Robust Identification of Breast Cancer Molecular Subtypes to Refine Prognosis

Benjamin Haibe-Kains[1,2]

[1]Functional Genomics Unit, Institut Jules Bordet

[2]Machine Learning Group, Université Libre de Bruxelles

Research Groups

Functional Genomics Unit - Jules Bordet Institute

- Head: Prof. Christos Sotiriou.
- 8 researchers (1 prof, 5 postDocs, 2 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
- International scientific collaborations : Genome Institute of Singapore, John Radcliffe Hospital, Karolinska Institute and Hospital, MD Anderson Cancer Center, Netherlands Cancer Institute, Swiss Institute of Bioinformatics, NCI/NIH, Gustave-Roussy Institute, IDDI.

Research Groups

Machine Learning Group - Université Libre de Bruxelles

- Head: Prof. Gianluca Bontempi.
- 10 researchers (2 prof, 4 postDoc, 4 PhD students), 4 graduate students.
- Research topics : Bioinformatics, Classification, Regression, Time series prediction, Sensor networks.
- Website : http://www.ulb.ac.be/di/mlg.
- Scientific collaborations inside ULB : IRIDIA (Sciences Appliquées), 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).

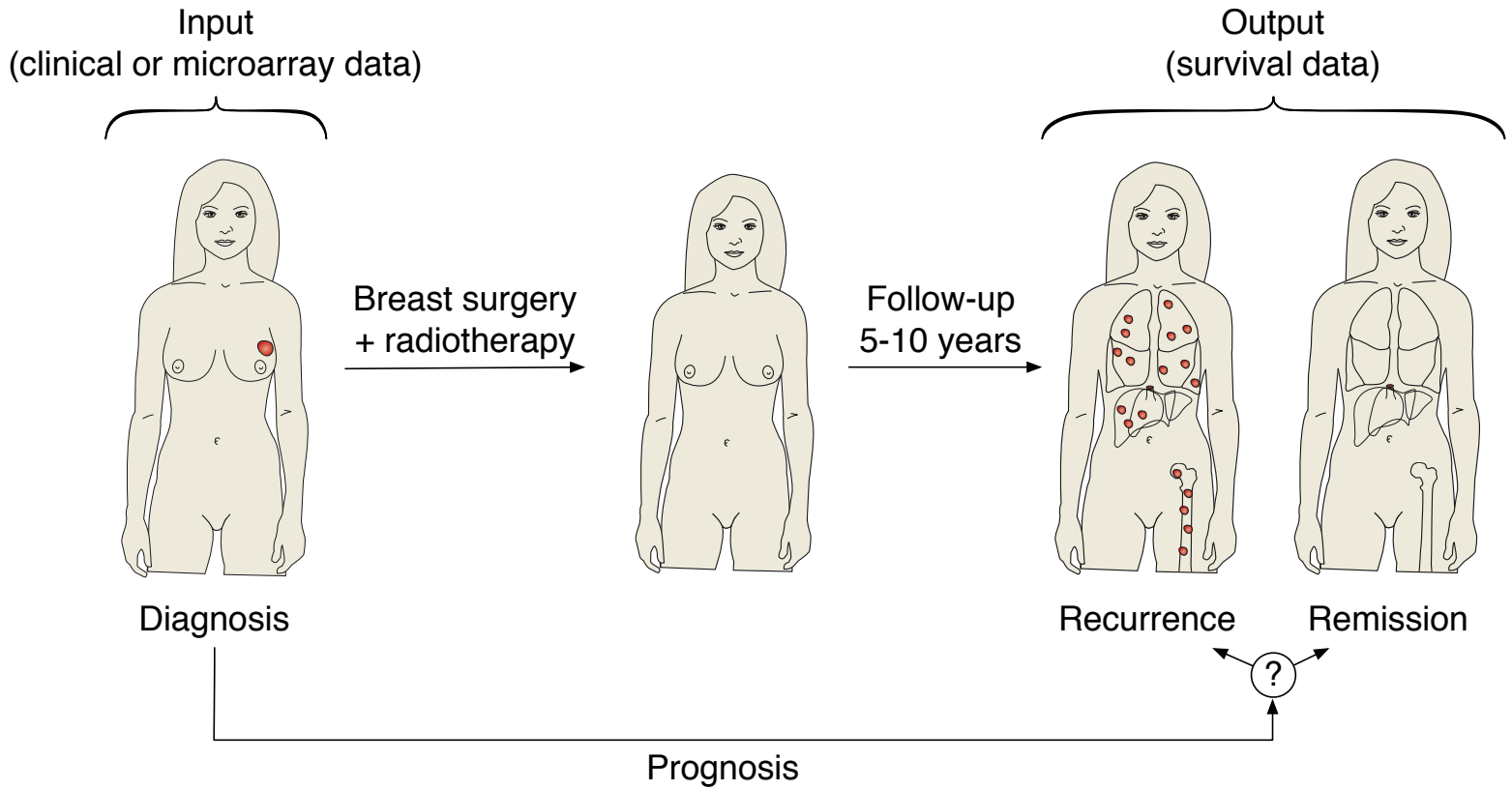
Outline

Introduction

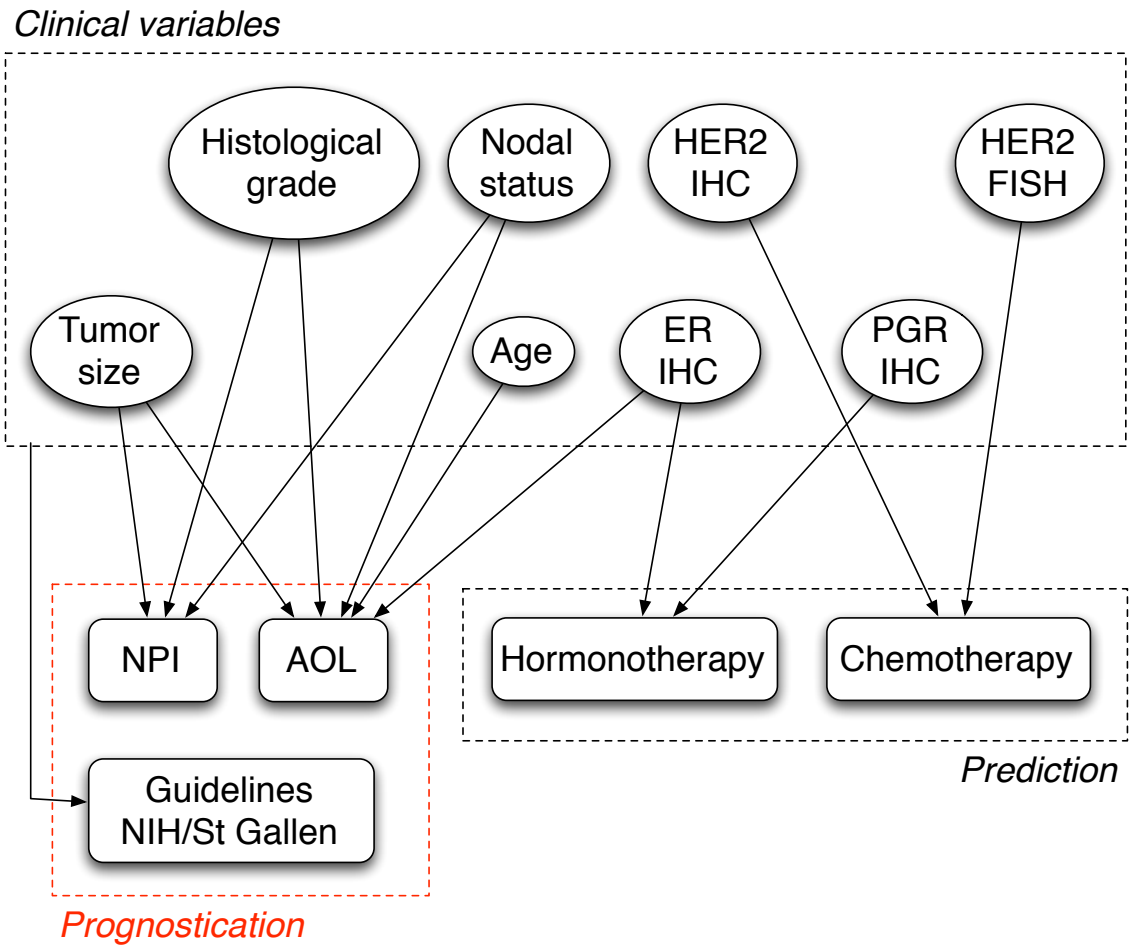
Breast Cancer

- Huge public health issue:
  - One of the most diagnosed malignancy in the Western World.
  - 1 out of 9 women will develop a breast cancer during her lifetime.

- Prognostication:



Traditional Approach for Prognostication

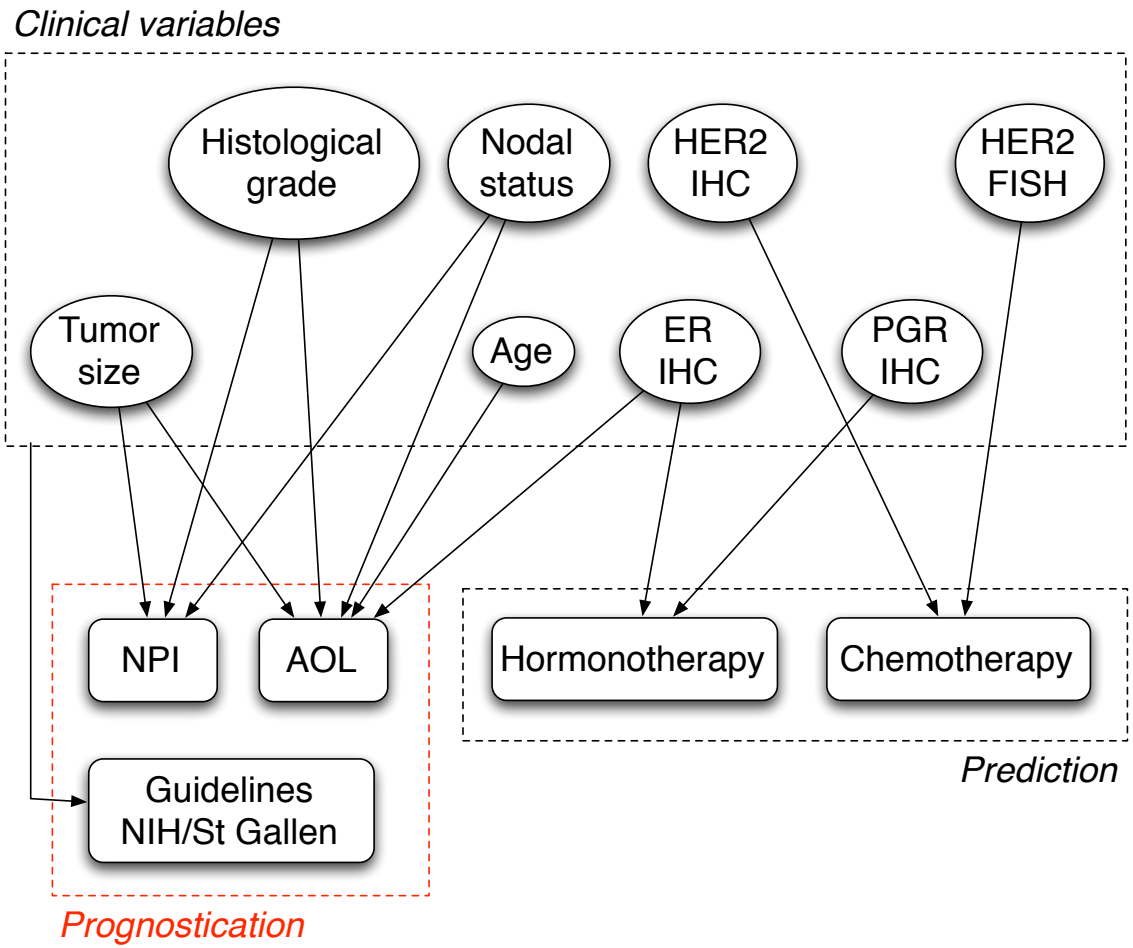


- NPI $= 0.2 \times$ tumor size $+$ grade $+$ nodal status
- AOL uses a life table technique considering age, nodal status, tumor size, and ER.

Problem: Prognostic clinical models predict numerous **low-risk** patients with early breast cancer (nodal status $= 0$) as **high-risk**.

➠ **Overtreatment**

Traditional Approach for Prognostication



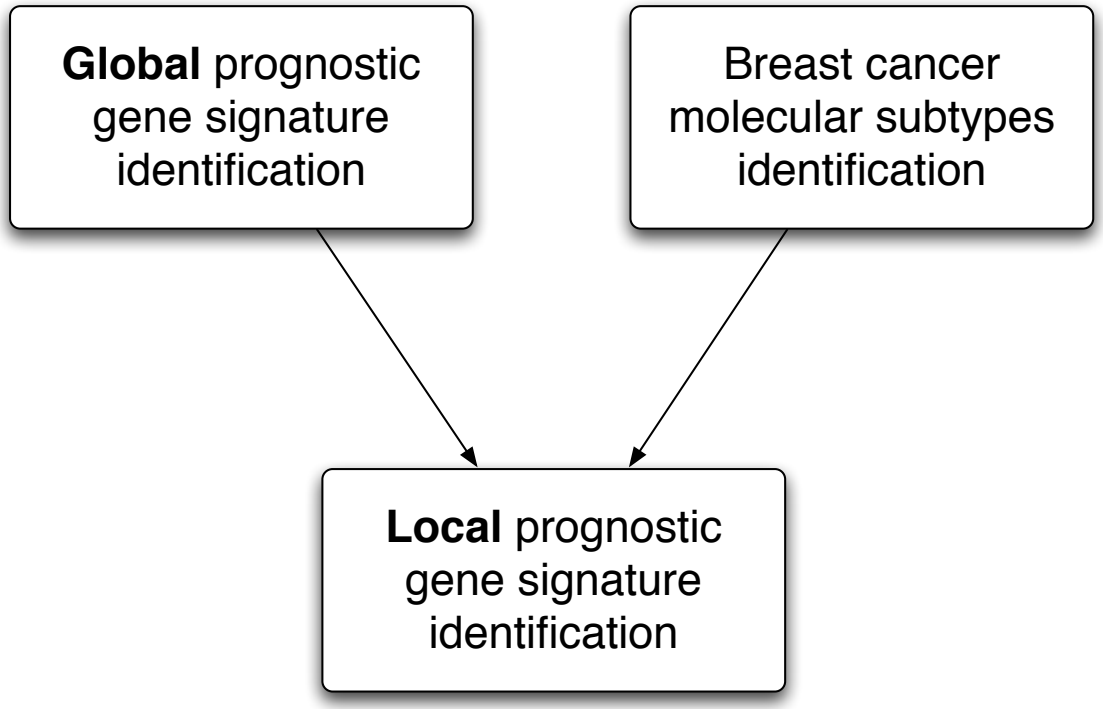
- NPI $= 0.2 \times$ tumor size $+$ grade $+$ nodal status
- AOL uses a life table technique considering age, nodal status, tumor size, and ER.

Problem: Prognostic clinical models predict numerous **low-risk** patients with early breast cancer (nodal status $= 0$) as **high-risk**.

➟ **Overtreatment**

Gene Expression Profiling Approach for Prognostication

- Improvement of breast cancer prognostication by using *machine learning* techniques to analyze microarray and survival data.
- Our methodology is composed of three main parts:



Global Prognostic Gene Signatures

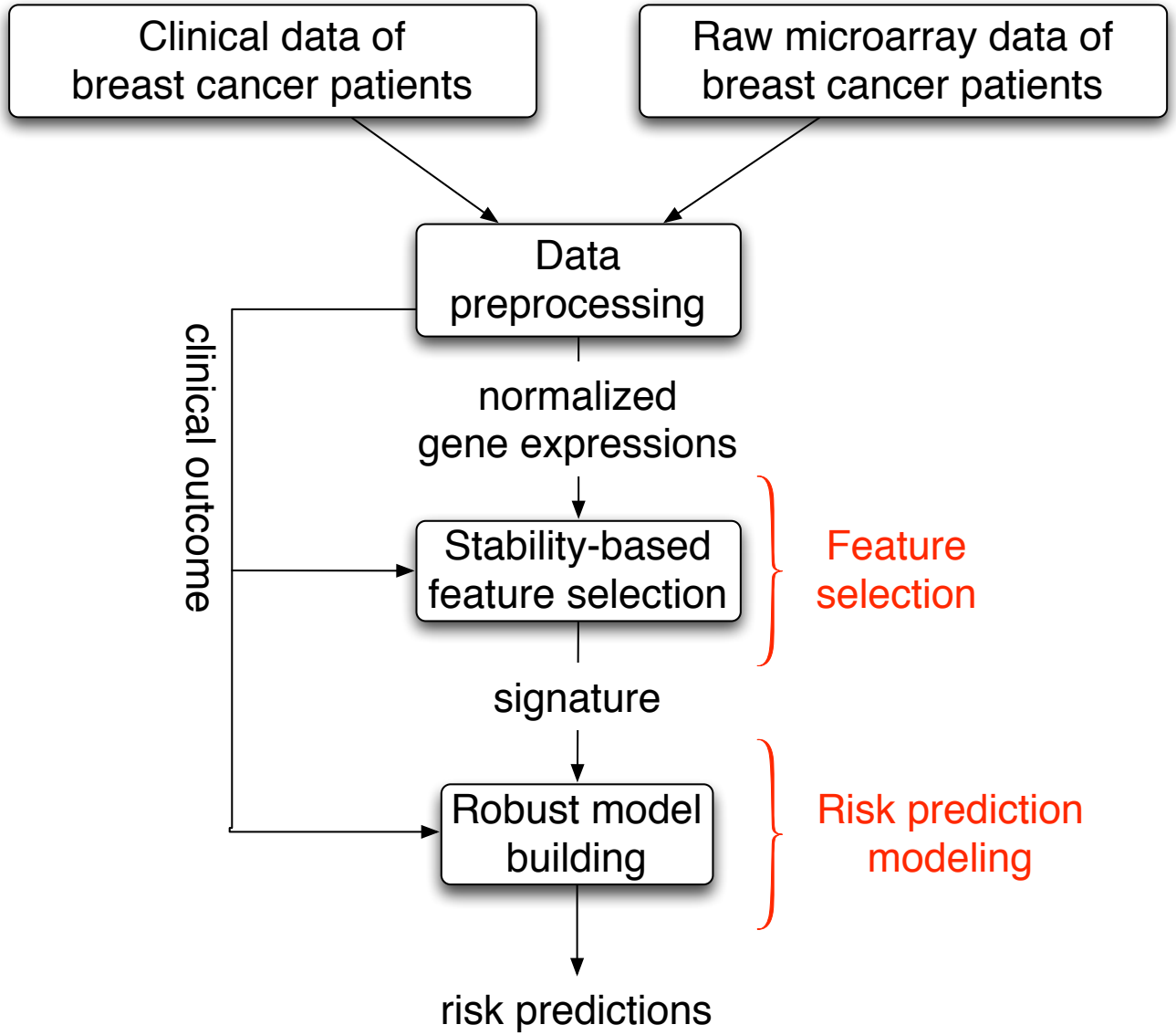
Global Prognostic Gene Signature Identification

- Objective: **Identification of prognostic gene signatures without taking into account the presence of molecular subtypes**.

- The idea is to identify prognostic gene signatures and their corresponding risk prediction models exhibiting the following characteristics:
  - Good performance with independent data.
  - Interpretable from a biological point of view.
  - Useable with data generated by different microarray platforms and/or normalization techniques.

Global Prognostic Gene Signature Identification

Methodology



Feature Selection

- Desirable properties of feature selection:
  - Computation efficiency (thousand of features).
  - Robustness to overfitting (not dataset dependent).
  - Intuitive tuning of (few) hyperparameters.

- For microarray data, **feature ranking** was shown to be efficient [Wessels et al., 2005].
- For prognostication, we used the **concordance index** [Harrell et al., 1996] as scoring function:

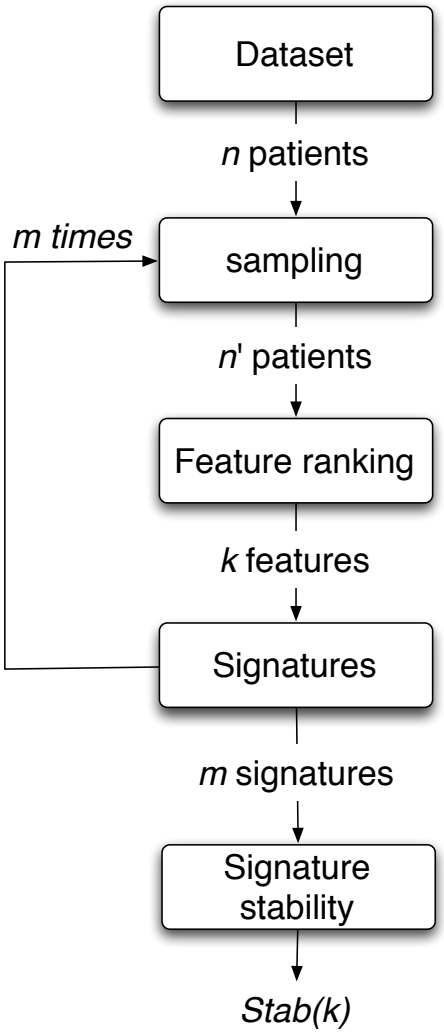
$$C\text{-index}(x_i, y) = \frac{\sum_{k,l \in \Omega} 1\{x_{ki} > x_{li}\}}{|\Omega|}$$

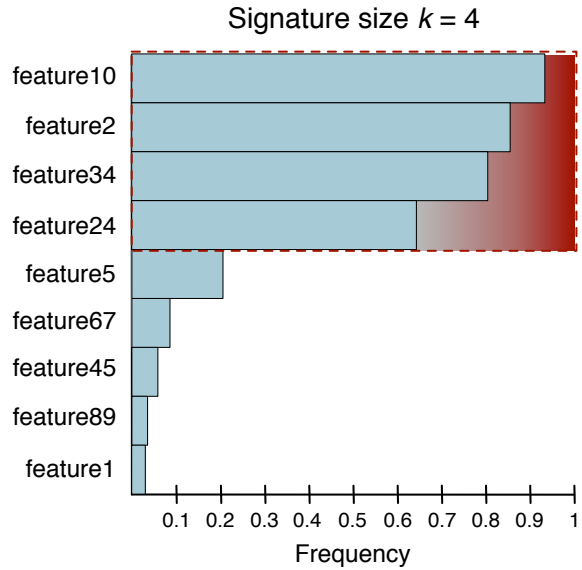
where x_i are the expressions of gene i, y are the survival data and Ω is the set of all the pairs of *comparable* patients $\{k, l\}$.

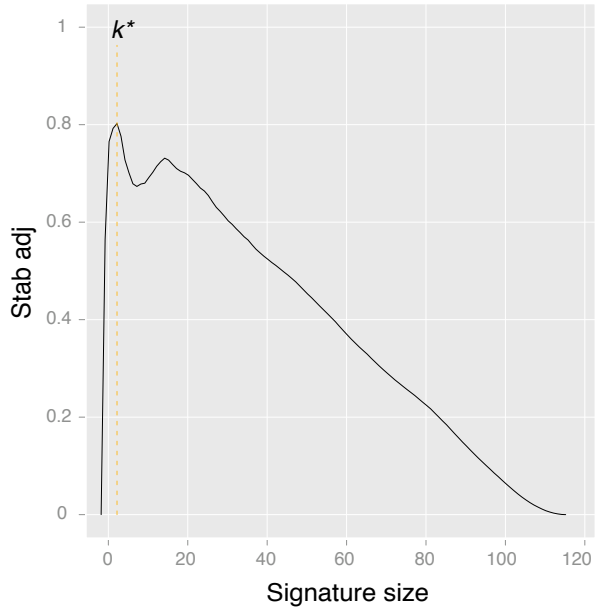
- Signature size?

Stability-Based Feature Ranking

Aim: To select a relevant and stable (low variance) set of features







Let X be the set of p features and $freq(x_j)$ be the number of sampling steps in which x_j has been selected out of m samplings without replacement. X is sorted into $\{x_{(1)}, \ldots, x_{(p)}\}$ where $freq(x_{(i)}) \geq freq(x_{(j)})$ if $i > j$.

$$Stab(k) = \frac{\sum_{l=1}^{k} freq(x_{(l)})}{k\, m} \qquad Stab_{adj}(k) = \max\left\{0, Stab(k) - \frac{k}{p}\right\}$$

$$\Rightarrow \quad k^* = \underset{k}{\operatorname{argmax}}\, Stab_{adj}(k)$$

Stability-Based Feature Ranking

Pros & Cons

- Pros
  - Computational scalability.
  - Reduction in risk of overfitting (low variance).
  - Easy and intuitive tuning of the hyperparameter.

- Cons
  - Potential risk of bias since feature ranking is unable to deal with **redundancy** and **complementarity** of the features [Meyer, 2008].

This method was first applied in a study related to tamoxifen resistance in breast cancer [Loi et al., 2008].

Risk Prediction Modeling

- Risk prediction models can be:
  - Linear or non-linear.
  - Univariate or multivariate.
- Intrinsic nature of microarray data evokes the risk of overfitting of non-linear multivariate prediction models.
- But univariate linear prediction models are not able to take into account for the multiple interactions underlying tumorigenesis.

➠ Trade-off?

- Since 2002, large comparative studies have been conducted to identify successful prediction models for class discovery and classification
- . . . leaving aside risk prediction (survival models) in breast cancer.

Comparative Study of Risk Prediction Models

- We compared numerous linear risk prediction models in the context of breast cancer prognostication using microarray and survival data [Haibe-Kains et al., 2008b].
  - "Complex" methods, although promising in the training set, yielded poor performance in independent sets (overfitting).
  - The loss of interpretability deriving from the use of overcomplex methods may be not sufficiently counterbalanced by significantly better performance.

- The few nonlinear risk prediction models (e.g. survival nnet [Eleuteri et al., 2003] and cSVM [van Belle et al., 2007]) developed in the field also yielded poor performance (data not shown).

Robust Model Building

- So we focussed on linear multivariate linear risk prediction models and relied on a simple additive combination scheme [Kittler et al., 1998] and *equal weights* linear regression [Wainer, 1976, Green, 1977]:

$$r = \sum_{i=1}^{k} \beta_i' x_i$$

where r is the *risk* of a patient, x_i are the gene expressions, and

$$\beta_i' = \begin{cases} -\frac{1}{k} & \text{if } x_i \text{ is positively correlated with survival} \\ +\frac{1}{k} & \text{otherwise} \end{cases}$$

- The use of equal weights (coefficients)
  - reduces the risk of overfitting,
  - ensures the coefficients estimation to not be influenced by outliers,
  - and causes only modest expected loss in accuracy.

Robust Model Buidling

Pros & Cons

- Pros
  - Additive models are multivariate models with low variance.
  - Less sensitive to redundancy that traditional multivariate models (collinearity).
  - Low computational cost (especially in combination with feature ranking).
  - Equal weights regression (i.e. signed average in our case) facilitates the computation of risk predictions in different microarray platforms with missing probes.

- Cons
  - Unable to deal with complementarity of features.

This method was used in the study related to **GGI** = **G**ene expression **G**rade **I**ndex [Sotiriou et al., 2006].

Breast Cancer Molecular Subtypes

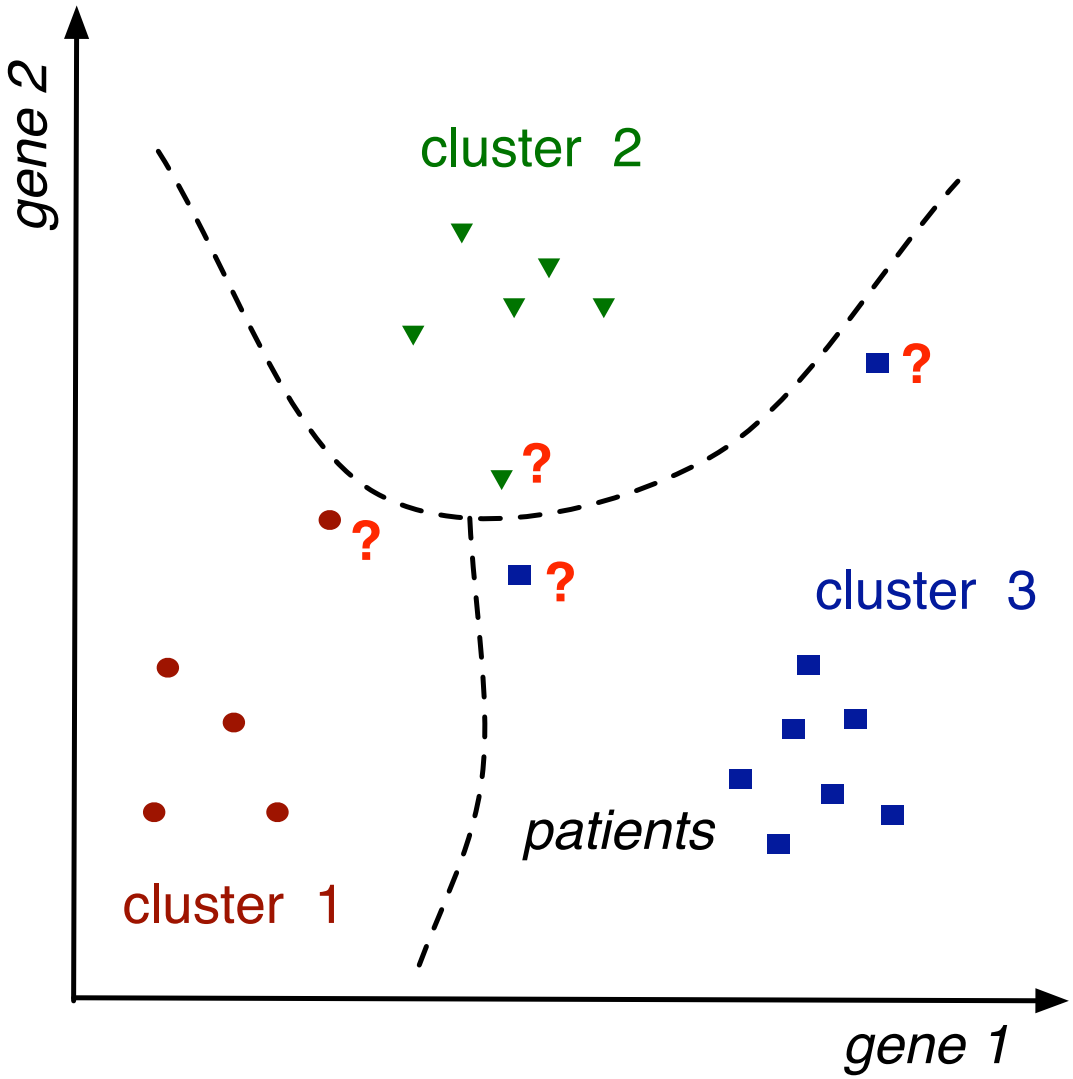
Molecular Subtypes Identification

- Objective: **Robust identification of breast cancer molecular subtypes along with estimation of classification uncertainty**.

- The idea is to extent the study of Perou *et al.* by building a clustering model exhibiting the following characteristics:
  - Estimation of classification uncertainty.
  - Good performance in independent dataset.
  - Useable with data generated by different microarray platforms and/or normalization techniques.

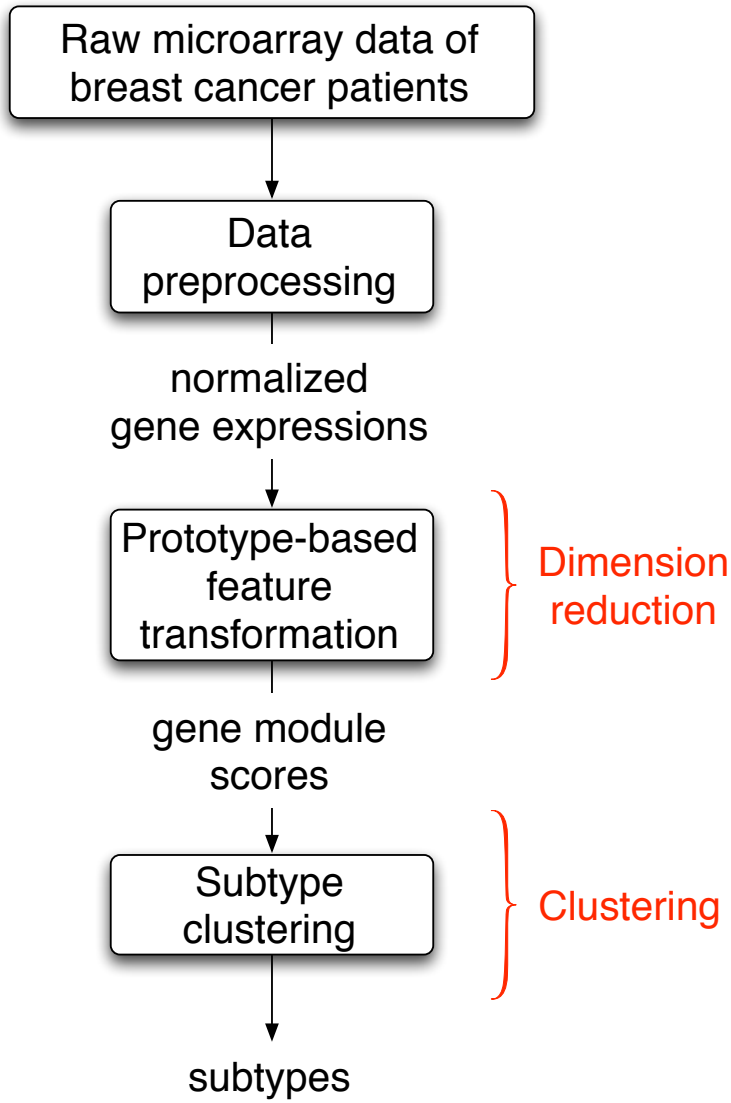
Molecular Subtypes Identification

Classification uncertainty



Molecular Subtypes Identification

Methodology



Dimension Reduction

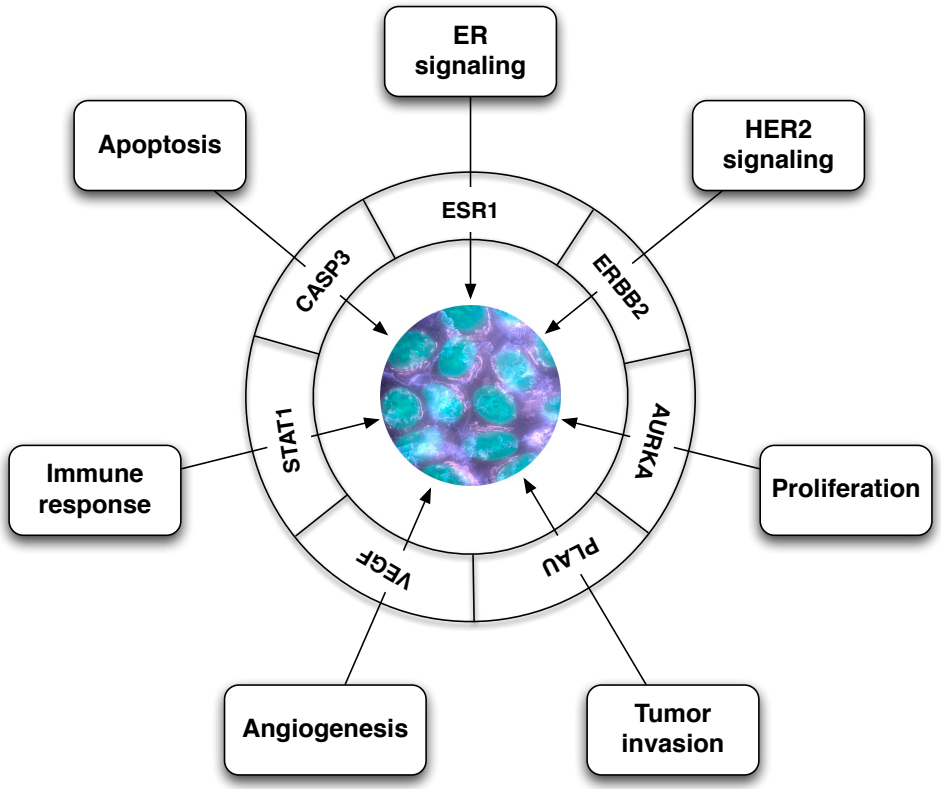
- Aim: To reduce the dimensionality of microarray data while keeping information relevant for subtypes identification.

- We used a prototype-based feature transformation [Desmedt et al., 2008, Haibe-Kains, 2009] to identify sets of genes **specifically** co-expressed to key biological processes in breast cancer.
- The method is composed of 2 main steps:
  1. Prototype-based clustering: Selection of prototypes and identification of *gene modules*.
  2. Summarization: Each cluster is summarized by a single feature called *gene module score* (weighted average of genes in cluster).

Gene Modules

Results

- We applied this method with 7 prototypes in two large datasets (NKI & VDX) [Desmedt et al., 2008].



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

Subtype Clustering

- Aim: To develop of a robust clustering model able to estimate classification uncertainty.

- Low dimensional input space defined by **ESR1** and **ERBB2** module scores (see [Kapp et al., 2006]).
- We developed a model-based clustering:
  - Probability to belong to subtype j:

$$\Pr(j|x_i) = \frac{\pi_j \mathcal{N}(x_i; \mu_j, \Sigma_j)}{\sum_{j=1}^{m} \pi_j \mathcal{N}(x_i; \mu_j, \Sigma_j)}$$

    where m is the number of Gaussians, π_j is the prior probability of x_i to be generated by the j^{th} Gaussian $\mathcal{N}(x_i; \mu_j, \Sigma_j)$
  - Hyperparameters:
    - ★ Number of clusters (Gaussians): BIC [Schwarz, 1978].
    - ★ Mean μ_j and covariance matrices Σ_j: EM algorithm [Dempster et al., 1977].

Subtype Clustering

Pros & Cons

- Pros
  - The low dimensional space increases the model robustness (low feature-to-sample ratio).
  - The low dimensional space facilitates the representation of the results and their interpretation.
  - Estimation of classification uncertainty + ability to perform *soft partitioning* of the data.
  - Easy to use this clustering model to predict the subtype of the tumor of a new patient (unlike hierarchical clustering).
- Cons
  - Although the use of ESR1 and ERBB2 module scores is beneficial (see above), it prevents us to find any new subtypes.

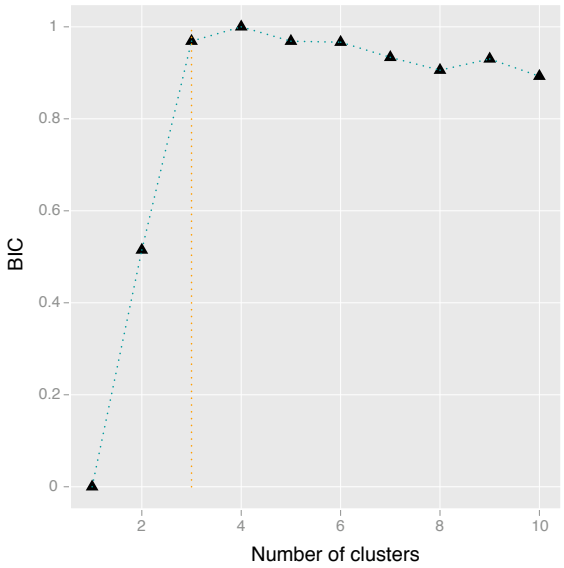
Subtype Clustering

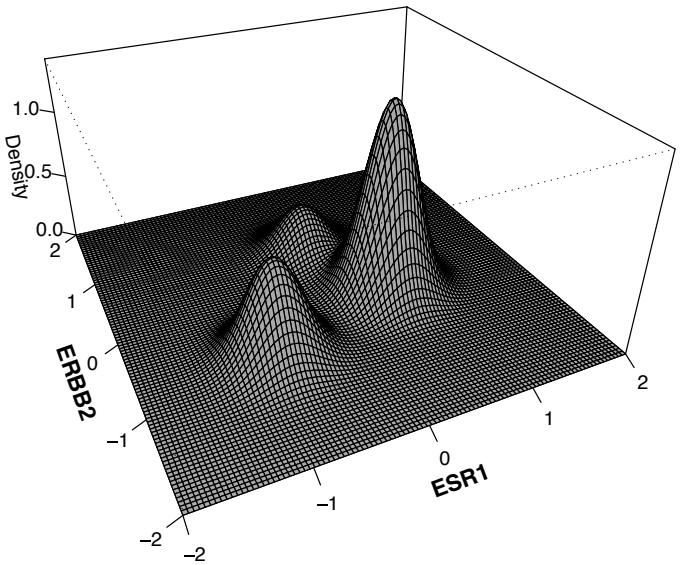
Results

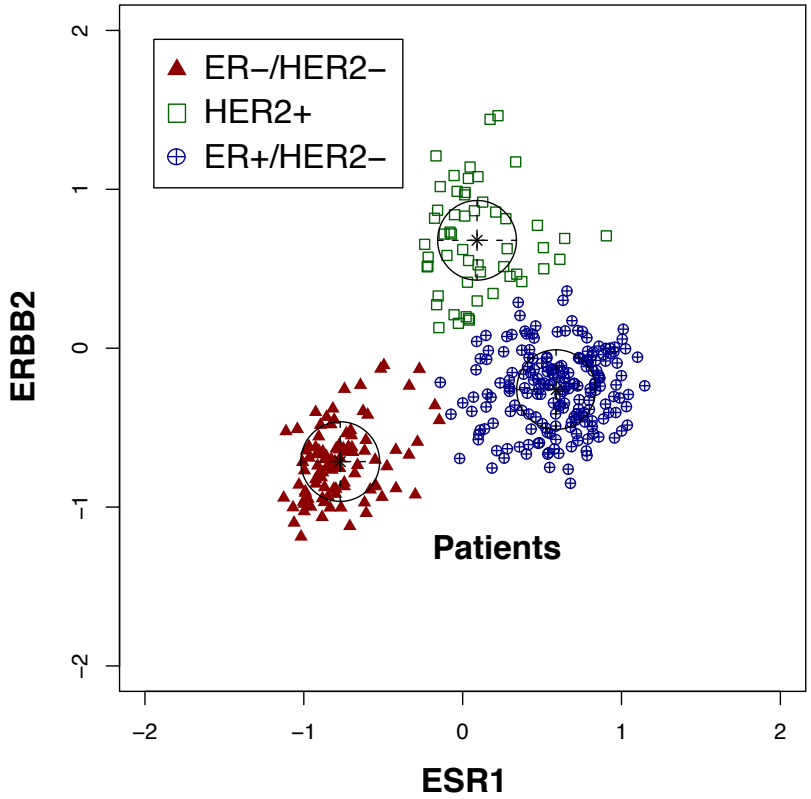
- Training set: 344 patients with early breast cancer (VDX).
- Validation set:
  - Clustering: 17 independent datasets ($>$ 3400 patients).
  - Prognosis: 745 untreated patients with early breast cancer from 5 datasets (NKI, TBG, UPP, UNT and MAINZ).

Subtype Clustering Model

Training set (VDX)

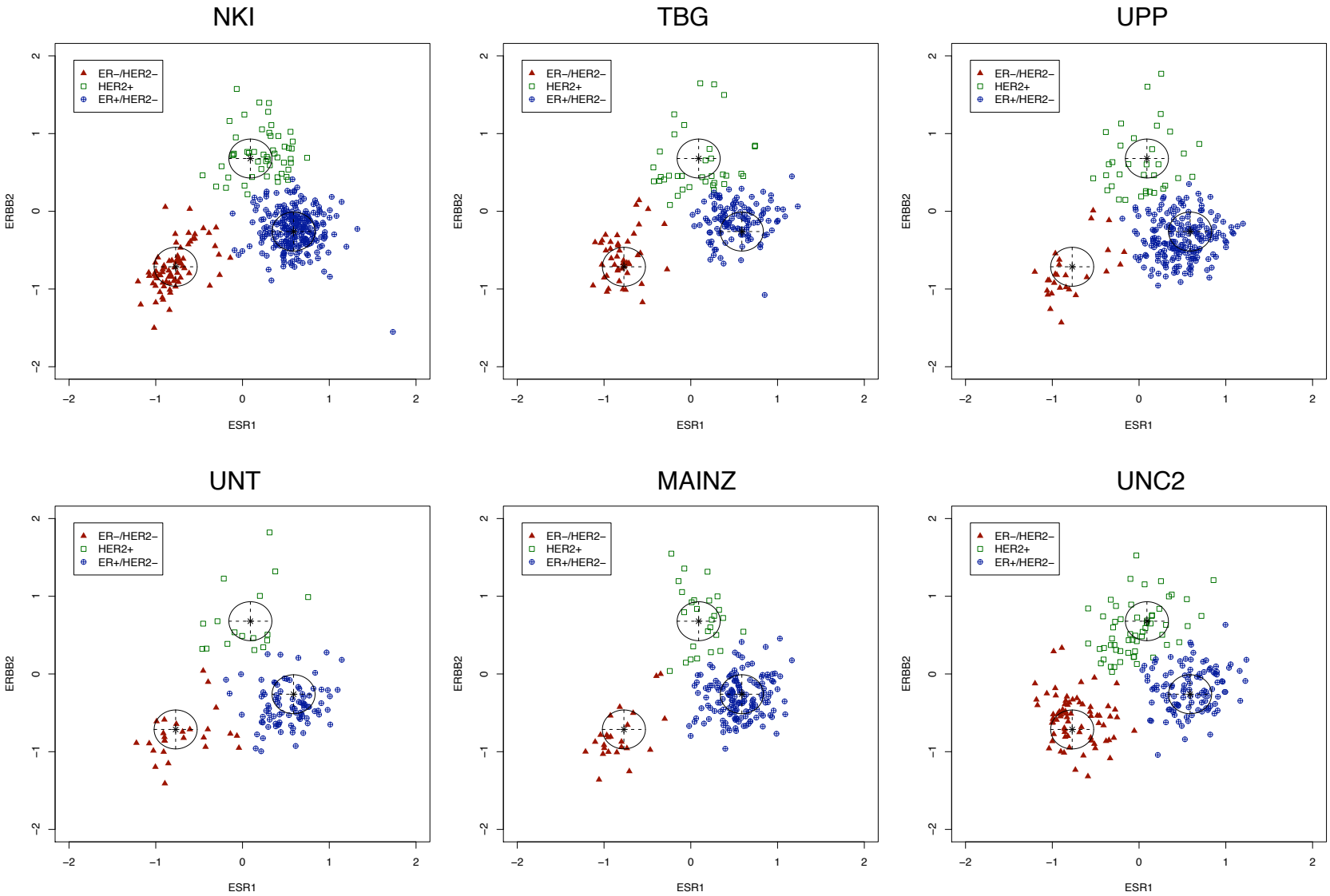






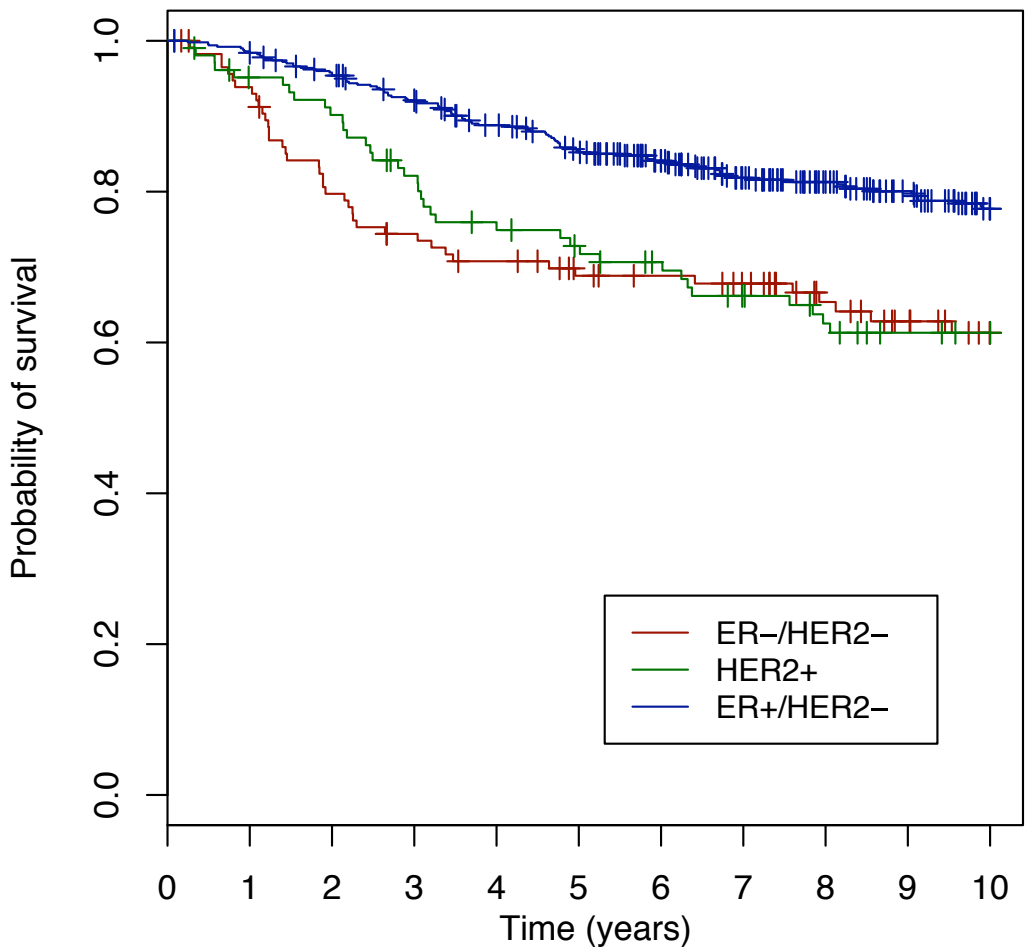
Subtype Clustering Model

Independent datasets



Subtype Clustering

Prognosis of early breast cancer patients



No. At Risk

ER–/HER2–	116	108	91	83	78	71	68	64	53	46	37
HER2+	105	97	91	81	73	69	64	58	52	47	44
ER+/HER2–	503	494	474	450	425	401	355	312	276	251	217

Local Prognostic Gene Signatures

Local Prognostic Gene Signature Identification

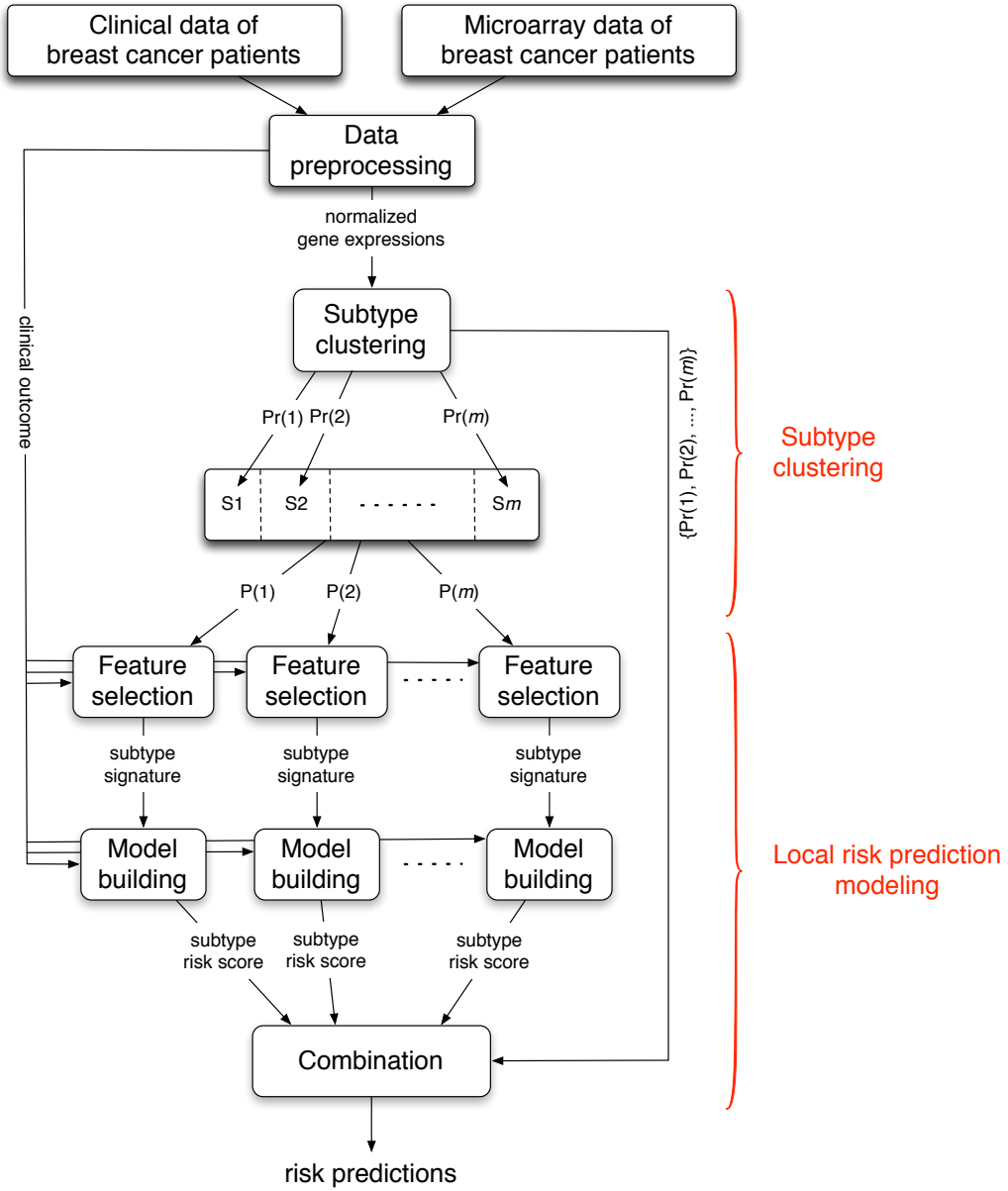
- Objective: **Identification of prognostic gene signatures taking into account the presence of molecular subtypes**.

- [Wang et al., 2005] conducted a study similar to [van't Veer et al., 2002] but they took into account the molecular heterogeneity of breast cancer wrt ER status (first *local signatures*).
- ➟ Prognostic gene signature: GENE76.

- Open issues:
  - The local models are developed for two subgroups (ER- and ER+) only, without considering the heterogeneity of the HER2+ subgroup.
  - Hard partitioning of the population of breast cancer patients.
  - The local model for ER- subgroup was trained on few samples and yielded poor performance in validation studies [Foekens et al., 2006, Desmedt et al., 2007].

Local Prognostic Gene Signature Identification

Methodology



Local Model Network

Aim: To develop a prognostic model taking into account breast cancer molecular subtypes

- We adopted a *divide-and-conquer* strategy, through a *modular modeling* approach, and developed a Local Model Network [Johansen and Foss, 1993] for prognostication:

$$r = \sum_{j=1}^{m} \rho_j(x, \theta_j) h_j(x, \alpha_j)$$

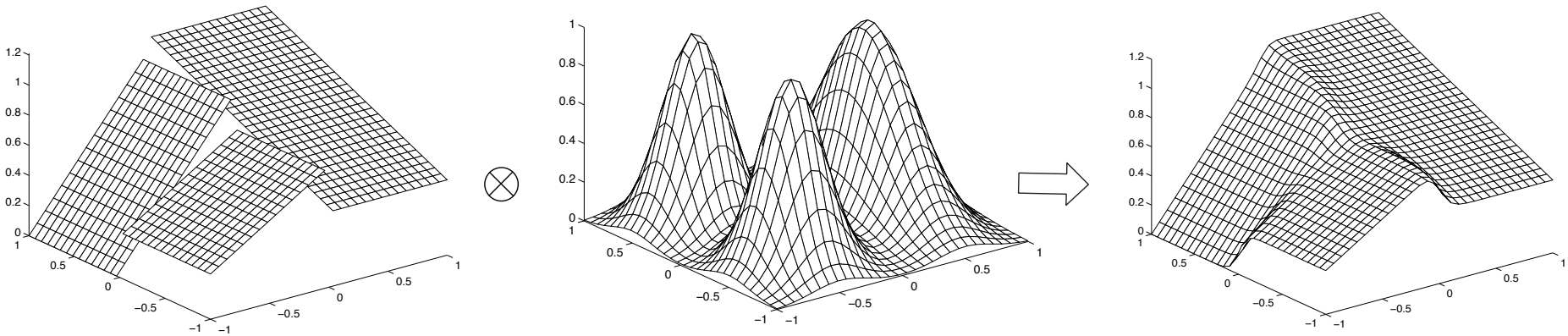
where x and r stand for the gene expressions and patients' risk respectively, $\rho_j(x, \theta_j)$ is the j^{th} basis function of parameters θ_j, $h_j(x, \alpha_j)$ is the j^{th} local risk prediction model of parameters α_j.

- The ρ_j are constrained to satisfy

$$\sum_{j=1}^{m} \rho_j(x, \theta_j) = 1$$

Local Model Network

- Example of LMN:



- The j^{th} basis function is defined as the probability to belong to the j^{th} breast cancer molecular subtype (*subtype clustering*):

$$\rho_j(x) = \frac{\pi_j \mathcal{N}(x; \mu_j, \Sigma_j)}{\sum_{j=1}^{m} \pi_j \mathcal{N}(x; \mu_j, \Sigma_j)}$$

- The j^{th} local model is defined by our robust model:

$$h_j(x) = \sum_{i=1}^{k(j)} \beta_i' x_i$$

where signature size k now depends on subtype j (*local feature selection*).

Local Feature Selection

- In order to take advantage of the subtypes identification, we adapted the stability-based feature ranking to identify the genes prognostic in specific subtypes.

- We introduced a **weighted version** of the concordance index:

$$C_{wted}(x_i, y, \rho_j) = \frac{\sum_{k,l \in \Omega} w_{kl} \, 1\{x_{ki} > x_{li}\}}{\sum_{k,l \in \Omega} w_{kl}}$$

where x_i are the expressions of gene i, y are the survival data and $w_{kl} = \rho_j(x_k)\rho_j(x_l)$ is the weight for the pair of *comparable* patients $\{k, l\}$ with $\rho_j(x)$ being the probability for a patient's tumor x to belong to the subtype j.

Local Model Network

Pros & Cons

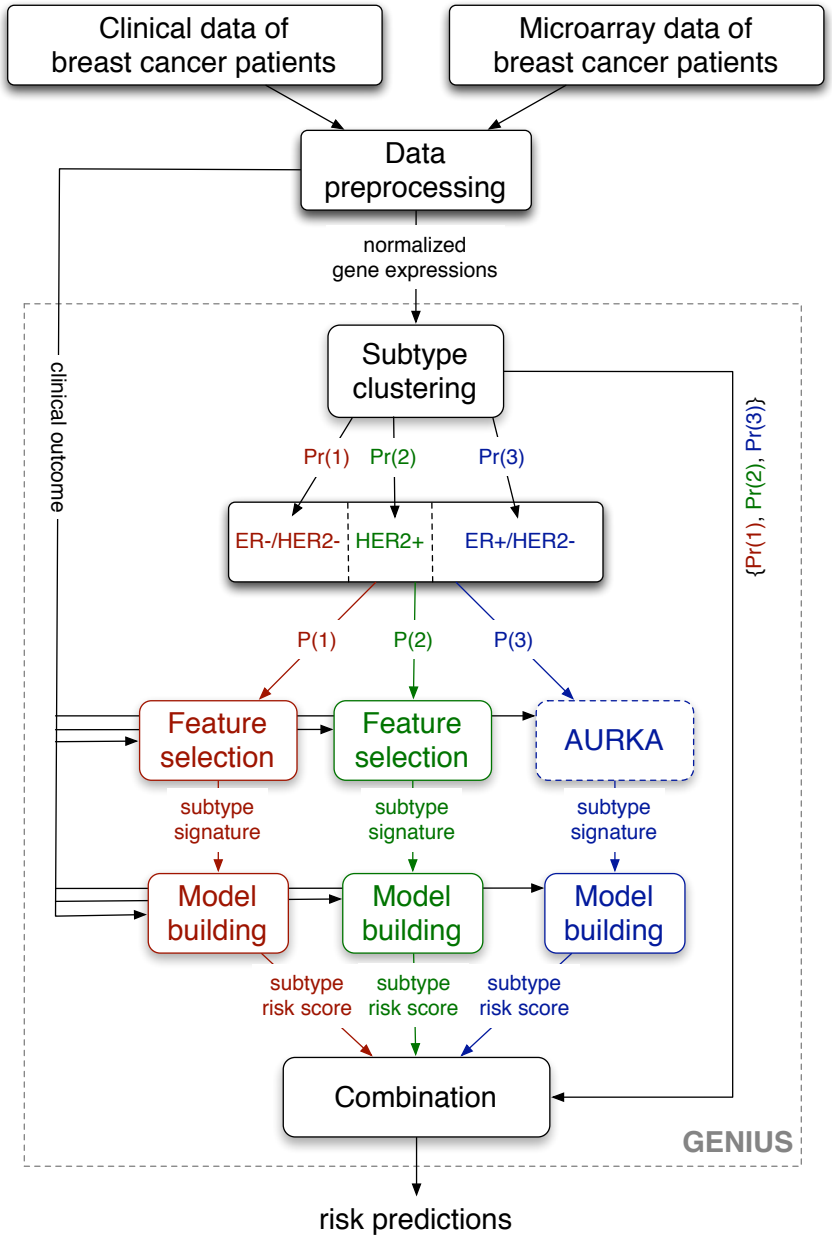
- Pros
  - Use of robust linear risk prediction models to perform nonlinear modeling.
  - Easy interpretation for doctors (breast cancer molecular subtypes + corresponding risk predictions).
  - Potentially more biological insights since the molecular heterogeneity of breast cancer is taken into account.

Local Prognostic Gene Signature Identification

GENIUS

- **GENIUS** = **G**ene **E**xpression prog**N**ostic **I**ndex **U**sing **S**ubtypes.
  - Aim: Refining breast cancer prognosis according to molecular subtypes.
- GENIUS refers to:
  - The local prognostic gene signatures identified through local stability-based feature ranking.
  - The risk prediction model using robust model building and Local Model Network.
- Training set:
  - Input data: 22,283 (1050 after filtering) gene expressions of 344 untreated patients with early breast cancer (VDX).
  - Output data: Survival data.
- Validation set: 745 untreated patients with early breast cancer from 5 datasets (NKI, TBG, UPP, UNT and MAINZ).

GENIUS


Clinical data of breast cancer patients
Microarray data of breast cancer patients
Data preprocessing
normalized gene expressions
clinical outcome
Subtype clustering
{Pr(1), Pr(2), Pr(3)}
Pr(1)
Pr(2)
Pr(3)
ER-/HER2-
HER2+
ER+/HER2-
P(1)
P(2)
P(3)
Feature selection
Feature selection
AURKA
subtype signature
subtype signature
subtype signature
Model building
Model building
Model building
subtype risk score
subtype risk score
subtype risk score
Combination
GENIUS
risk predictions

GENIUS

Results

- Local prognostic gene signatures identified for the ER-/HER2- and HER2+ subtypes involved different genes but are enriched in genes related to immune response.
- GENIUS yielded significantly better performance than all the state-of-the-art prognostic gene signatures in the global population. GENIUS was not significantly superior in all subtypes but it was the only signature that performed well whatever the subtype.
- GENIUS significantly outperformed prognostic clinical models in the global population. Its superiority almost reached significance in all the subtypes.

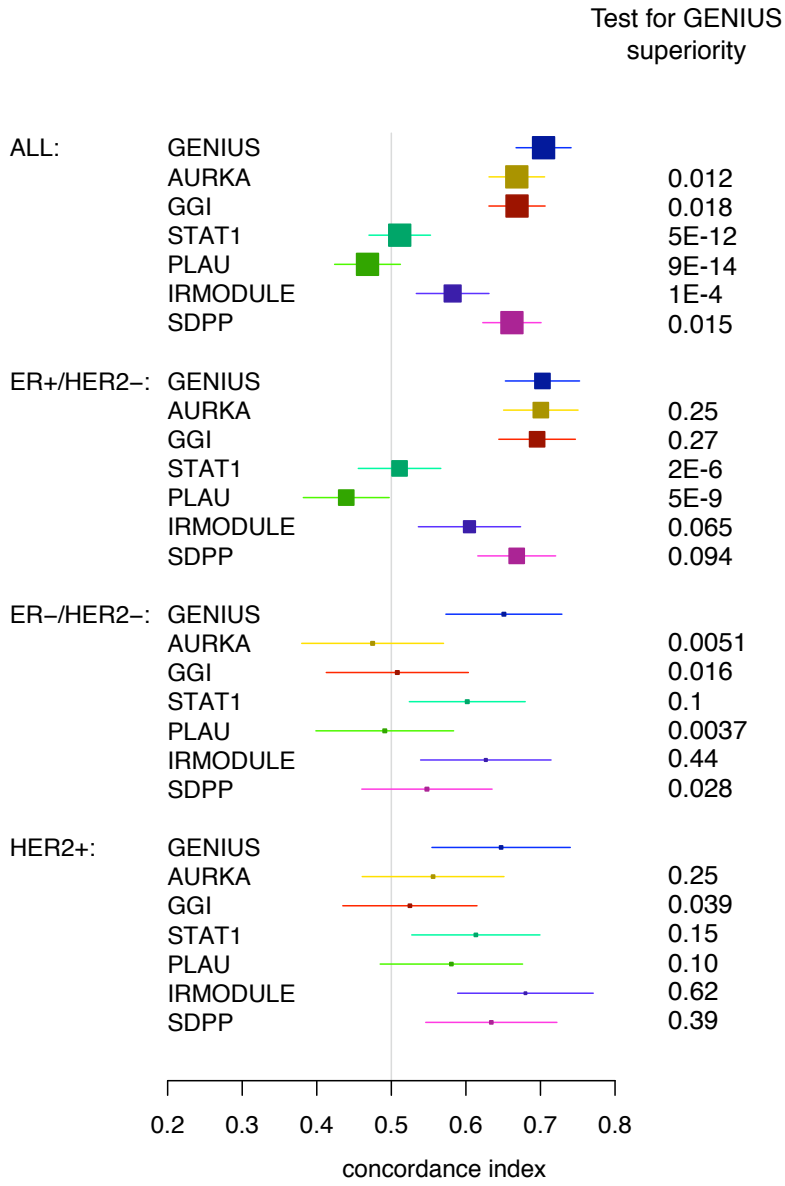
GENIUS

Results

- Local prognostic gene signatures identified for the ER-/HER2- and HER2+ subtypes involved different genes but are enriched in genes related to immune response.
- GENIUS yielded significantly better performance than all the state-of-the-art prognostic gene signatures in the global population. GENIUS was not significantly superior in all subtypes but it was the only signature that performed well whatever the subtype.
- GENIUS significantly outperformed prognostic clinical models in the global population. Its superiority almost reached significance in all the subtypes.

GENIUS

Performance comparison with prognostic gene signatures



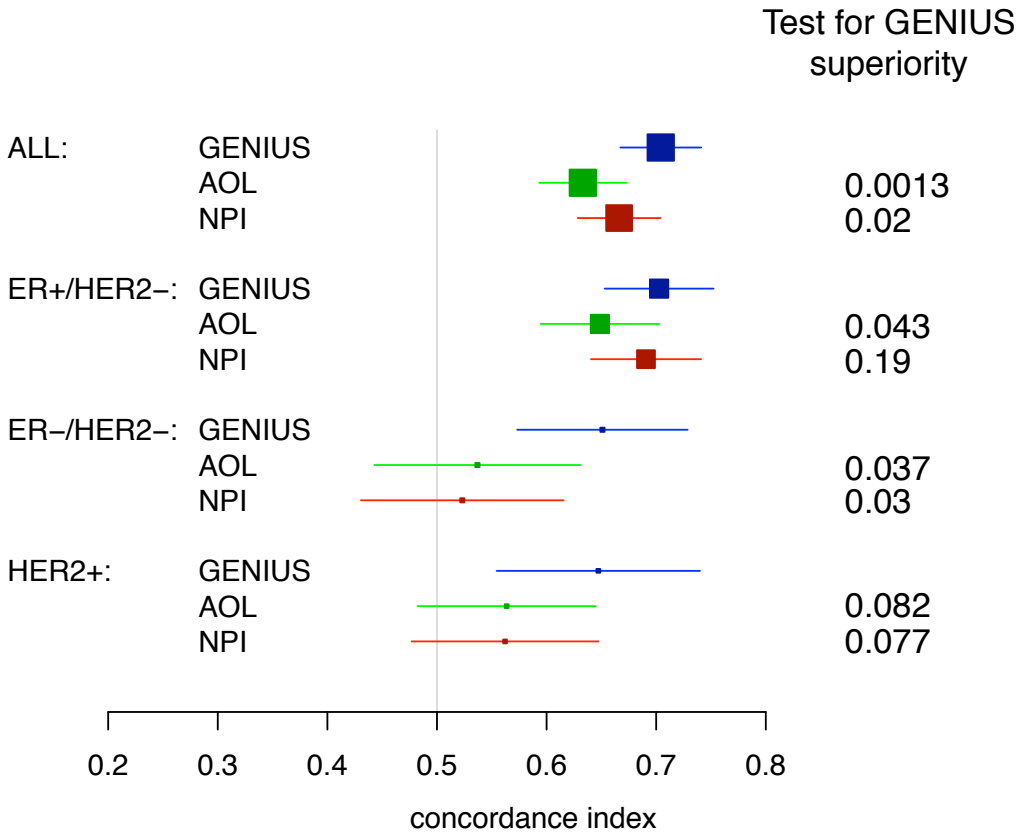
GENIUS

Results

- Local prognostic gene signatures identified for the ER-/HER2- and HER2+ subtypes involved different genes but shared genes related to immune response.
- GENIUS yielded significantly better performance than all the state-of-the-art prognostic gene signatures in the global population. GENIUS was not significantly superior in all subtypes but it was the only signature that performed well whatever the subtype.
- GENIUS significantly outperformed prognostic clinical models in the global population. Its superiority did not reach significance in all the subtypes.

GENIUS

Performance comparison with prognostic clinical models



Conclusions

Biological Insights

- From global prognostic gene signatures we highlighted the importance of proliferation when robustly quantified through gene expression profiling [Sotiriou et al., 2006, Wirapati et al., 2008, Desmedt et al., 2008].

- From local prognostic gene signatures we showed that the prognostic factors dramatically depend on the molecular subtypes, i.e. proliferation for ER+/HER2-, immune response for ER-/HER2- and HER2+ and angiogenesis for HER2+
  [Wirapati et al., 2008, Desmedt et al., 2008, Haibe-Kains, 2009].

- Since the ER+/HER2- subtype is the most common one ($\approx 60\%$ of the patients), we showed that the prognostic ability of most global gene signatures (e.g. GENE70, GENE76 or GGI) is driven by proliferation-related genes
  [Wirapati et al., 2008, Desmedt et al., 2008, Haibe-Kains et al., 2008b].

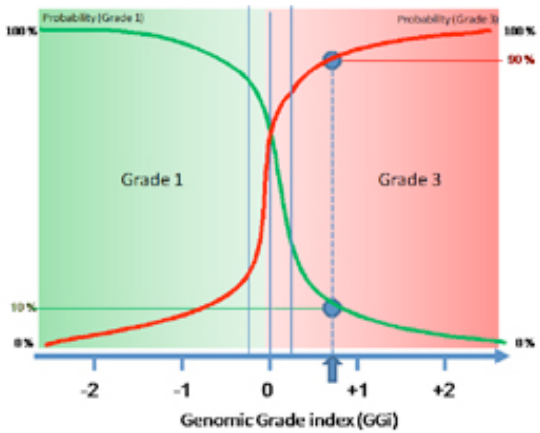
Translational Research

IPSOGEN
PROFILING BREAST CANCER

MAPQUANT DX | MAPQUANT DX GENOMIC GRADE | PARTNERS | ORDERING & LOGISTICS | NEWS & EVENTS | CORPORATE WEBSITE | LEUKEMIA PRODUCTS

Improving the clinical value of tumor grading

Map*Quant* Dx™ Genomic Grade is the cornerstone of the Map*Quant* Dx™ assay series. It is the very first, microarray-based and clinically-validated, molecular diagnostic test to accurately measure tumor grade, a consensus indicator of tumor proliferation, risk of metastasis and response to chemotherapy.



Resolving histological grading uncertainty

Tumor grade is a decision factor in most national & international guidelines to breast cancer treatment. It is generally recommended to treat high-grade "grade 3" breast carcinoma with chemotherapy because they are chemosensitive and will often recur otherwise. By contrast, most low-grade "grade 1" tumors should not be treated with chemotherapy because they have a good prognosis and often are chemo-insensitive.

A critical clinical issue is how to treat the 50% of breast cancers tested today as intermediate grade?

The Map*Quant* Dx™ Genomic Grade test now allows to resolve more than 80% of these uncertain "grade 2" tumors into "grade 1" or "grade 3" tumors, potentially sparing useless chemotherapy treatments to tens of thousands patients a year.

Clinical utility of the Genomic Grade index

Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis.
J Natl Cancer Inst. 2006 Feb 15;98(4):262-72.
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, VandeVijver MJ , Bergh J , Piccart M , Delorenzi M.

Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade.
J Clin Oncol. 2007 Apr1; 25(10):1239-46.
Loi S, Haibe-Kains B, Desmedt C, Lallemand F, Tutt AM, Gillet C, Ellis P, Harris A, Bergh J, Foekens JA, Klijn JG, Larsimont D, Buyse M, Bontempi G, Delorenzi M, Piccart MJ, Sotiriou C.

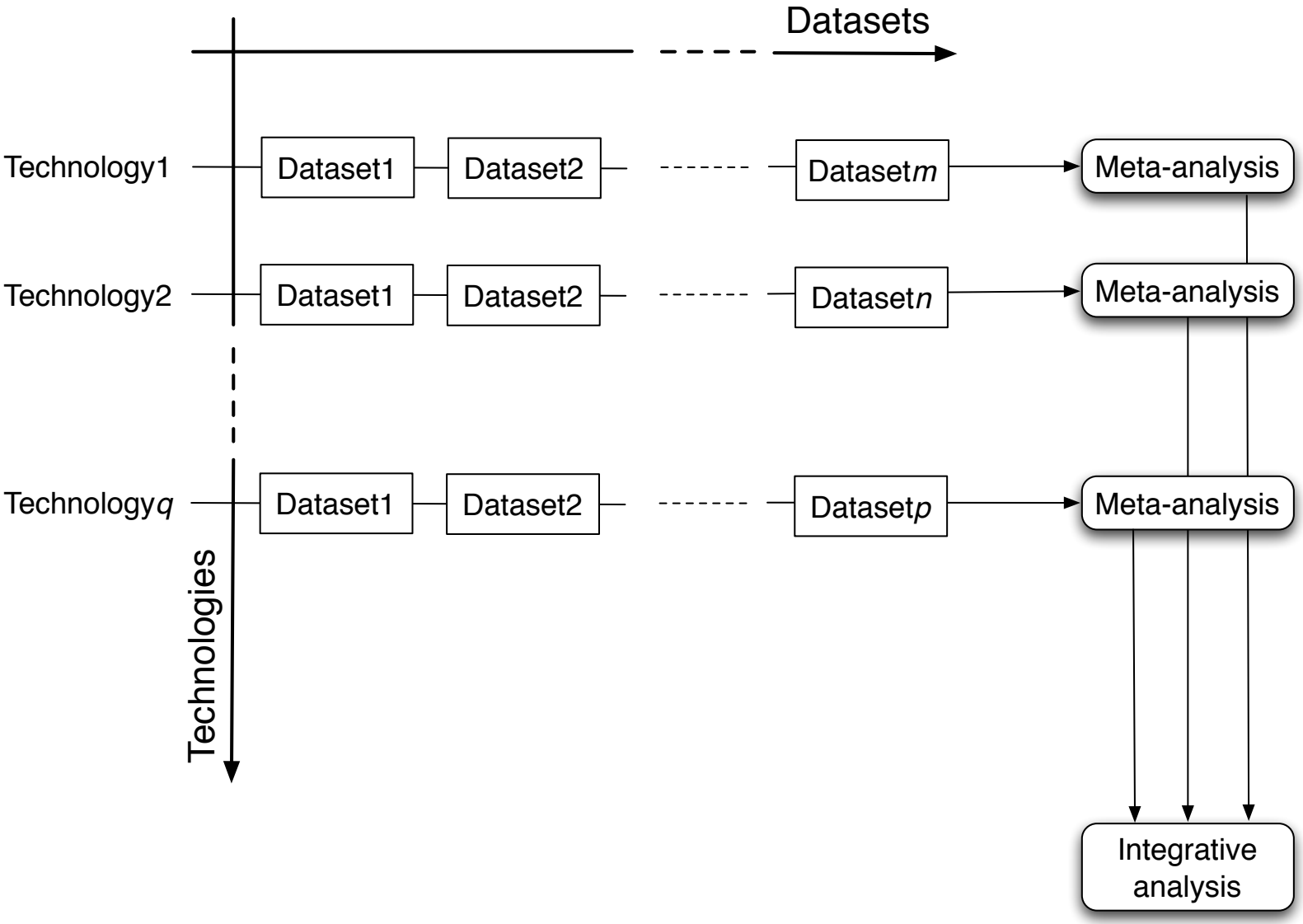
Software

- `survcomp` R package available from CRAN:
  - Implementation of 6 performance criteria:
    1. Hazard ratio
    2. D index
    3. Concordance index
    4. Time-dependent ROC curve
    5. Cross-validated partial likelihood
    6. Brier score
  - Implementation of 2 statistical tests for performance comparison:
    - ★ Paired student t test for 1, 2, and 3
    - ★ Wilcoxon signed-rank test for 4, 5, and 6.

- SWEAVE code: The LaTeX and R codes performing the whole analysis of the microarray and survival data are publicly available for [Haibe-Kains et al., 2008a, Haibe-Kains et al., 2008b].

Future of Bioinformatics

Integrative bioinformatics



Thank you for your attention.

This presentation is available from http://www.ulb.ac.be/di/map/bhaibeka/papers/haibekains2009robust_talk_dfci.pdf.

Bibliography

Bibliography I


Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977).
Maximum likelihood from incomplete data via the EM algorithm.
Journal of the Royal Statistical Society, B, 39(1):1–38.


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.


Eleuteri, A., Tagliaferri, R., Milano, L., De Placido, S., and De Laurentiis, M. (2003).
A novel neural network-based survival analysis model.
Neural Netw, 16(5-6):855–864.

Bibliography II


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).


Green, B. F. (1977).
Parameter sensitivity in multivariate methods.
Multivariate Behavioral Research, 12(3):263–287.


Haibe-Kains, B. (2009).
Identification and Assessment of Dene Signatures in Human Breast Cancer.
PhD thesis, Université Libre de Bruxelles.


Haibe-Kains, B., Desmedt, C., Loi, S., Delorenzi, M., Sotiriou, C., and Bontempi, G. (2008a).
Computational Intelligence in Clinical Oncology : Lessons Learned from an Analysis of a Clinical Study, volume 122 of *Studies in Computational Intelligence*, chapter 10, pages 237–268.
Springer-Verlag Berlin/Heidelberg.

Bibliography III


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.


Harrell, F. J., Lee, K., and Mark, D. (1996).
Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.
Stat Med, 15(4):361–387.


Johansen, T. A. and Foss, B. A. (1993).
Constructing NARMAX models using ARMAX models.
International Journal of Control, 58:1125–1153.


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.


Kittler, J., Hatef, M., Duin, R., and Matas, J. (1998).
On combining classifiers.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3):226–238.

Bibliography IV


Loi, S., Haibe-Kains, B., Desmedt, C., Wirapati, P., Lallemand, F., Tutt, A., Gillet, C., Ellis, P., Ryder, K., Reid, J., Daidone, M., Pierotti, M., Berns, E., Jansen, M., Foekens, J., Delorenzi, M., Bontempi, G., Piccart, M., and Sotiriou, C. (2008).
Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen.
BMC Genomics, 9(1):239.


Meyer, P. E. (2008).
Information-Theoretic Variable Selection and Network Inference from Microarray Data.
PhD thesis, Université Libre de Bruxelles.


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.


Schwarz, G. (1978).
Estimating the dimension of a model.
Annals of Statistics, 6:461–464.

Bibliography V

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., 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.

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

Bibliography VI


van Belle, V., Pelckmans, K., Suykens, J. A., and van Huffel, S. (2007).
Support vector machines for survival analysis.
In *Third International Conference on Computational Intelligence in Medicine and Healthcare*.


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.


Wainer, H. (1976).
Estimating coefficients in linear models: It don't make no nevermind.
Psychological Bulletin, 83(2):213–217.

Bibliography VII


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.


Wessels, L. F. A., Reinders, M. J. T., Hart, A. A. M., Veenman, C. J., Dai, H., He, Y. D., and van't Veer, L. J. (2005).
A protocol for building and evaluating predictors of disease state based on microarray data.
Bioinformatics, 21(19):3755–3762.

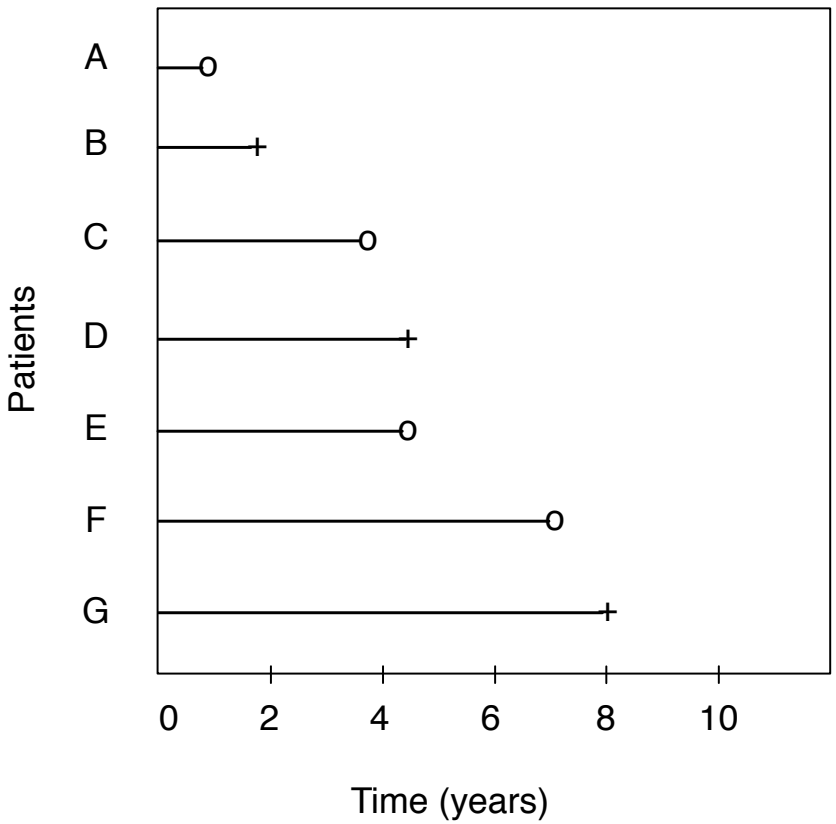

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Marks, J. R., and Nevins, J. R. (2001).
Predicting the clinical status of human breast cancer by using gene expression profiles.
PNAS, 98(20):11462–11467.


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.

Appendix

Survival Data

- Retrospective observation plan (data retrieved from hospital DB).
  - Censoring



  - Survival data

Patient id	Time (years)	Event
A	1	1
B	2	0
C	4	1
D	5	0
E	5	1
F	7	1
G	8	0

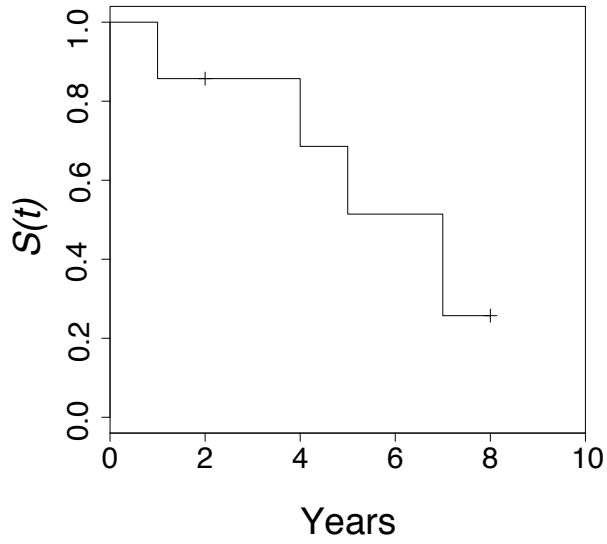
Survival analysis

Censoring in survival data requires specific methods to deal with.

Survival Distributions and Performance

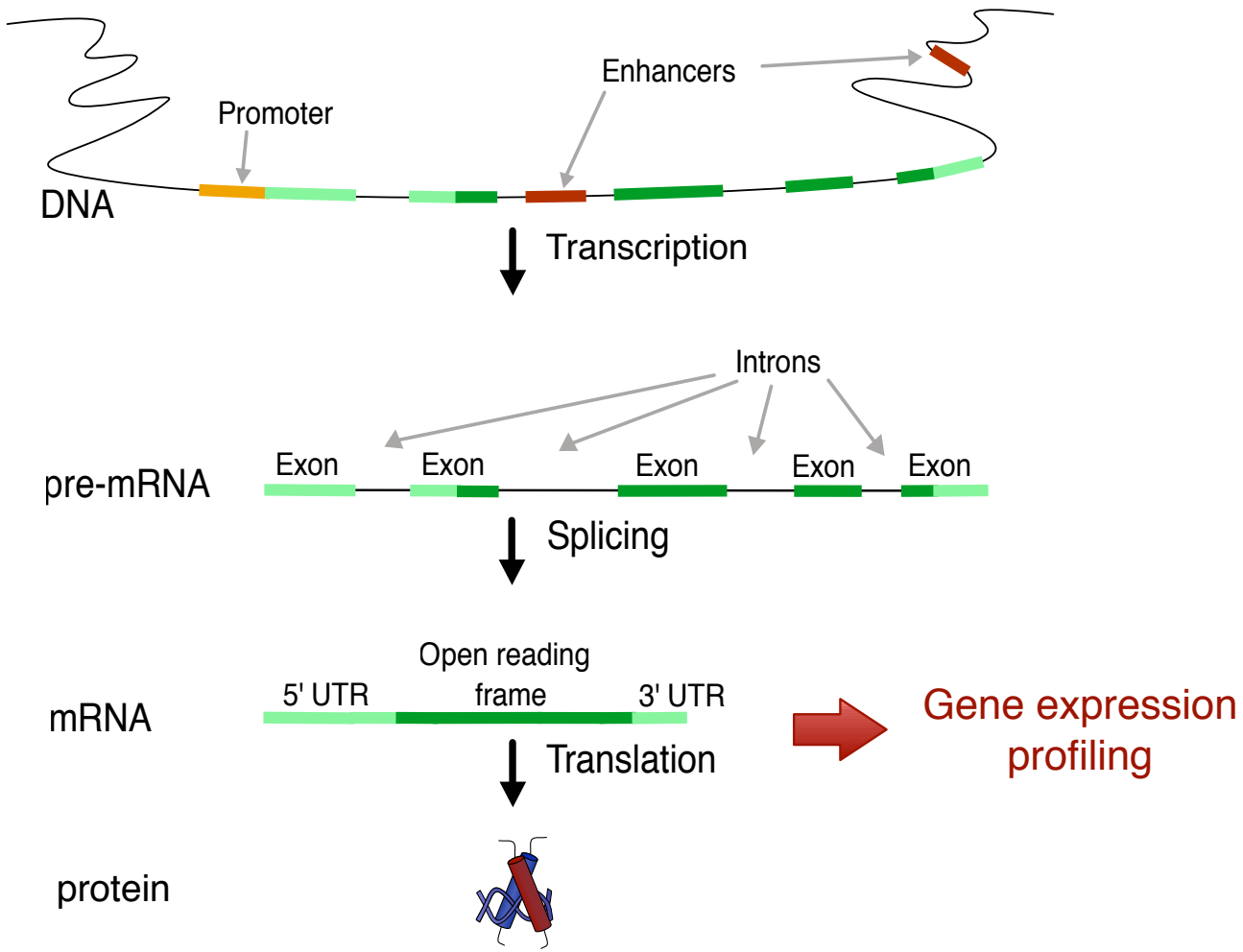
- Time of event occurrence are realization of a random variable **t**.
- Survival distributions:
  - $S(t) = 1 - \Pr\{\mathbf{t} \leq t\}$.
  - $h(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t \leq \mathbf{t} < t + \Delta t \,|\, \mathbf{t} \geq t\}}{\Delta t}$

Kaplan-Meier survival curve



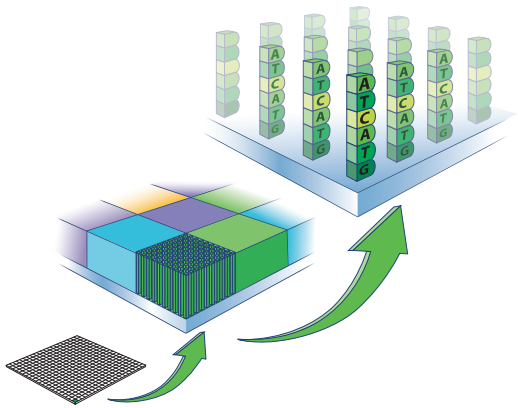
- Influence of covariates x:
  - Proportional hazards model (Cox):
    $h_i(t) = \underbrace{\lambda_0(t)}_{\text{unspecified baseline}} \exp(\underbrace{\beta x_i}_{\text{risk}})$
    where x_i is the gene expression profile of patient i in the thesis.
- Performance:
  - Concordance index: Estimate of the probability that, for a pair of randomly chosen comparable patients, the patient with the higher risk prediction will experience an event before the lower risk patient.
  - ➠ C-index $\in [0, 1]$ (higher is better in the thesis).

Central Dogma of Molecular Biology

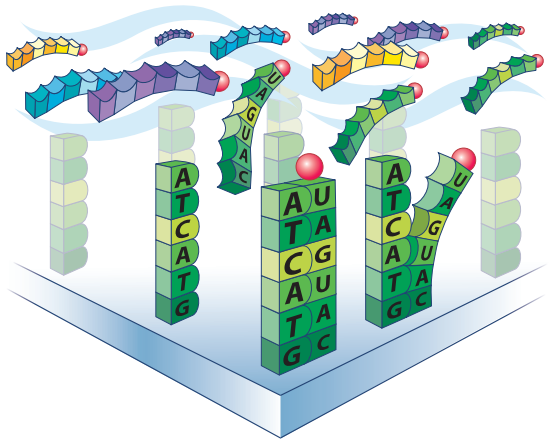


Microarray Technology

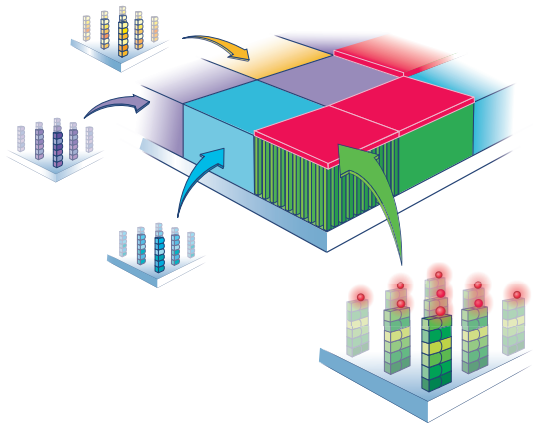
Microarray chip



Hybridization



Detection



Microarray Data

- Few samples (dozens to hundreds patients).
  - Microarray technology is expensive.
  - Frozen tumor samples are rare (biobank).
- Numerous gene expressions are measured (tens of thousands genes).
  - 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 requiring feature selection).

- Microarray is a complex technology.

➠ High level of noise in the measurements.

- Biology is complex (e.g. gene co-expressions due to pathways).

➠ Redundancy.

- Microarray data matrix

DATA	x_1	x_2	...	x_p
$patient_1$	1.3	3.2	...	7
$patient_2$	2.7	1.2	...	2.3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$patient_n$	5.4	1.4	...	9.1

Datasets

Dataset	Technology	Survival	Treatment	Patients	Probes
VDX	Affymetrix	YES	untreated	344	22,283
NKI	Agilent	YES	heterogeneous	345	24,481
STNO2	cDNA Stanford	YES	heterogeneous	122	7,787
NCI	cDNA NCI	YES	heterogeneous	99	6,878
MGH	Arcturus	YES	hormono	60	11,421
MSK	Affymetrix	YES	heterogeneous	99	22,283
UPP	Affymetrix	YES	heterogeneous	251	22,283
STK	Affymetrix	YES	heterogeneous	159	22,283
UNT	Affymetrix	YES	untreated	137	22,283
UNC2	Agilent	YES	heterogeneous	248	21,495
DUKE	Affymetrix	YES	heterogeneous	171	12,625
CAL	Affymetrix	YES	heterogeneous	118	22,283
TBG	Affymetrix	YES	untreated	198	22,283
NCH	Agilent	YES	heterogeneous	135	17,086
DUKE2	Affymetrix	NO	chemo	160	61,359
MAINZ	Affymetrix	YES	untreated	200	22,283
TAM	Affymetrix	YES	hormono	354	44,928
TAM2	Aminolink	YES	hormono	155	21,332
LUND2	Swegene	NO	hormono	105	27,648
LUND	Swegene	NO	heterogeneous	143	26,824
MUG	Operon	NO	NA	152	16,783

State-of-the-art: [van't Veer et al., 2002]

- Study of [van't Veer et al., 2002]:
  - Training set:
    - ★ Input data: 24,481 (5000 after filtering) gene expressions of 78 untreated patients with early breast cancer.
    - ★ Output data: Dichotomization of survival data (5 years).
  - Feature selection: Feature ranking (correlation).
  - Model building: Nearest centroid classifier.
  - Hyperparameter tuning: Supervised tuning of the signature size (cross-validation).
  - Validation set:
    - ★ Internal: Cross-validation !!!
    - ★ External: 295 treated and untreated patients with breast cancer at various stages [van de Vijver et al., 2002] (NKI) !!!

State-of-the-art: [Perou et al., 2000, Sorlie et al., 2001]

- Study of [Perou et al., 2000, Sorlie et al., 2001]:
  - Training set:
    - ★ Input data: 8,102 gene expressions of 84 patients with breast cancer at various stages.
  - Feature selection: Feature ranking based on variance (*intrinsic gene list*).
  - Model building: Hierarchical clustering.
  - Hyperparameter tuning: Number of clusters is selected based on visualization assessment !!!
- Validation:
  - Internal: None !!!
  - External: 38 new patients in [Sorlie et al., 2003] + datasets from [van't Veer et al., 2002] and [West et al., 2001] !!!

State-of-the-art: [Wang et al., 2005]

- Study of [Wang et al., 2005]:
  - Training set:
    - ★ Input data: 22,283 (17,819 after filtering) gene expressions of 115 untreated patients with early breast cancer.
    - ★ Output data: "Full" survival data and dichotomization (5 years) !!!
  - Feature selection: Feature ranking based on significance of univariate Cox's models (bootstrap) for ER- and ER+ separately.
  - Model building: Combination of univariate Cox's models.
  - Hyperparameter tuning: Signature size tuned by optimizing the sensitivity and specificity in the training set.
  - Validation set:
    - ★ Internal: 171 patients (single "random" split) !!!
    - ★ External: 180 untreated patients with early breast cancer.

Comparative Study of Risk Prediction Models

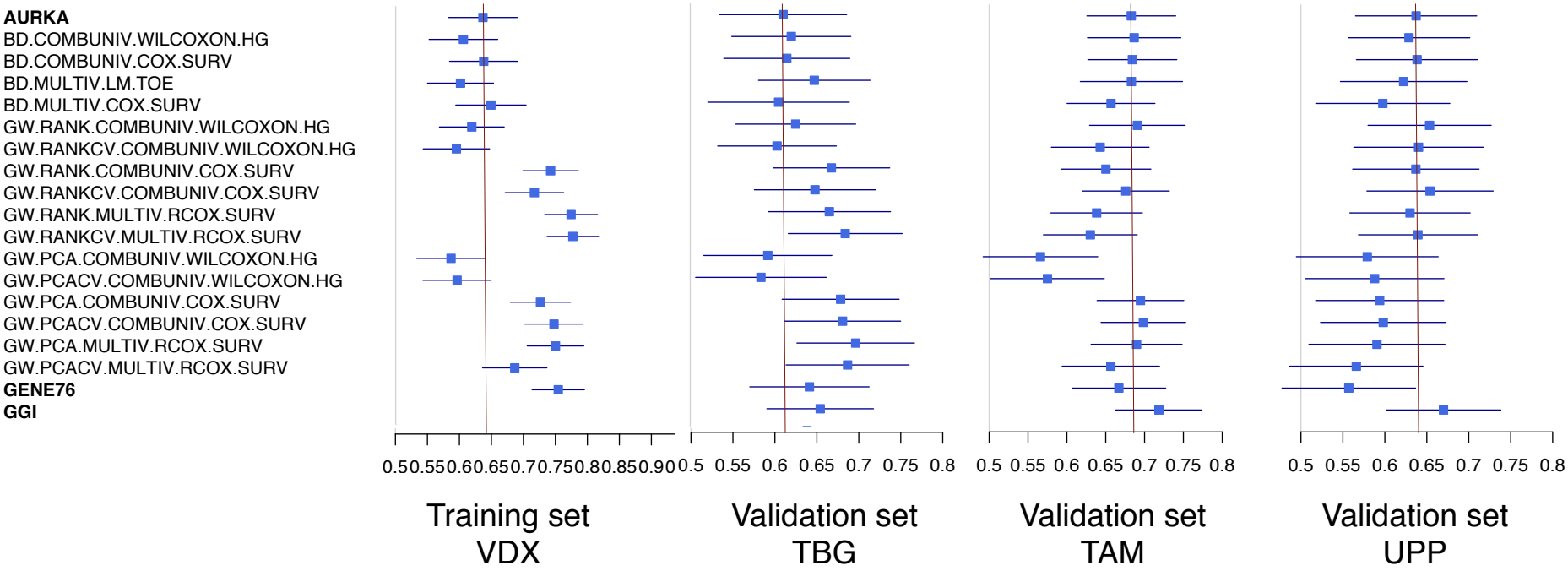
- We compared numerous linear risk prediction models in the context of breast cancer prognostication using microarray and survival data [Haibe-Kains et al., 2008b].
- Risk prediction models:

	Genotype	Dim. reduction	Structure	Learning algo.	Phenotype
1	**AURKA**				
2	BD		COMBUNIV	WILCOXON	HG
3	BD		COMBUNIV	COX	SURV
4	BD		MULTIV	LM	TOE
5	BD		MULTIV	COX	SURV
6	GW	RANK(CV)	COMBUNIV	WILCOXON	HG
7	GW	RANK(CV)	COMBUNIV	COX	SURV
8	GW	RANK(CV)	MULTIV	RCOX	SURV
9	GW	PCA(CV)	COMBUNIV	WILCOXON	HG
10	GW	PCA(CV)	COMBUNIV	COX	SURV
11	GW	PCA(CV)	MULTIV	RCOX	SURV
12				**GENE76**	
13				**GGI**	

Comparative Study of Risk Prediction Models

Results

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



- ➟ "Complex" methods, although promising in the training set, yielded poor performance in independent sets (overfitting).
- ➟ The loss of interpretability deriving from the use of overcomplex methods may be not sufficiently counterbalanced by significantly better performance.

Global Prognostic Gene Signature Identification

Biological validation

- **GGI** = **G**ene expression **G**rade **I**ndex.
  - Aim: Understanding the molecular basis of histological grade to improve prognosis.
- GGI refers to:
  - The signature of 128 probes selected through feature ranking with fixed threshold.
  - The robust model.
- Training set:
  - Input data: 22,283 gene expressions of 64 ER+ tamoxifen treated patients with breast cancer at various stages.
  - Output data: Binary class defined by histological grade 1 or 3.
- Validation set: 745 untreated patients with early breast cancer from 5 datasets (NKI, TBG, UPP, UNT and MAINZ).

- Histological grade:
  - HG1 and HG3 tumors have distinct gene expression profiles characterized by proliferation-related genes, HG2 tumors being heterogeneous.
  - GGI is strongly predictive of HG1 vs HG3.

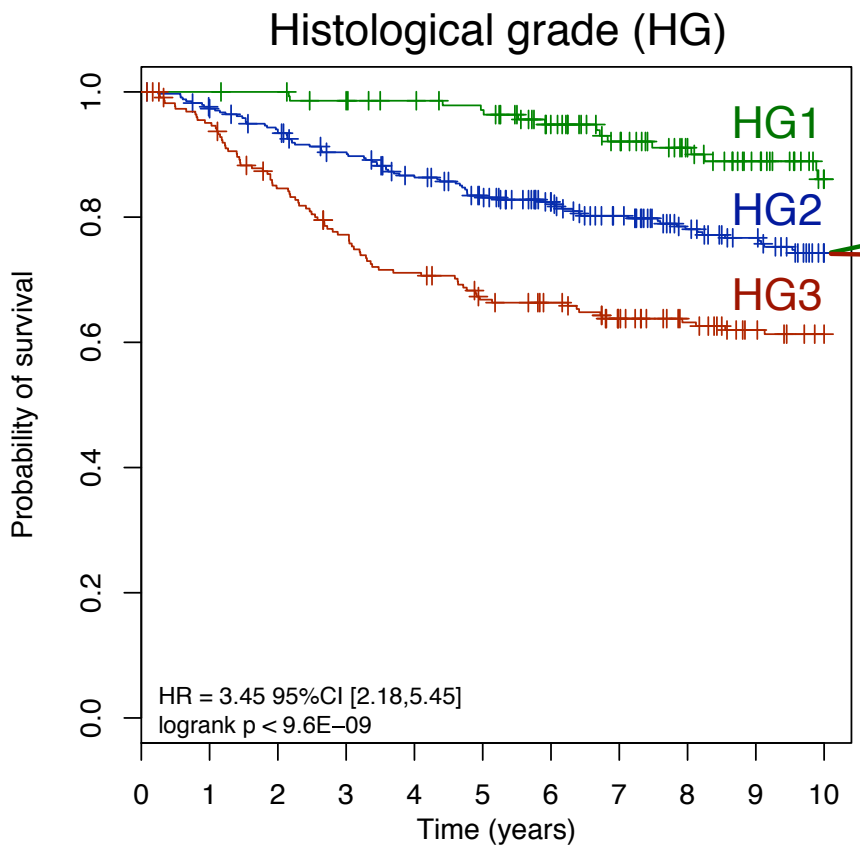
- Prognosis:
  - Strong prognostic value: GGI is strongly associated with distant metastasis free survival (DMFS). The three-category histological grading system could be replaced with a two-catehory one using GG to improve prognosis.
  - Proliferation and modeling: We showed in [Haibe-Kains et al., 2008b] that GGI is the only risk prediction model to outperform the single proliferation gene AURKA.
  - Similar performance compared state-of-the-art gene signatures: GGI is competitive with GENE70 [van't Veer et al., 2002] and GENE76 [Wang et al., 2005].

GGI

Results

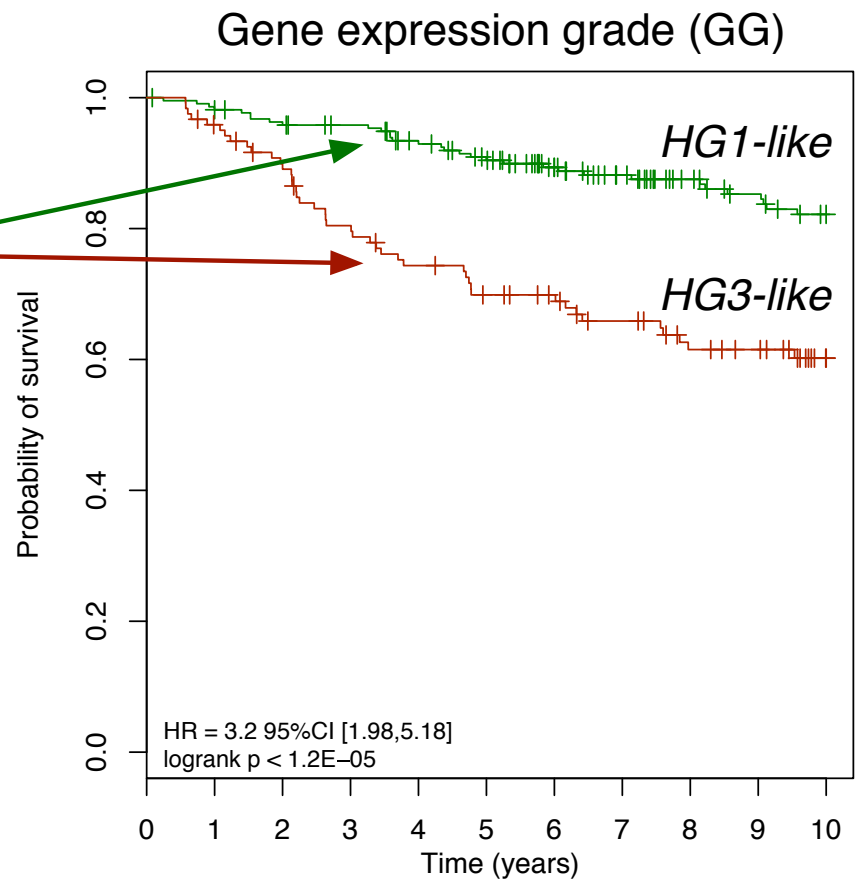
- Histological grade:
  - HG1 and HG3 tumors have distinct gene expression profiles characterized by proliferation-related genes, HG2 tumors being heterogeneous.
  - GGI is strongly predictive of HG1 vs HG3.

- Prognosis:
  - Strong prognostic value: GGI is strongly associated with distant metastasis free survival (DMFS). The three-category histological grading system could be replaced with a two-catehory one using GG to improve prognosis.
  - Proliferation and modeling: We showed in [Haibe-Kains et al., 2008b] that GGI is the only risk prediction model to outperform the single proliferation gene AURKA.
  - Similar performance compared state-of-the-art gene signatures: GGI is competitive with GENE70 [van't Veer et al., 2002] and GENE76 [Wang et al., 2005].



No. At Risk

HG1	144	144	144	139	136	132	115	101	86	74	58
HG2	338	326	311	295	274	257	227	203	176	163	141
HG3	226	212	185	166	153	139	133	119	107	96	89



No. At Risk

HG2/GG1	217	212	206	202	190	180	156	139	121	112	102
HG2/GG3	121	115	105	94	85	78	73	65	56	52	39

GGI

Results

- Histological grade:
  - HG1 and HG3 tumors have distinct gene expression profiles characterized by proliferation-related genes, HG2 tumors being heterogeneous.
  - GGI is strongly predictive of HG1 vs HG3.

- Prognosis:
  - Strong prognostic value: GGI is strongly associated with distant metastasis free survival (DMFS). The three-category histological grading system could be replaced with a two-catehory one using GG to improve prognosis.
  - Proliferation and modeling: We showed in [Haibe-Kains et al., 2008b] that GGI is the only risk prediction model to outperform the single proliferation gene AURKA.
  - Similar performance compared state-of-the-art gene signatures: GGI is competitive with GENE70 [van't Veer et al., 2002] and GENE76 [Wang et al., 2005].

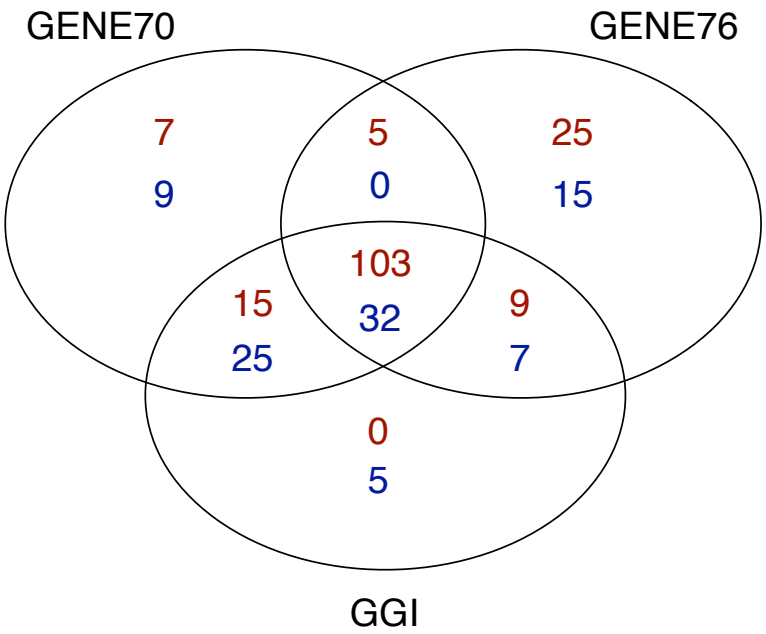
- Histological grade:
  - HG1 and HG3 tumors have distinct gene expression profiles characterized by proliferation-related genes, HG2 tumors being heterogeneous.
  - GGI is strongly predictive of HG1 vs HG3.

- Prognosis:
  - Strong prognostic value: GGI is strongly associated with distant metastasis free survival (DMFS). The three-category histological grading system could be replaced with a two-catehory one using GG to improve prognosis.
  - Proliferation and modeling: We showed in [Haibe-Kains et al., 2008b] that GGI is the only risk prediction model to outperform the single proliferation gene AURKA.
  - Similar performance compared state-of-the-art gene signatures: GGI is competitive with GENE70 [van't Veer et al., 2002] and GENE76 [Wang et al., 2005].

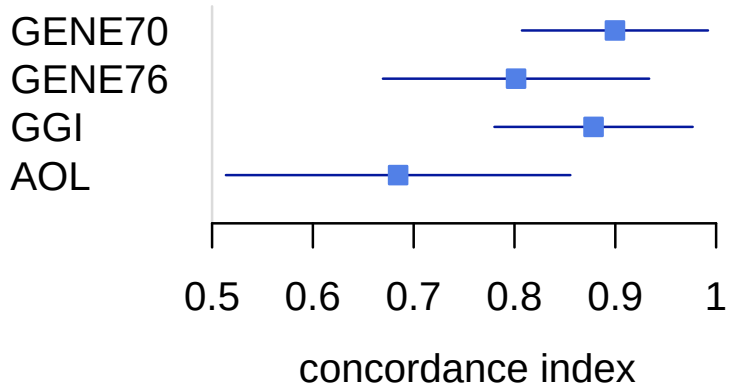
GGI

Performance comparison (TBG)

- Concordance in predictions:



- Similar performance:



	Test difference
GENE70 vs GENE76	0.15
GENE70 vs GGI	0.53
GENE76 vs GGI	0.22

	Test superiority
GENE70 vs AOL	0.007
GENE76 vs AOL	0.13
GGI vs AOL	0.015

- GGI is simple:
  - From a statistical point of view: robust model building.
  - From a biological point of view: GGI signature is mainly composed of proliferation-related genes.

- GGI yielded similar performance than gene signatures including genes related to many biological processes (GENE70 and GENE76).
  - ➠ Actually, we showed in [Wirapati et al., 2008] that the prognostic value of these signatures is mainly driven by the proliferation-related genes.

- The studies related to GGI challenged the use of risk prediction models of high statistical and biological complexity

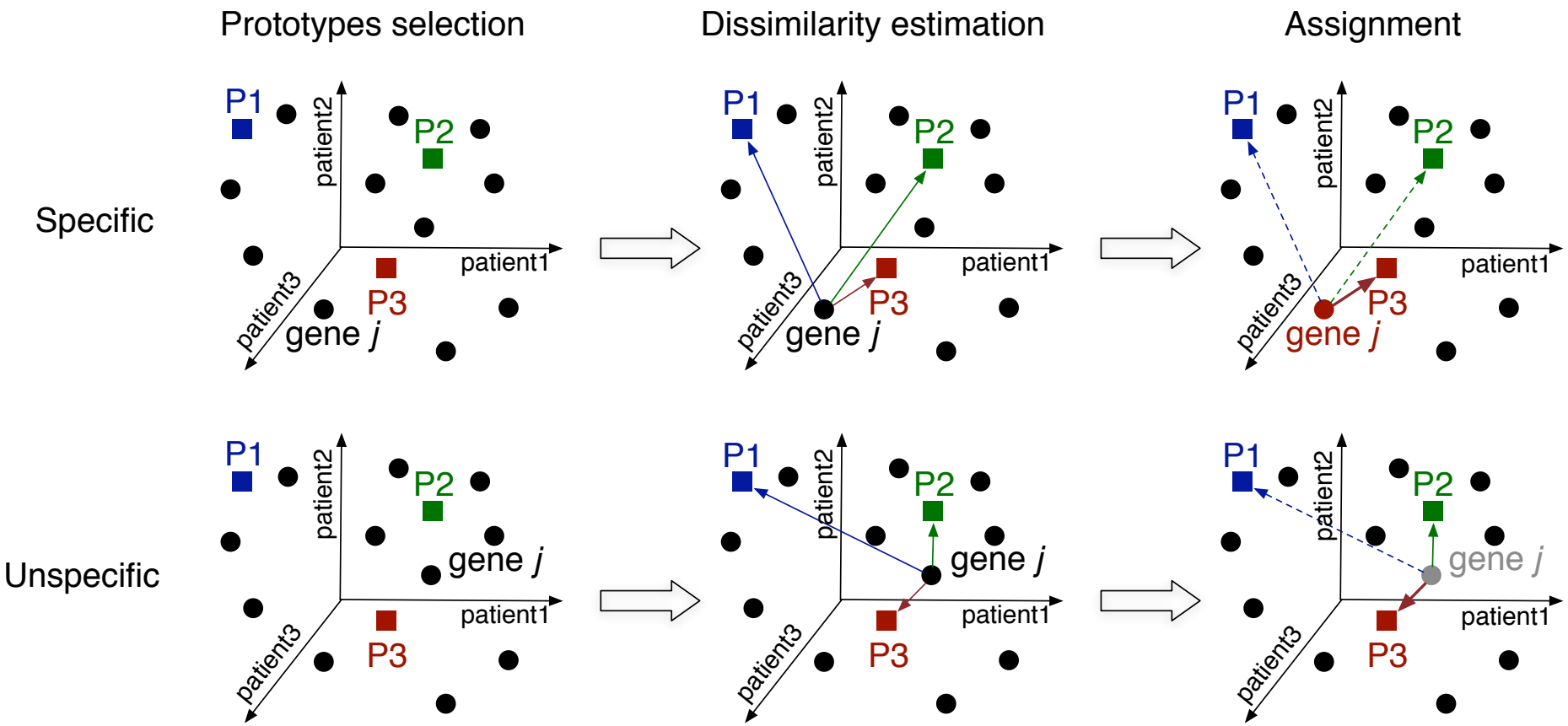
  [Sotiriou et al., 2006, Haibe-Kains et al., 2008b, Desmedt et al., 2008].

Molecular Subtypes Identification

State-of-the-Art

- State-of-the-art:
  - [Perou et al., 2000, Sorlie et al., 2001] were the first to study the molecular heterogeneity of breast cancer in order to build a subtype classifier.
  - Open Issues:
    - ★ Lack of estimation of classification uncertainty.
    - ★ The use of large number of genes led to a clustering model prone to overfitting (low robustness).
    - ★ The (number of) clusters were subjectively selected.
    - ★ Lack of existing statistics to evaluate the robustness of the clustering model in independent datasets.
    - ★ Association with survival tested in heterogeneous populations of breast cancer patients.
- [Kapp et al., 2006] confirmed the importance of ER and HER2 signaling pathways.

Prototype-Based Clustering



Clustering Performance

- Performance assessment of clustering is a difficult task since the "truth" is hidden.
- In [Tibshirani and Walther, 2005], the authors introduced a new framework viewing clustering as a supervised learning problem in which the "true" class labels have to be estimated.

- Let X and Y be the training and validation sets, respectively.
- Let $C_X(Y)$ denote the use, in the dataset Y, of a clustering model C fitted on dataset X.
- Let $D[C_\cdot(.)]$ be the co-membership matrix of the clustering $C_\cdot(.)$.
- Idea:
  1. Cluster $Y \rightarrowtail C_Y(Y)$.
  2. Cluster $X \rightarrowtail C_X(X)$.
  3. Cluster Y using $C_X \rightarrowtail C_X(Y)$.
  4. Compare $C_X(Y)$ and $C_Y(Y)$ through the co-membership matrix D $\rightarrowtail$*prediction strength*.

Subtype Clustering vs Perou's Method

Prediction strength in independent datasets (higher is better)

Dataset	Subtype clustering	Perou's method			
	3 clusters	2 clusters	3 clusters	4 clusters	5 clusters
NKI	**1.00**	0.89	0.42	0.25	0.20
TBG	**0.83**	0.92	0.39	0.24	0.22
UPP	**0.87**	0.71	0.51	0.33	0.22
UNT	**0.89**	0.78	0.55	0.36	0.00
STNO2	**0.69**	0.86	0.44	0.28	0.25
NCI	**0.83**	0.93	0.44	0.33	0.13
STK	**0.87**	0.61	0.38	0.28	0.25
MSK	**0.96**	0.77	0.66	0.20	0.00
UNC2	**0.87**	0.81	0.57	0.31	0.00
NCH	**0.82**	0.66	0.49	0.36	0.29
DUKE	**0.82**	0.57	0.42	0.37	0.42
DUKE2	**0.64**	0.92	0.63	0.47	0.00
MAINZ	**0.90**	0.68	0.39	0.24	0.18
CAL	**0.95**	0.84	0.41	0.31	0.00
LUND2	**0.87**	0.92	0.51	0.17	0.17
LUND	**0.81**	0.55	0.36	0.24	0.20
MUG	**0.49**	0.50	0.33	0.27	0.23
mean	**0.83**	0.76	0.47	0.30	0.16
sd	0.12	0.14	0.10	0.07	0.12