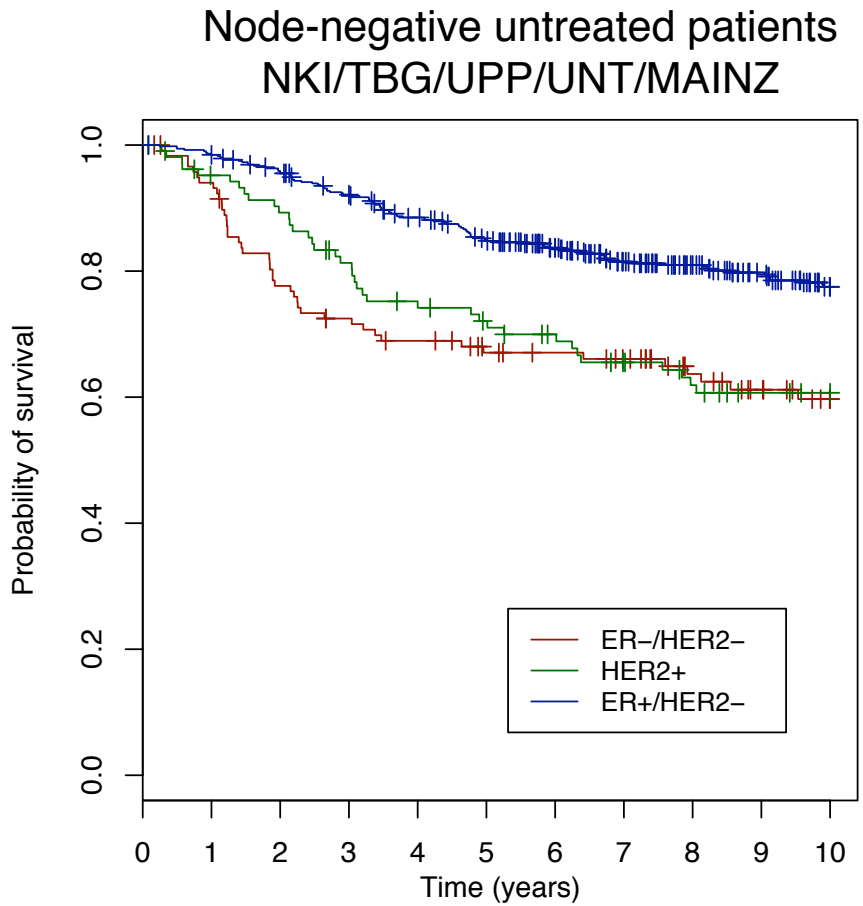


Breast Cancer Subtypes

Clinical Outcome

- ER-/HER2-: 20-25%
- HER2+: 15-20%
- ER+/HER2-: 60-70%

  of the global population of breast cancer patients.



No. At Risk

ER−/HER2−	119	111	91	83	78	71	68	64	53	46	37
HER2+	106	98	91	81	73	69	64	58	52	47	44
ER+/HER2−	516	507	487	462	435	410	363	319	282	257	223

- Advantages:
  - Simple model-based clustering:
    - ★ Easily applicable to new data.
    - ★ Returning for each patient the probability to belong to each subtype (*soft partitioning*).
  - Low dimensional space:
    - ★ Stability/robustness of the clustering model.
    - ★ Low computational cost to fit the model.
    - ★ Simple visualization of the results.

- Disadvantage:
  - Low dimensional space: which dimension could we add in order to find another robust subtype?

Part III

Prognostic Gene Signatures

Prognostic Gene Signatures

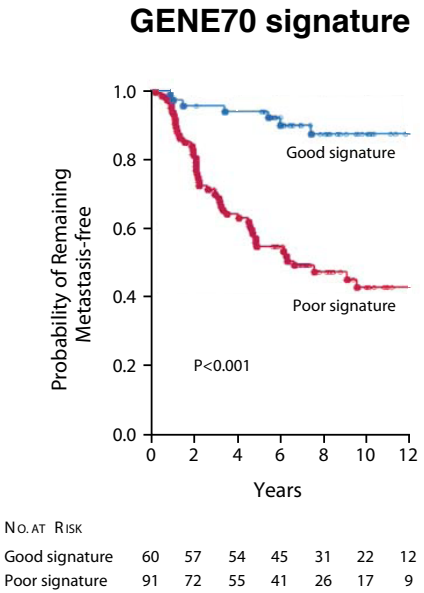
- Use of microarray technology to improve current prognostic models (NIH/St Gallen guidelines, NPI, AOL).

- A typical microarray analysis dealing with breast cancer prognostication involves 5 key steps:
  1. Data preprocessing: quality controls and normalization.
  2. Filtering: discard the genes exhibiting low expressions and/or low variance.
  3. Identification of a list of prognostic genes (called a *gene signature*).
  4. Building of a prognostic model, i.e. combination of the genes from the signature in order to predict the clinical outcome of the patients.
  5. Validation of the model performance and comparison with current prognostic models.

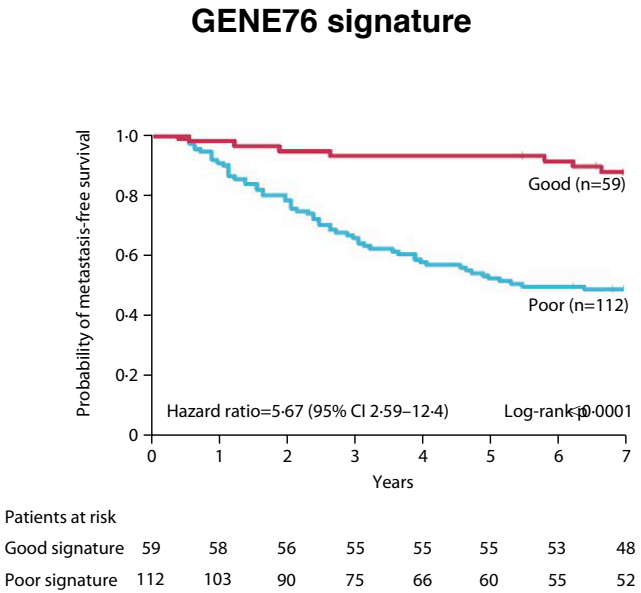
Prognostic Gene Signatures

Fishing Expedition

- Prognostic models derived from gene expression data by looking for genes associated with clinical outcome without any a priori biological assumption [van't Veer et al., 2002, Wang et al., 2005].



van't Veer et al.
van de Vijver



Wang et al.

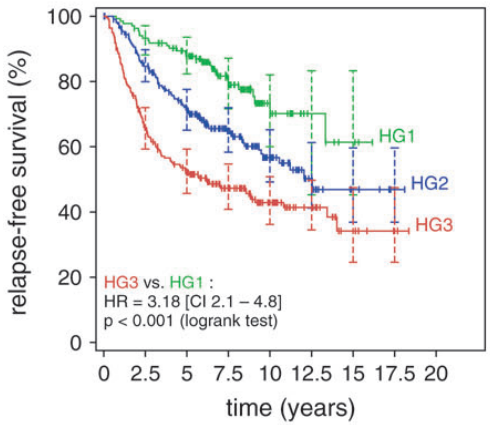
- Promising results but some criticisms from a statistical point of view.

Prognostic Gene Signatures

Hypothesis-driven

- Prognostic models were also derived from gene expression data based on a biological assumption.
  - Example: GGI [Sotiriou et al., 2006] was designed to discriminate patients with low and high histological grade (proliferation).
  - GGI was able to discriminate patients with intermediate histological grade (HG2).

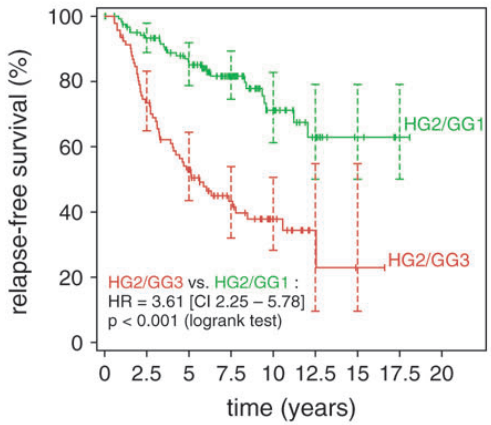
Histological grade



number at risk

HG1	134	123	107	59	23	8	4	
HG2	216	174	136	80	40	16	6	1
HG3	220	137	102	67	35	20	6	2
total	570	434	345	206	98	44	16	3

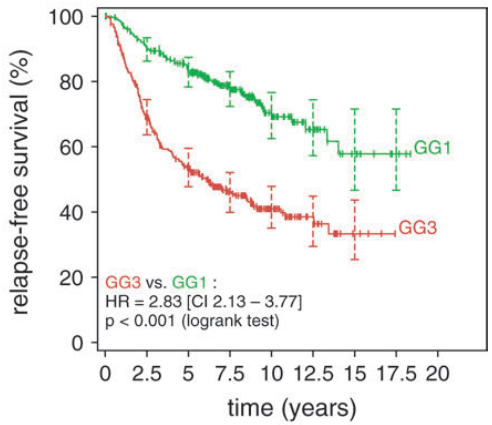
GGI (HG2)



number at risk

HG2/GG1	124	108	91	55	28	13	5	1
HG2/GG3	92	66	45	25	12	3	1	
total	216	174	136	80	40	16	6	1

GGI



number at risk

GG1	279	243	206	123	59	26	12	3
GG3	291	191	139	83	39	18	4	
total	570	434	345	206	98	44	16	3

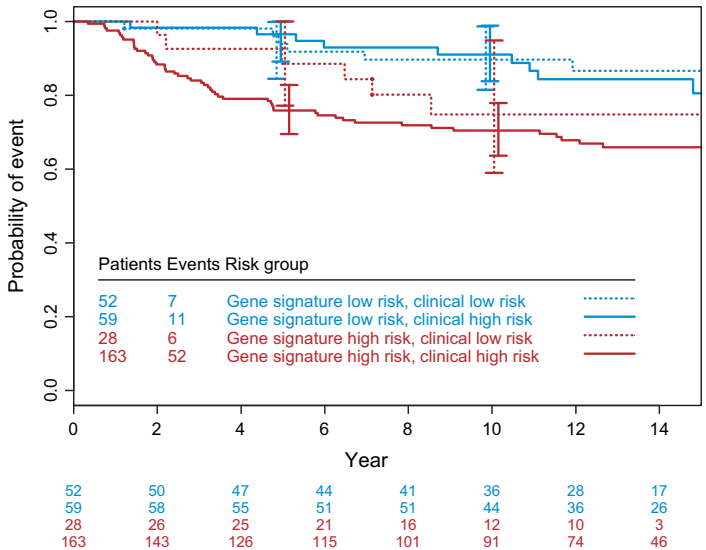
- These preliminary resulting were promising but validation was required.
- A first validation was published by the authors of the GENE70 and GENE76 signatures in [van de Vijver et al., 2002] and [Foekens et al., 2006] respectively.
- Our group was involved in a second validation:
  - Complete independence: the authors of the signatures were not aware of the clinical data of the patients in the dataset.
  - The statistical analyses were performed by an independent group.
  - Aim: validate definitively the prognostic power of these two models in order to start a large clinical trial called MINDACT (Microarray In Node negative Disease may Avoid ChemoTherapy).

Prognostic Gene Signatures

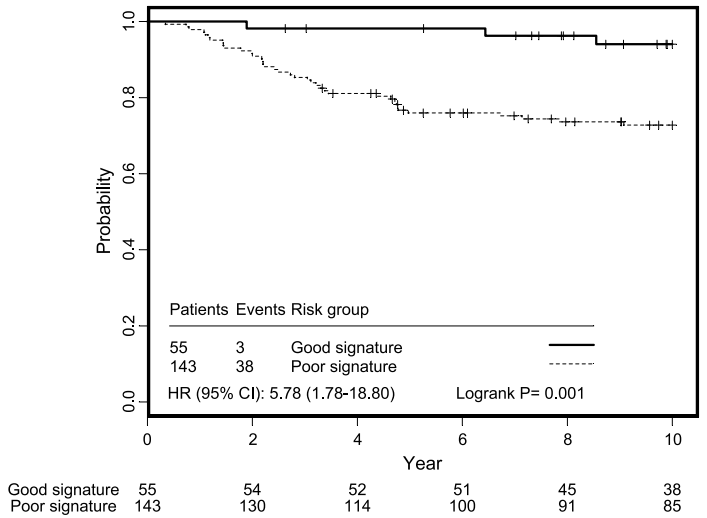
Independent Validation (cont.)

- Although the performance in this validation series was less impressive than in the original publications, GENE70 and GENE76 sufficiently improved the current clinical models to go ahead with MINDACT.

GENE70



GENE76

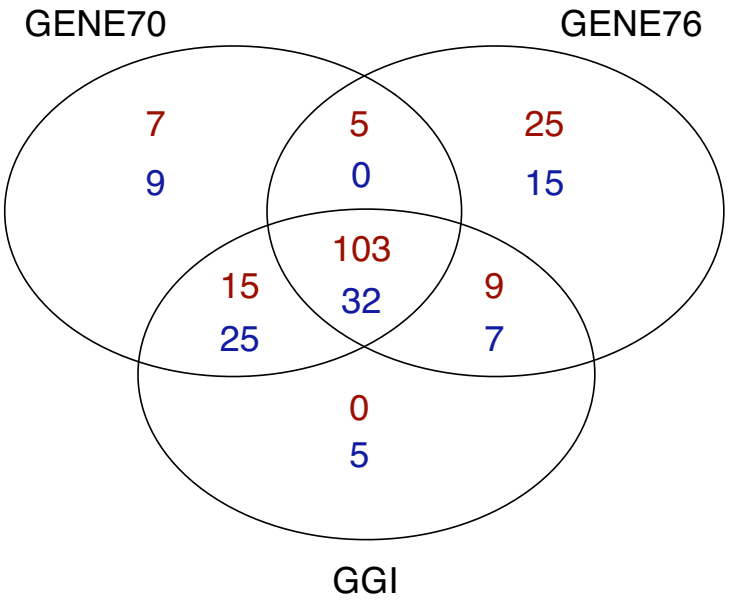


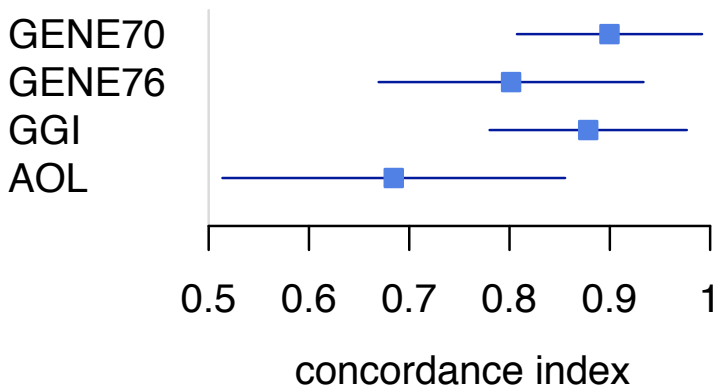
- Validation of GENE70 [Buyse et al., 2006] and GENE76 [Desmedt et al., 2007].

Prognostic Gene Signatures

Independent Validation (cont.)

- We sought to compare the GGI to the GENE70 and GENE76 signatures in this validation series
- ➠ GGI yields very similar performance [Haibe-Kains et al., 2008a].





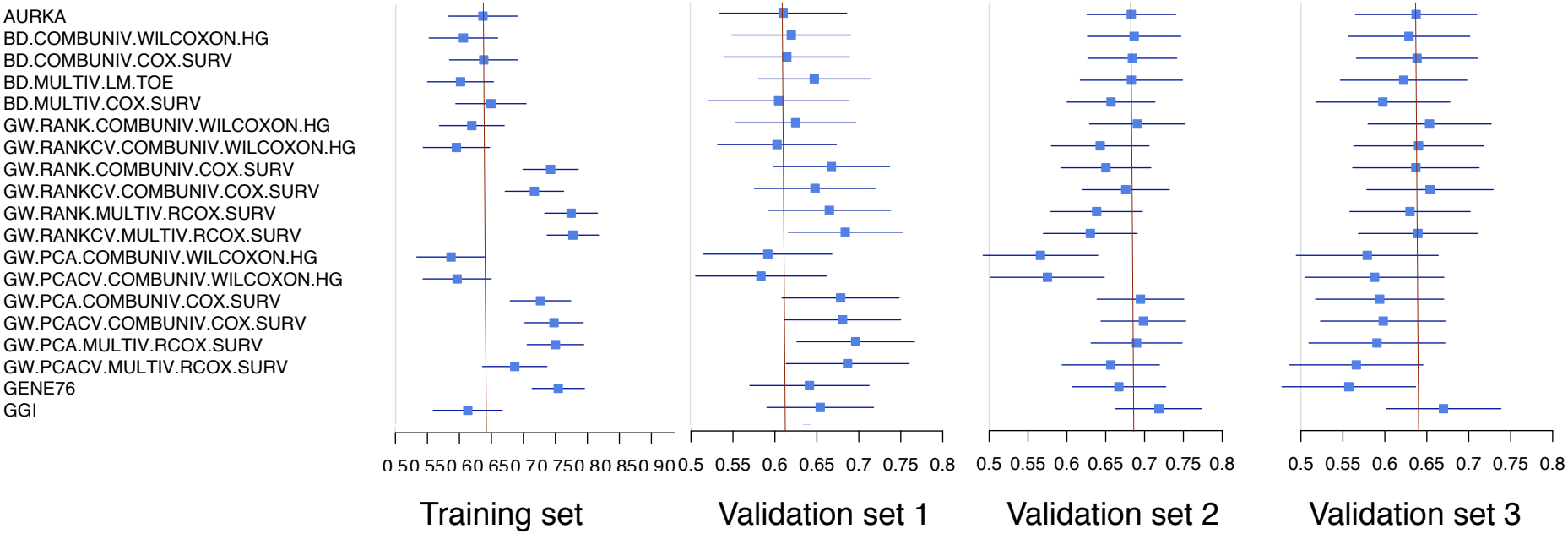
Prognostic Gene Signatures

A Single Gene?

- From the validation studies, we learned that GGI yields similar (sometimes better) performance than other gene signatures [Haibe-Kains et al., 2008a].
- Since GGI is a very simple model from a statistical and a biological (proliferation-related genes) points of view, we challenged the use of complex statistical methods for breast cancer prognostication.
- We compared simple to complex statistical methods to a single proliferation gene (AURKA) [Haibe-Kains et al., 2008b].

➠ Due to the complexity of microarray data, it is very hard to build prognostic models statistically better than AURKA.

- Forestplot of the concordance index for each method in the training set and the three validation sets:



Part IV

Subtypes and Prognosis

Prognosis in Specific Subtypes

- The first publications attempted to build a prognostic model from the global population of breast cancer patients.

- In 2005, Wang et al. were the first to divide the global population based on ER status:
  - As breast cancer biology is very different according to the ER status (IHC), prognostic models might be different too.
  - They built a prognostic model for each subgroup of patients (ER+ and ER-).
  - To make a prediction, they used one of the two models depending on the ER-status of the tumor.
  - Unfortunately the group of ER- tumors was too small and their corresponding model was not generalizable.

Prognosis in Specific Subtypes
(cont.)

- Recently, Teschendorff et al. built a new prognostic model for ER- tumors [Teschendorff et al., 2007] and validated it using large datasets [Teschendorff and Caldas, 2008].
  - The signature is composed of 7 immune-related genes.
- We showed in two meta-analyses [Wirapati et al., 2008, Desmedt et al., 2008] that:
  - Proliferation (AURKA) was the most prognostic factor in ER+/HER2- tumors and the common driving force of the early gene signatures.
    - ⋆ Actually, these signatures (e.g. GENE70, GENE76, GGI) are prognostic in ER+/HER2- tumors only.
  - Immune response (STAT1) is prognostic in ER-/HER2- and HER2+ tumors.
  - Tumor invasion (PLAU or uPA) is prognostic in HER2+ tumors.
- Finak et al. introduced a stroma-derived prognostic predictor (SDPP) particularly efficient in HER2+ tumors [Finak et al., 2008].

New Prognostic Model

- Since current prognostic models/gene signatures are limited to some subtypes, we plan to develop a new prognostic model integrating the breast cancer subtypes identification in order to:
  - Build a prognostic gene signatures specifically targeting each subtype.
  - Build a global prognostic model able to predict the risk of the patients whatever the tumor subtype (ER-/HER2-, HER2+ or ER+/HER2-).

- We plan to assess and to compare the performance of this new model to current prognostic models using the thorough statistical framework developed in [Haibe-Kains et al., 2008b].

- This new prognostic model is called *GENIUS*, standing for

Gene **E**xpression prog**N**ostic **I**ndex **U**sing **S**ubtypes ☺

Part V

Conclusion

Conclusion

- Numerous studies confirmed the great potential of gene expression profiling using microarrays to better understand cancer biology and to improve current prediction models.
- This technology becomes more and more mature (MAQC [MAQC Consortium, 2006]) and is now ready for clinical applications.
- The promising results of early publications were validated in different independent studies.
- Recent meta-analyses successfully recapitulated the main discoveries made these late decades and refined our knowledge on breast cancer biology.

Conclusion (cont.)

- We benefit from this strong basis to go a step further to improve breast cancer prognosis using microarrays.
  - Prognostic models/gene signatures in specific subtypes [Teschendorff et al., 2007, Desmedt et al., 2008, Finak et al., 2008].
  - Development of GENIUS, a prognostic model integrating breast cancer molecular subtypes identification [manuscript in preparation].

- A major issue remains: "How to combine these microarray prognostic models with clinical variables?"
  - Several studies showed the additional information of tumor size, nodal status, ...
  - However, we currently lack of data to fit robust prognostic models combining microarray and clinical variables.

Thank you for your attention.

This presentation is available from http://www.ulb.ac.be/di/map/bhaibeka/papers/haibekains2008gene.pdf.

Part VI

Bibliography

Bibliography I

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., Joshi, M.-B., Harpole, D., Lancaster, J. M., Berschuk, A., Olson Jr, J. A., Marks, J. R., Dressman, H. K., West, M., and Nevins, J. R. (2006).

Oncogenic pathway signatures in human cancers as a guide to targeted therapies.

Nature, 439:353–356.

Bonnefoi, H., Potti, A., Delorenzi, M., Mauriac, L., Campone, M., Tubiana-Hulin, M., Petit, T., Rouanet, P., Jassem, J., Blot, E., Becette, V., Farmer, P., André, S., Acharya, C. R., Mukherjee, S., Cameron, D., Bergh, J., Nevins, J. R., and Iggo, R. (2007).

Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the eortc 10994/big 00-01 clinical trial.

Lancet Oncology, 8(12):1071–1078.

Bibliography II

Buyse, M., Loi, S., van't Veer, L., Viale, G., Delorenzi, M., Glas, A. M., Saghatchian d'Assignies, M., Bergh, J., Lidereau, R., Ellis, P., Harris, A., Bogaerts, J., Therasse, P., Floore, A., Amakrane, M., Piette, F., Rutgers, E., Sotiriou, C., Cardoso, F., and Piccart, M. J. (2006).

Validation and Clinical Utility of a 70-Gene Prognostic Signature for Women With Node-Negative Breast Cancer.

J. Natl. Cancer Inst., 98(17):1183–1192.

Chin, K., Devries, S., Fridlyand, J., Spellman, P., Roydasgupta, R., Kuo, W. L., Lapuk, A., Neve, R., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B., Esserman, L., Albertson, D., Waldman, F., and Gray, J. (2006).

Genomic and transcriptional aberrations linked to breast cancer pathophysiologies.

Cancer cell, 10:529–41.

Bibliography III

Desmedt, C., Haibe-Kains, B., Wirapati, P., Buyse, M., Larsimont, D., Bontempi, G., Delorenzi, M., Piccart, M., and Sotiriou, C. (2008).

Biological Processes Associated with Breast Cancer Clinical Outcome Depend on the Molecular Subtypes.

Clin Cancer Res, 14(16):5158–5165.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M. S., Bergh, J., Lidereau, R., Ellis, P., Harris, A. L., Klijn, J. G., Foekens, J. A., Cardoso, F., Piccart, M. J., Buyse, M., and Sotiriou, C. (2007).

Strong Time Dependence of the 76-Gene Prognostic Signature for Node-Negative Breast Cancer Patients in the TRANSBIG Multicenter Independent Validation Series.

Clin Cancer Res, 13(11):3207–3214.

Finak, G., Bertos, N., Pepin, F., Sadekova, S., Souleimanova, M., Zhao, H., Chen, H., Omeroglu, G., Meterissian, S., Omeroglu, A., Hallett, M., and Park, M. (2008).

Stromal gene expression predicts clinical outcome in breast cancer.

Nat Med, 14(5):518–527.

Bibliography IV

Foekens, J. A., Atkins, D., Zhang, Y., Sweep, F. C., Harbeck, N., Paradiso, A., Cufer, T., Sieuwerts, A. M., Talantov, D., Span, P. N., Tjan-Heijnen, V. C., Zito, A. F., Specht, K., Hioefler, H., Golouh, R., Schittulli, F., Schmitt, M., Beex, L. V., Klijn, J. G., and Wang, Y. (2006).

Multicenter validation of a gene expression–based prognostic signature in lymph node–negative primary breast cancer.

Journal of Clinical Oncology, 24(11).

Haibe-Kains, B., Desmedt, C., Piette, F., Buyse, M., Cardoso, F., van't Veer, L., Piccart, M., Bontempi, G., and Sotiriou, C. (2008a).

Comparison of prognostic gene expression signatures for breast cancer.

BMC Genomics, 9(1):394.

Haibe-Kains, B., Desmedt, C., Sotiriou, C., and Bontempi, G. (2008b).

A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all?

Bioinformatics, 24(19):2200–2208.

Bibliography V

Hoadley, K., Weigman, V., Fan, C., Sawyer, L., He, X., Troester, M., Sartor, C., Rieger-House, T., Bernard, P., Carey, L., and Perou, C. (2007).

Egfr associated expression profiles vary with breast tumor subtype.

BMC Genomics, 8(1):258.

Kapp, A., Jeffrey, S., Langerod, A., Borresen-Dale, A.-L., Han, W., Noh, D.-Y., Bukholm, I., Nicolau, M., Brown, P., and Tibshirani, R. (2006).

Discovery and validation of breast cancer subtypes.

BMC Genomics, 7(1):231.

Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A. M., Gillet, C., Ellis, P., Harris, A., Bergh, J., Foekens, J. A., Klijn, J. G., Larsimont, D., Buyse, M., Bontempi, G., Delorenzi, M., Piccart, M. J., and Sotiriou, C. (2007).

Definition of Clinically Distinct Molecular Subtypes in Estrogen Receptor-Positive Breast Carcinomas Through Genomic Grade.

J Clin Oncol, 25(10):1239–1246.

Bibliography VI

MAQC Consortium (2006).

The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements.

Nat Biotech, 24(9):1151–1161.

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Pioner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T., and Bergh, J. (2005).

An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival.

PNAS, 102(38):13550–13555.

Minn, A. J., Gupta, G. P., Siegel, P. M., Bos, P. D., Shu, W., Giri, D. D., Viale, A., Olshen, A. B., Gerald, W. L., and Massague, J. (2005).

Genes that mediate breast cancer metastasis to lung.

Nature, 436(7050):518–524.

Bibliography VII

Naderi, A., Teschendorff, A. E., Barbosa-Morais, N. L., Pinder, S. E., Green, A. R., andJ . F. Robertson, D. G. P., Aparicio, S., Ellis, I. O., Brenton, J. D., and Caldas, C. (2007).

A gene-expression signature to predict survival in breast cancer across independent data sets.

Oncogene, 26:1507–1516.

Pawitan, Y., Bjohle, J., Amler, L., Borg, A., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., Liu, E. T., Miller, L., Nordgren, H., Ploner, A., Sandelin, K., Shaw, P. M., Smeds, J., Skoog, L., Wedren, S., and Bergh, J. (2005).

Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts.

Breast Cancer Research, 7(6):953–964.

Bibliography VIII

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A.-L., Brown, P. O., and Botstein, D. (2000).

Molecular portraits of human breast tumours.

Nature, 406(6797):747–752.

Pusztai, L., Mazouni, C., Anderson, K., Wu, Y., and Symmans, W. F. (2006).

Molecular Classification of Breast Cancer: Limitations and Potential.

Oncologist, 11(8):868–877.

Schmidt, M., Bohm, D., von Torne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H.-A., Hengstler, J. G., Kolbl, H., and Gehrmann, M. (2008).

The Humoral Immune System Has a Key Prognostic Impact in Node-Negative Breast Cancer.

Cancer Res, 68(13):5405–5413.

Bibliography IX

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisher, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese., J. C., Brown, P. O., Botstein, D., Eystein, P. L., and Borresen-Dale, A. L. (2001).

Gene expression patterns breast carcinomas distinguish tumor subclasses with clinical implications.

Proc. Matl. Acad. Sci. USA, 98(19):10869–10874.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geister, S., Demeter, J., Perou, C., Lonning, P. E., Brown, P. O., Borresen-Dale, A. L., and Botstein, D. (2003).

Repeated observation of breast tumor subtypes in indepedent gene expression data sets.

Proc Natl Acad Sci USA, 1(14):8418–8423.

Sotiriou, C., Neo, S. Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S., Harris, A. L., and Liu, E. T. (2003).

Breast cancer classification and prognosis based on gene expression profiles from a population-based study.

Proc. Natl. Acad. Sci., 100(18):10393–10398.

Bibliography X

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M. J., Bergh, J., Piccart, M., and Delorenzi, M. (2006).

Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis.

J. Natl. Cancer Inst., 98(4):262–272.

Teschendorff, A. and Caldas, C. (2008).

A robust classifier of high predictive value to identify good prognosis patients in er-negative breast cancer.

Breast Cancer Research, 10(4):R73.

Teschendorff, A., Miremadi, A., Pinder, S., Ellis, I., and Caldas, C. (2007).

An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer.

Genome Biology, 8(8):R157.

Bibliography XI

Tibshirani, R. and Walther, G. (2005).

Cluster validation by prediction strength.

Journal of Computational and Graphical Statistics, 14(3):511–528.

van de Vijver, M. J., He, Y. D., van't Veer, L., Dai, H., Hart, A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002).

A gene expression signature as a predictor of survival in breast cancer.

New England Journal of Medicine, 347(25):1999–2009.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhiven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002).

Gene expression profiling predicts clinical outcome of breast cancer.

Nature, 415:530–536.

Bibliography XII

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., van Gelder, M. E. M., Yu, J., Jatkoe, T., Berns, E. M., Atkins, D., and Forekens, J. A. (2005).

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

Lancet, 365:671–679.

Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B., Desmedt, C., Ignatiadis, M., Sengstag, T., Schutz, F., Goldstein, D., Piccart, M., and Delorenzi, M. (2008).

Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures.

Breast Cancer Research, 10(4):R65.

Part VII

Appendix

Gene Expression Profiling Technologies

- There exist several technologies to measure the expression of genes.
- Low throughput technologies such as RT-PCR, allow for measuring the expression of a few genes.
- High throughput technologies, such as microarrays, allows for measuring simultaneously the expression of thousands of genes (whole genome).

- Microarray principles will be illustrated through the Affymetrix technology.

Microarray

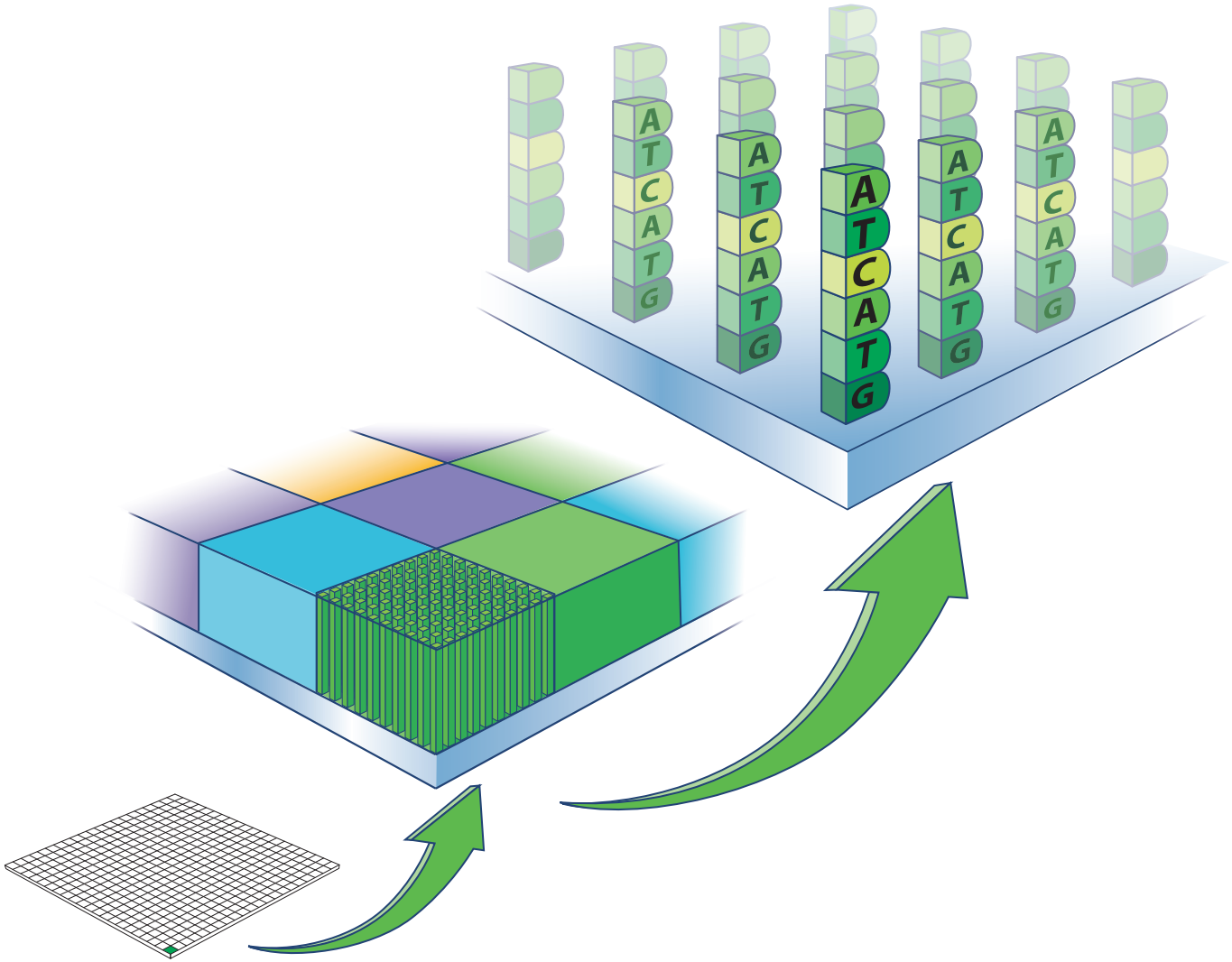
- A microarray is composed of
  - DNA fragments (*probes*) fixed on a solid support.
  - Ordered position of probes.
  - Principle of hybridization to a specific probe of complementary sequence.
  - Molecular labeling.

➠ Simultaneous detection of thousands of sequences in parallel.

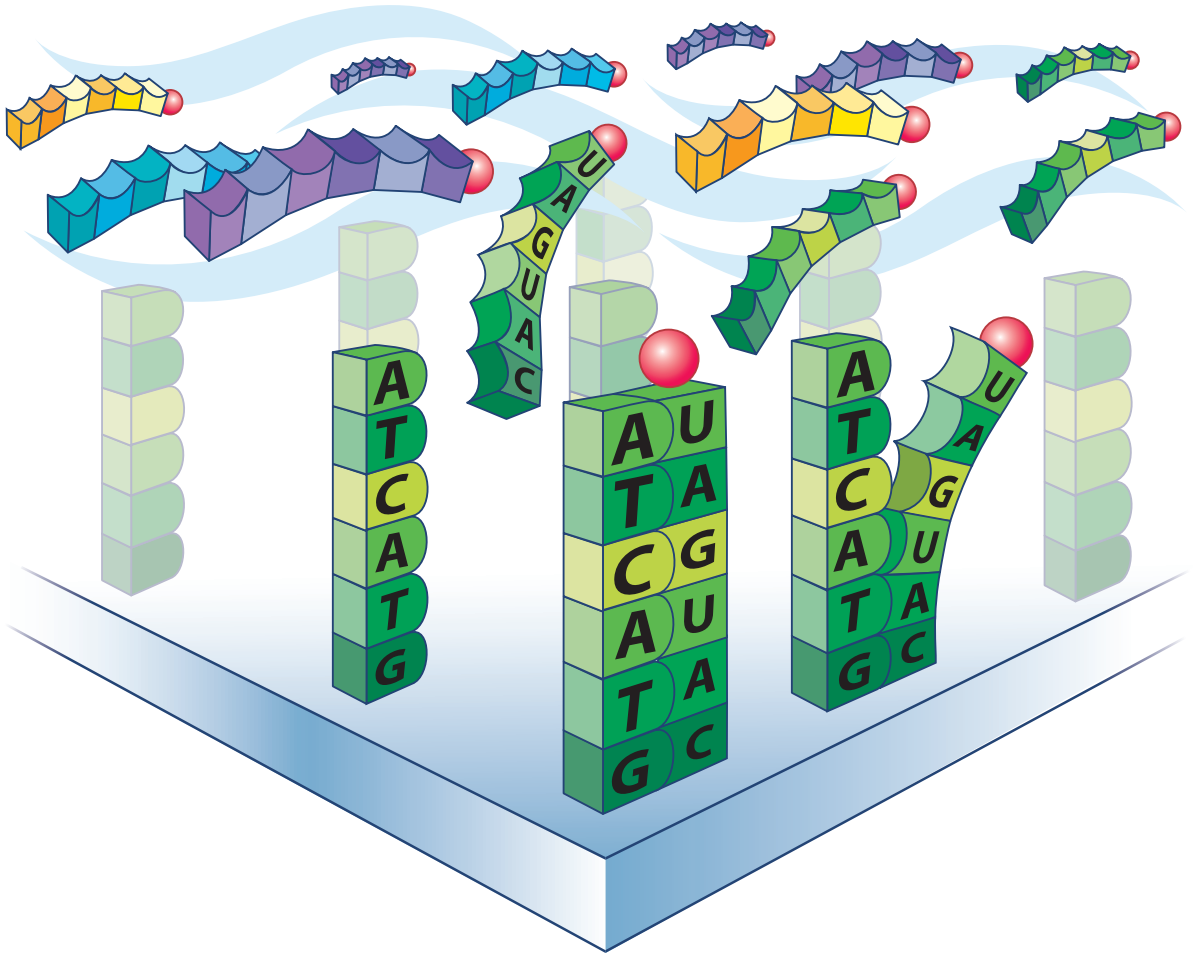
Affymetrix GeneChip

Affymetrix GeneChip

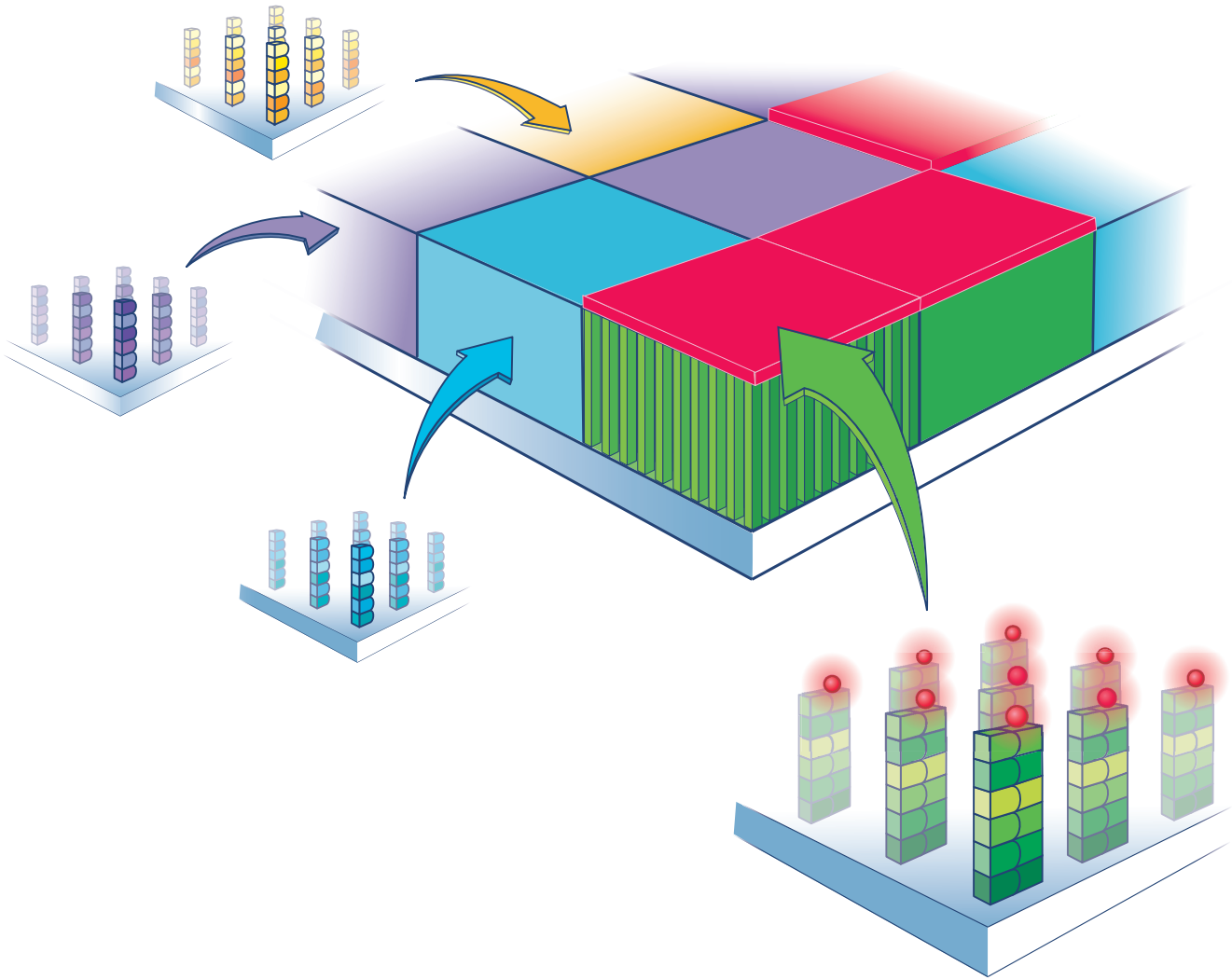
Probes



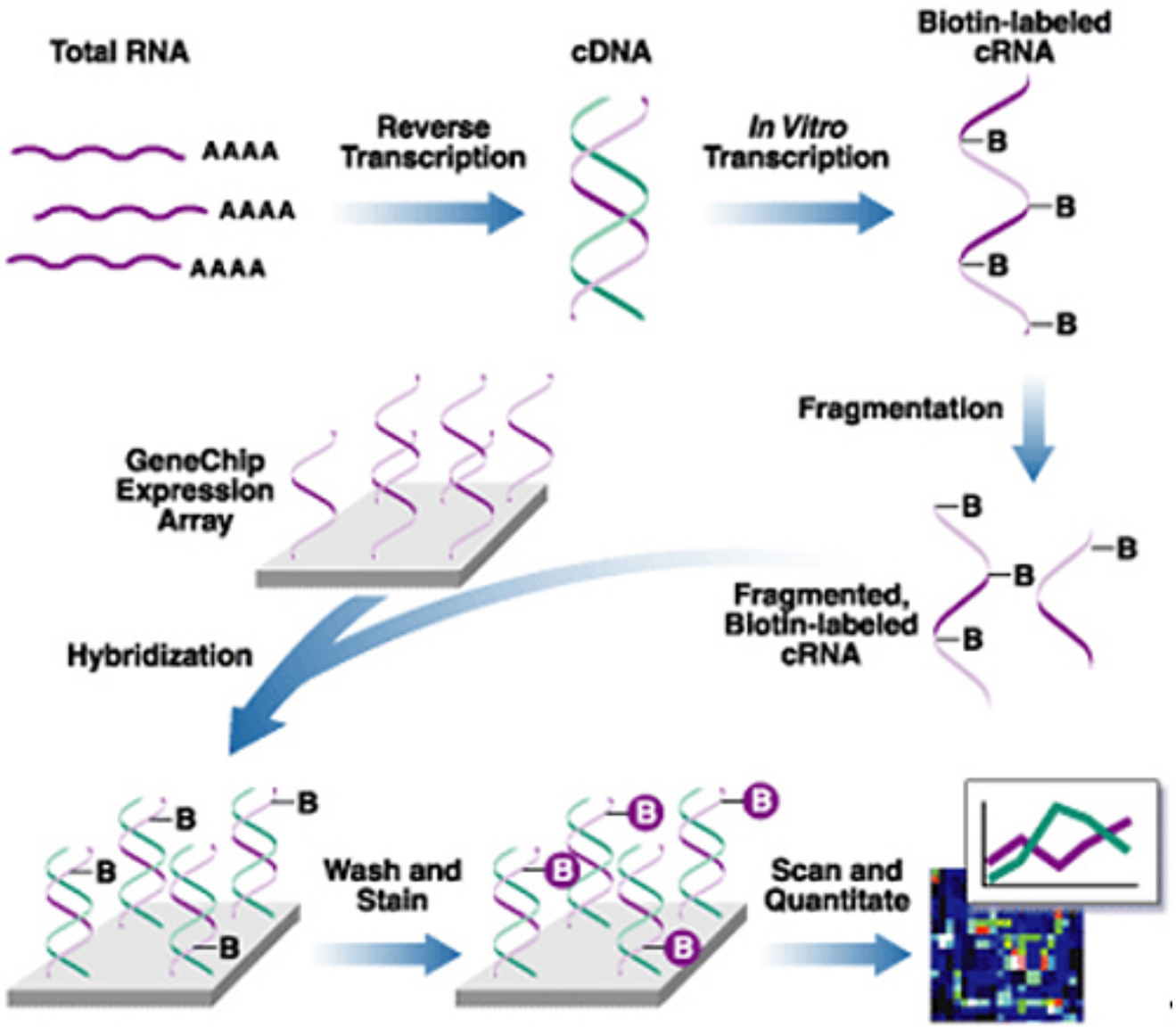
Hybridization



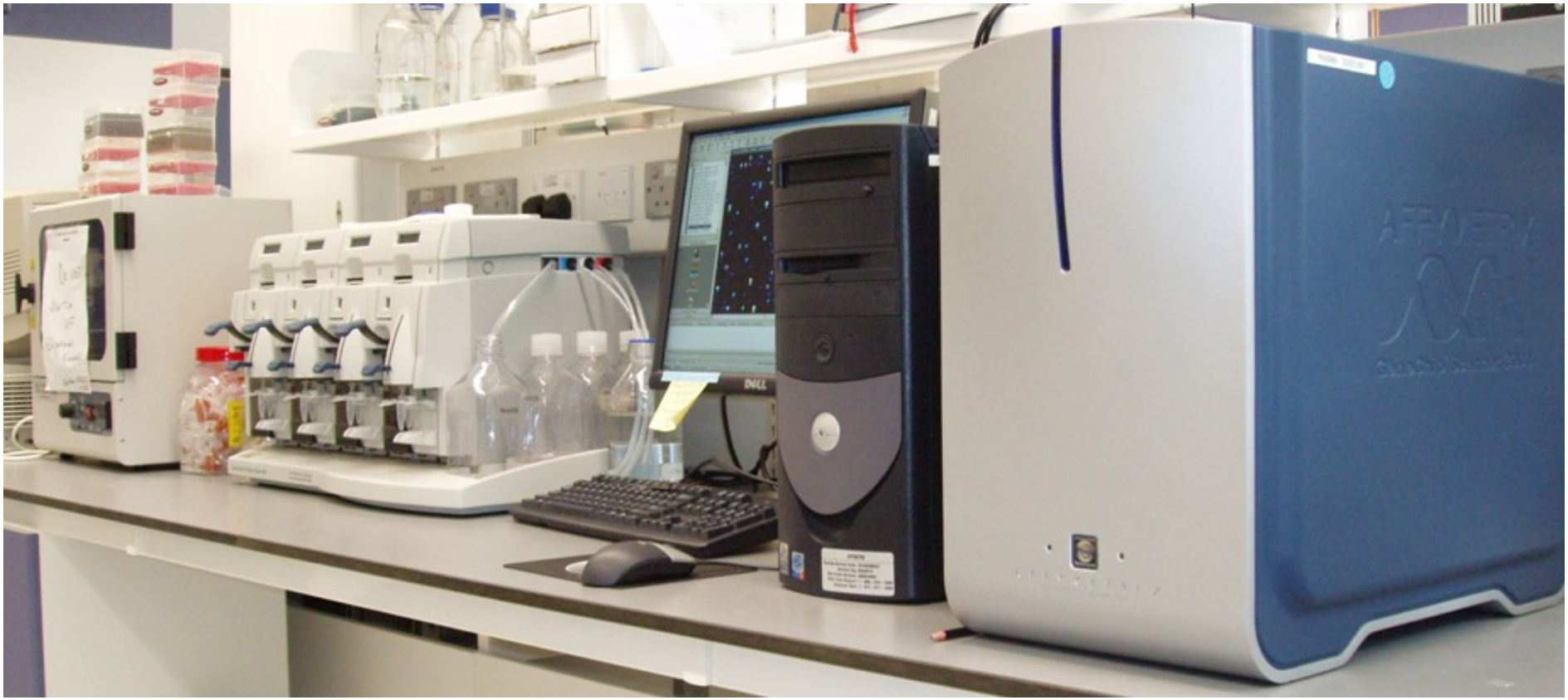
Detection



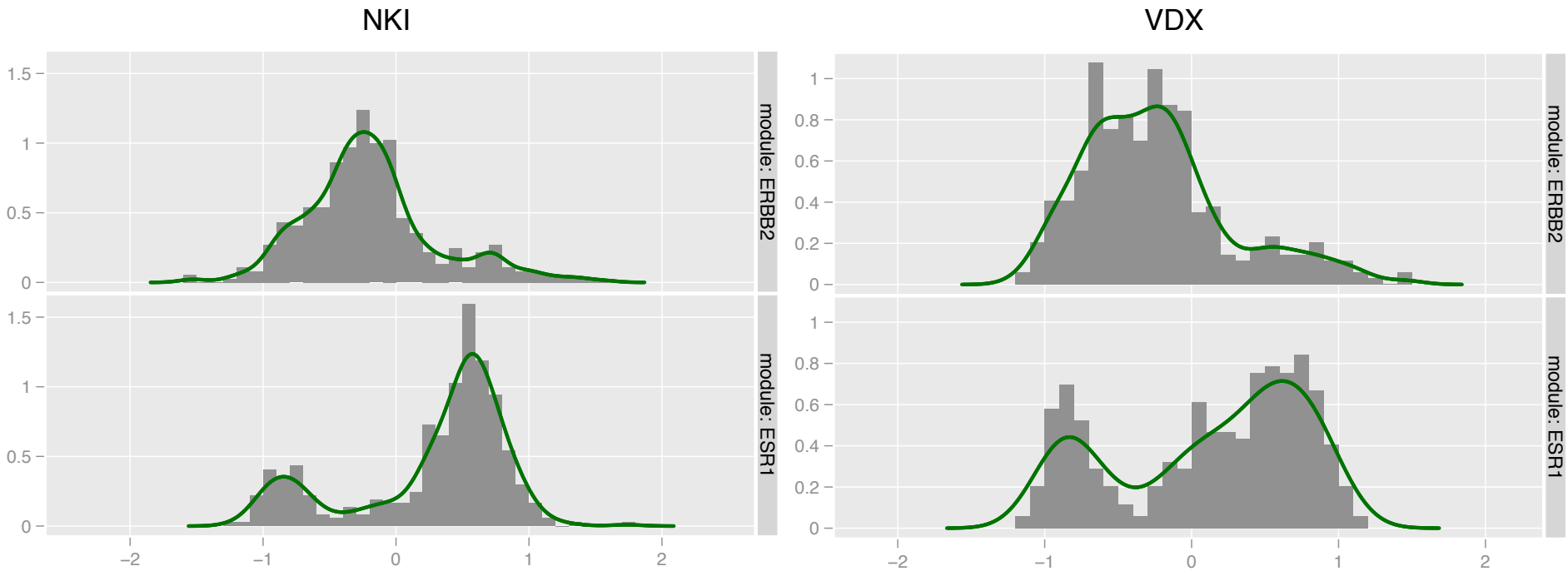
Affymetrix Design



Affymetrix Equipment



Mixture of Gaussians?



Tools

- Bioinformatics softwares
  - **R** is a widely used open source language and environment for statistical computing and graphics
  - **Bioconductor** is an open source and open development software project for the analysis and comprehension of genomic data
  - **Java Treeview** is an open source software for clustering visualization
  - **BRB Array Tools** is a software suite for microarray analysis working as an Excel macro

Links

- Personal webpage: http://www.ulb.ac.be/di/map/bhaibeka/
- Machine Learning Group: http://www.ulb.ac.be/di/mlg
- Functional Genomics Unit: http://www.bordet.be/en/services/medical/array/practical.htm
- Master in Bioinformatics at ULB and other belgian universities: http://www.bioinfomaster.ulb.ac.be/