Molecular subtypes identification to refine breast cancer prognosis

Benjamin Haibe-Kains[1,2]

[1] Functional Genomics Unit, Institut Jules Bordet

[2] Machine Learning Group, Université Libre de Bruxelles

October 16, 2008

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 (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).
- The MLG is part to the "Groupe de Contact FNRS" on Machine Learning and to CINBIOS: http://babylone.ulb.ac.be/Joomla/.

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.

Summary

- Introduction
  - Breast Cancer
  - Prognosis
  - Gene Expression Profiling
- Breast Cancer Molecular Subtypes
- Prognostic Gene Signatures
- Subtypes and Prognosis
  - GENIUS
- Conclusion

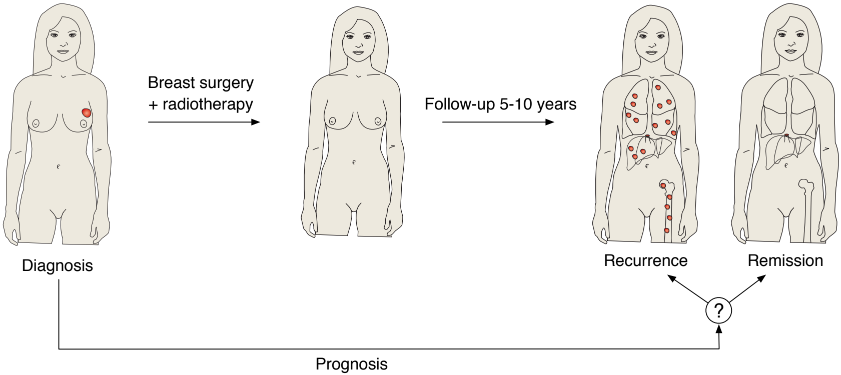
Part I

Introduction

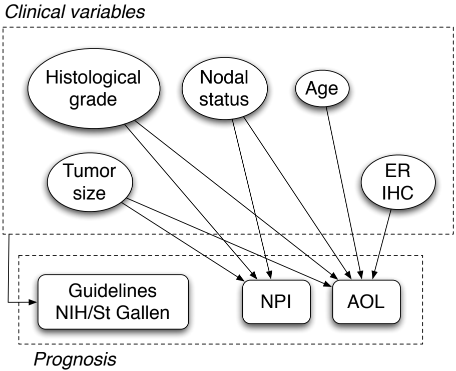
Breast Cancer

- 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



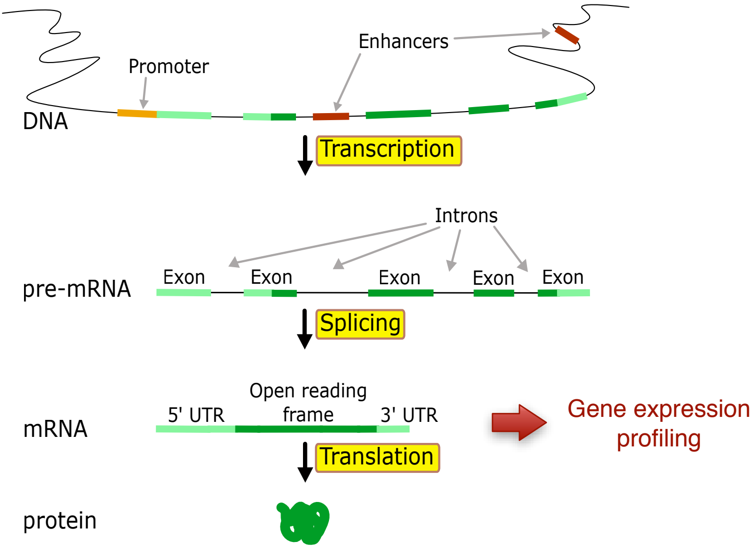
- Need to improve current clinical tools to detect patients who need adjuvant systemic therapy.

Potential of Genomic Technologies for Prognosis

- In the nineties, new biotechnologies emerged:
  - Human genome sequencing.
  - Gene expression profiling (low to high-throughput).

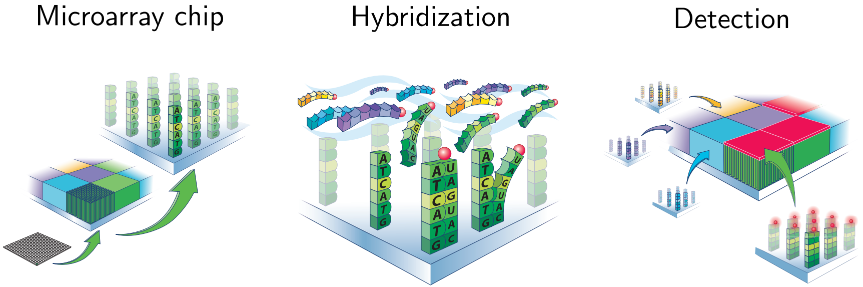
- Genomic data could be used to better understand cancer biology
- . . . and to build efficient prognostic models.

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 new 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 data.

- Biology is complex.

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

Microarray Data

Warning

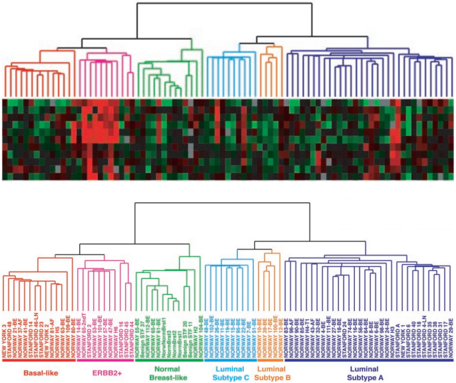
- You can easily find spurious patterns in the data, biologically ”meaningful”.
- Personal experience:
  - At the beginning of my thesis, I had accidentally mixed the patients labels, so the relation between input (gene expressions) and output (a mutation) was completely random.
  - I gave a list of genes differentially expressed between wild type and mutated patients, to the biologists in charge of the project and they found it very interesting (known genes, meaningful biological story).
  - When I saw my mistake, I corrected the bug and sent a new gene list
  - . . . and the results were even better!
- In conclusion, the complexity of microarray data and the biology behind should make you very critic and cautious with your results.

Part II

Breast Cancer Molecular Subtypes

Breast Cancer Subtypes

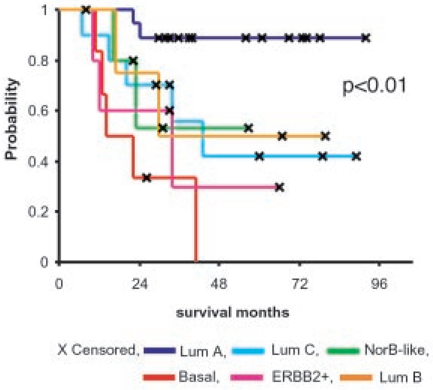
- Early microarray studies showed that BC is a molecularly heterogeneous disease [Perou et al., 2000; Sorlie et al., 2001, 2003; Sotiriou et al., 2003].
    - Hierarchical clustering on microarray data [Sorlie et al., 2001]:



Breast Cancer Subtypes

Clinical Outcome

- The molecular subtypes exhibited different clinical outcomes, suggesting that the biological processes involved in patients' survival might be different.



Breast Cancer Subtypes

Early Results

- These early studies showed similar results, i.e. ER and HER2 pathways 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.

Breast Cancer Subtypes

New Clustering Model

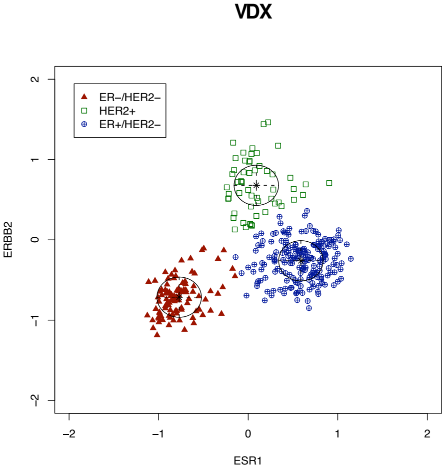
- Because of these limitations we sought to develop a simple method to identify the breast cancer subtypes.

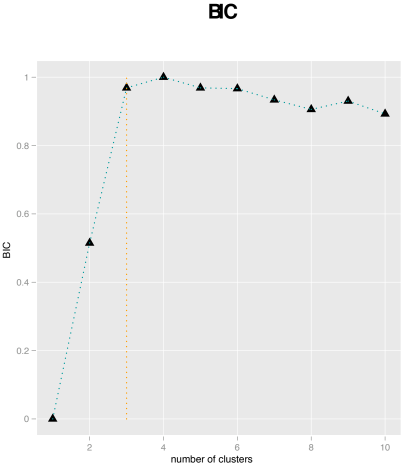
➠ We introduced a model-based clustering (mixture of Gaussians) in a two-dimensional space defined by the ESR1 and ERBB2 module scores [Wirapati et al., 2008; Desmedt et al., 2008].

  - We used the Bayesian information criterion (BIC) to select the most likely number of subtypes [Fraley and Raftery, 2002].
  - We validated our model (fitted on Wang et al. series) on 14 independent datasets in terms of number of clusters and prediction strength [Tibshirani and Walther, 2005].

New Clustering Model

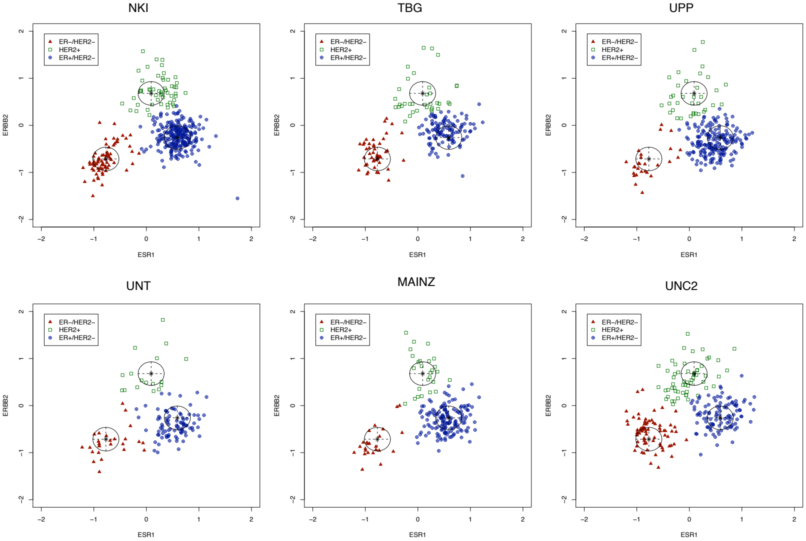
Training





New Clustering Model

Validation



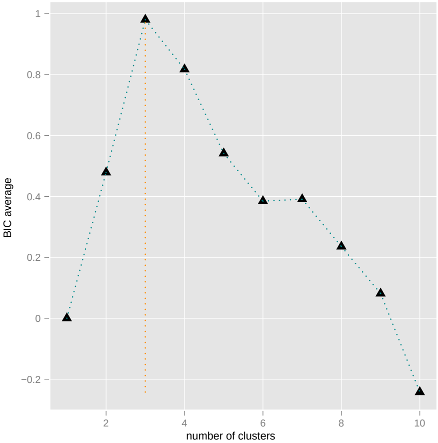
New Clustering Model

Validation: Prediction Strength

Dataset	ER-/HER2-	HER2+	ER+/HER2-
NKI	1.00	1.00	0.99
TBG	1.00	1.00	0.83
UPP	1.00	0.93	0.87
UNT	1.00	0.89	0.92
MAINZ	1.00	1.00	0.90
STNO2	1.00	0.69	0.97
NCI	0.85	0.83	0.93
MSK	1.00	1.00	0.96
STK	1.00	0.91	0.87
DUKE	1.00	0.82	0.92
UNC2	1.00	0.87	0.96
CAL	1.00	1.00	0.95
DUKE2	1.00	0.64	0.95
NCH	1.00	0.82	0.98

New Clustering Model

Validation: Number of Clusters

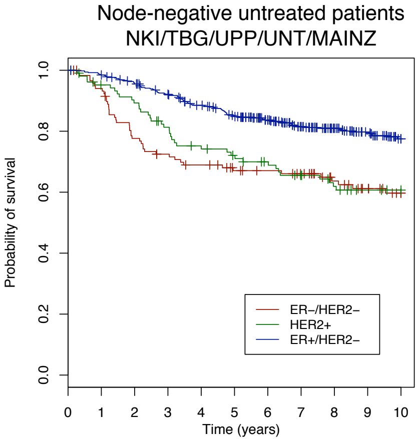


Breast Cancer Subtypes

Clinical Outcome

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

of the global population of BC patients.



Node-negative untreated patients
NKI/TBG/UPP/UNT/MAINZ

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:
    - Low computational cost to fit the model.
    - Simple visualization of the results.
- Disadvantages:
  - 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).

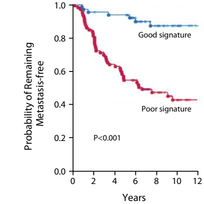
- A typical microarray analysis dealing with breast cancer prognostication involves 5 key steps:
  1. Data preprocessing: quality controls and normalization.
  2. Filtering: discard the genes exhibiting low expressions and/or low variance.
  3. Identification of a list of prognostic genes (called a *gene signature*).
  4. Building of a prognostic model, i.e. combination of the expression 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].

GENE70 signature

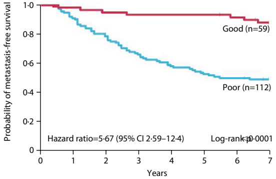


No. at Risk

Good signature	60	57	54	45	31	22	12
Poor signature	91	72	55	41	26	17	9

van't Veer et al.
van de Vijver

GENE76 signature



Patients at risk

Good signature	59	58	56	55	55	55	53	48
Poor signature	112	103	90	75	66	60	55	52

Wang et al.

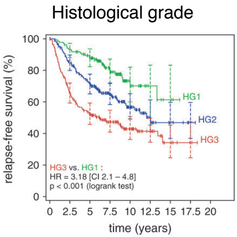
- Promising results but a lot criticisms from a statistical point of view.

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).

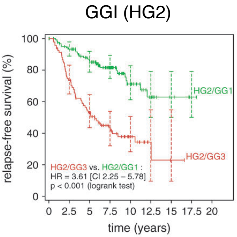
Histological grade



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

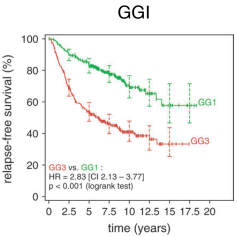
GGI (HG2)



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

Prognostic Gene Signatures

Independent Validation

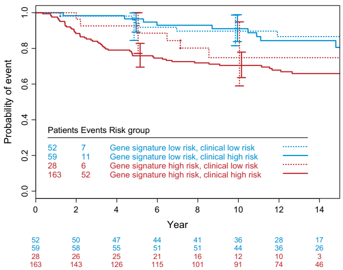
- These preliminary resulting were promising but validation was required.
- A first validation was published by the authors of the GENE70 and GENE76 signatures in [van de Vijver et al., 2002] and [Foekens et al., 2006] respectively.
- Our group was involved in a second validation:
  - Complete independence: 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

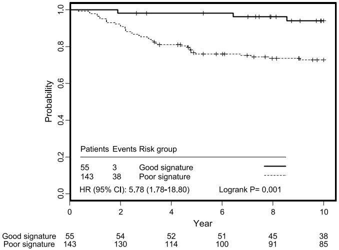
Independent Validation (cont.)

- 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

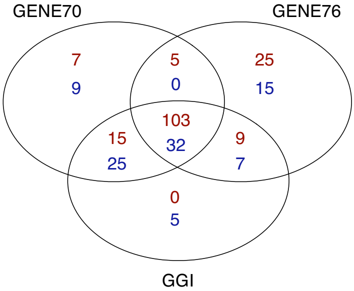


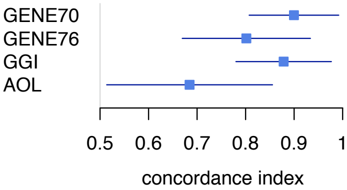
➡ Validation of GENE70 [Buyse et al., 2006] and GENE76 [Desmedt et al., 2007].

Prognostic Gene Signatures

Independent Validation (cont.)

- We sought to compare the GGI to the GENE70 and GENE76 signatures in this validation series
- . . . and showed that GGI has very similar performance [Haibe-Kains et al., 2008b].





Part IV

Subtypes and Prognosis

Prognosis in Specific Subtypes

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

- In 2005, Wang et al. were the first to divide the global population based on ER status:
  - As BC biology is very different according to the ER status, 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 [Teschendorff and Caldas, 2008] using large datasets.
  - 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 early 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

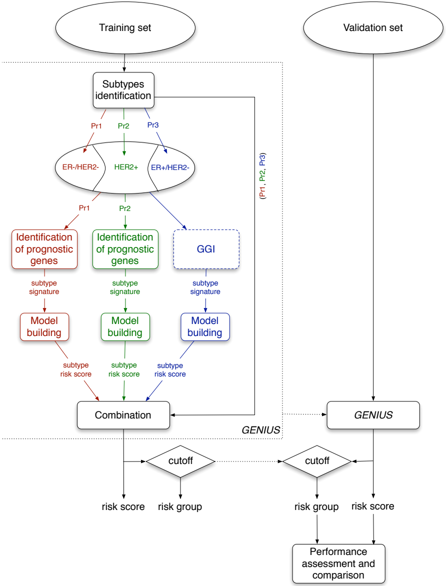
- Since current prognostic models/gene signatures are limited to some subtypes, we sought to develop a new prognostic model integrating the breast cancer subtypes identification in order to:
  - Build a prognostic gene signatures specifically targeting each subtype.
  - Build a global prognostic model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

- We assessed the performance and compared it to current prognostic models using the thorough statistical framework developed in [Haibe-Kains et al., 2008a].

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

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

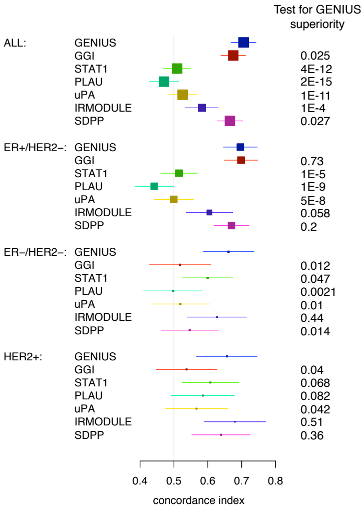
GENIUS


Training set
Validation set
Subtypes identification
Pr1
Pr2
Pr3
ER-/HER2-
HER2+
ER+/HER2-
(Pr1, Pr2, Pr3)
Pr1
Pr2
Identification of prognostic genes
Identification of prognostic genes
GGI
subtype signature
subtype signature
subtype signature
Model building
Model building
Model building
subtype risk score
subtype risk score
subtype risk score
Combination
GENIUS
GENIUS
cutoff
cutoff
risk score
risk group
risk group
risk score
Performance assessment and comparison

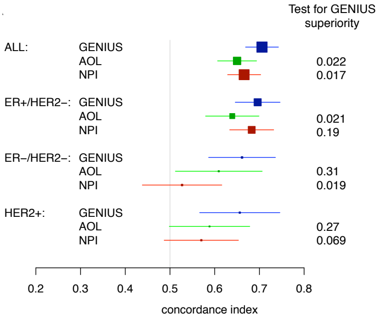
- We trained GENIUS on VDX:
  - 286 node-negative untreated BC patients.
- We assessed the performance in an independent dataset composed of
  - 765 node-negative untreated patients
  - coming from 5 different datasets (NKI, TBG, UPP, UNT and MAINZ).
- Risk score prediction: continuous value.
- Risk group prediction: binary value (application of a cutoff on the risk score).

GENIUS

Risk Score Prediction



State-of-the-art gene signatures



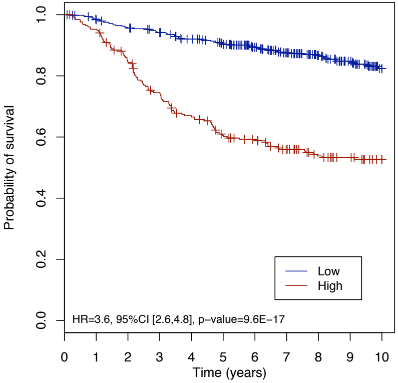
clinical prognostic indices

GENIUS

Risk Group Prediction

GENIUS

Global population

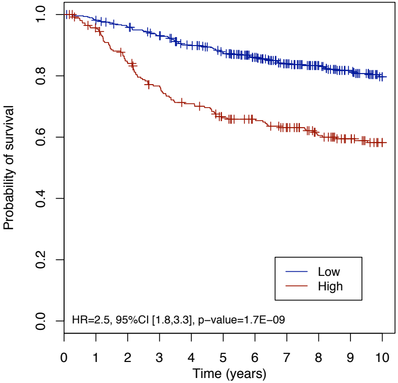


No. At Risk

Low	488	474	459	445	427	409	364	325	286	257	223
High	253	241	208	180	158	140	131	116	100	92	81

GGI

Global population



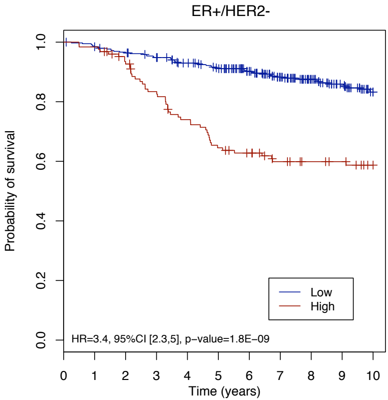
No. At Risk

Low	484	473	458	439	414	394	348	306	270	245	215
High	257	242	209	186	171	155	146	135	116	104	89

GENIUS

Risk Group Prediction (cont.)

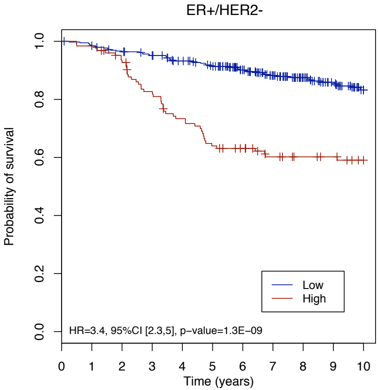
GENIUS



No. At Risk

Low	391	384	374	364	349	335	294	258	226	204	176
High	125	124	113	99	87	76	71	62	57	54	47

GGI



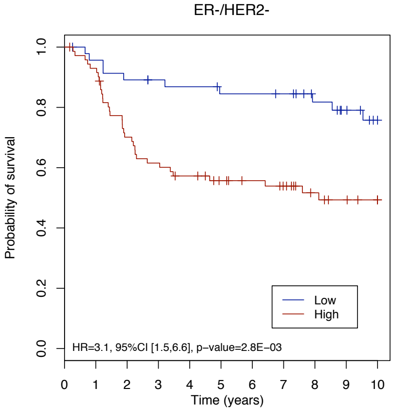
No. At Risk

Low	391	384	373	364	349	335	294	258	226	204	175
High	125	124	114	99	87	76	71	62	57	54	48

GENIUS

Risk Group Prediction (cont.)

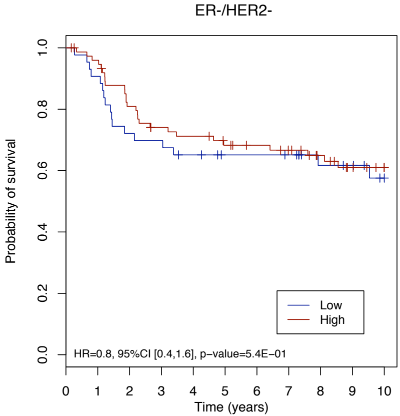
GENIUS


ER-/HER2-

No. At Risk

Low	47	45	42	40	39	37	37	36	31	27	21
High	72	67	50	44	40	35	32	29	23	20	16

GGI


ER-/HER2-

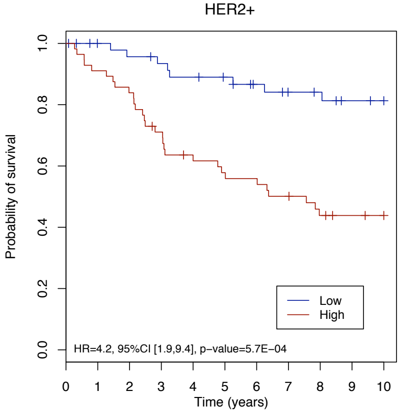
No. At Risk

Low	43	40	32	31	28	25	25	24	19	18	13
High	76	72	60	53	51	47	44	41	35	29	24

GENIUS

Risk Group Prediction (cont.)

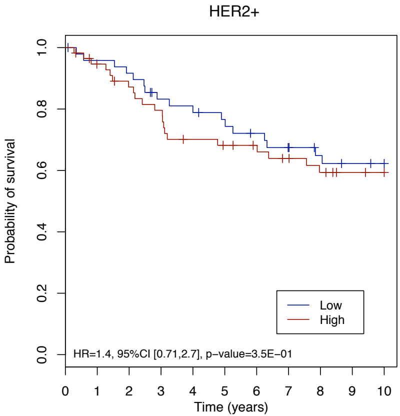
GENIUS



No. At Risk

	0	1	2	3	4	5	6	7	8	9	10
Low	50	47	45	43	41	39	35	32	31	28	26
High	56	52	47	39	33	31	30	27	22	20	18

GGI



No. At Risk

	0	1	2	3	4	5	6	7	8	9	10
Low	49	47	45	39	37	35	32	29	26	24	22
High	57	52	47	43	37	35	33	30	27	24	22

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 [shi, 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 BC 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/haibekains2008molecular.pdf.

Part VI

Bibliography

Bibliography I

The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotech*, 24(9):1151–1161, 2006. URL http://dx.doi.org/10.1038/nbt1239.

Marc Buyse, Sherene Loi, Laura van't Veer, Giuseppe Viale, Mauro Delorenzi, Annuska M. Glas, Mahasti Saghatchian d'Assignies, Jonas Bergh, Rosette Lidereau, Paul Ellis, Adrian Harris, Jan Bogaerts, Patrick Therasse, Arno Floore, Mohamed Amakrane, Fanny Piette, Emiel Rutgers, Christos Sotiriou, Fatima Cardoso, and Martine J. Piccart. 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, 2006. doi: 10.1093/jnci/djj329. URL http://jnci.oxfordjournals.org/cgi/content/abstract/jnci;98/17/1183.

C. A. Davis, F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Kuffner, and R. Zimmer. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, 22(19):2356–2363, 2006.

Bibliography II

Christine Desmedt, Fanny Piette, Sherene Loi, Yixin Wang, Francoise Lallemand, Benjamin Haibe-Kains, Giuseppe Viale, Mauro Delorenzi, Yi Zhang, Mahasti Saghatchian d'Assignies, Jonas Bergh, Rosette Lidereau, Paul Ellis, Adrian L. Harris, Jan G.M. Klijn, John A. Foekens, Fatima Cardoso, Martine J. Piccart, Marc Buyse, Christos Sotiriou, and on behalf of the TRANSBIG Consortium. 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, 2007. doi: 10.1158/1078-0432.CCR-06-2765. URL `http://clincancerres.aacrjournals.org/cgi/content/abstract/13/11/3207`.

Christine Desmedt, Benjamin Haibe-Kains, Pratyaksha Wirapati, Marc Buyse, Denis Larsimont, Gianluca Bontempi, Mauro Delorenzi, Martine Piccart, and Christos Sotiriou. Biological Processes Associated with Breast Cancer Clinical Outcome Depend on the Molecular Subtypes. *Clin Cancer Res*, 14(16):5158–5165, 2008. doi: 10.1158/1078-0432.CCR-07-4756. URL `http://clincancerres.aacrjournals.org/cgi/content/abstract/14/16/5158`.

Bibliography III

Greg Finak, Nicholas Bertos, Francois Pepin, Svetlana Sadekova, Margarita Souleimanova, Hong Zhao, Haiying Chen, Gulbeyaz Omeroglu, Sarkis Meterissian, Atilla Omeroglu, Michael Hallett, and Morag Park. Stromal gene expression predicts clinical outcome in breast cancer. *Nat Med*, 14(5):518–527, 2008. URL `http://dx.doi.org/10.1038/nm1764`.

J. A. Foekens, D. Atkins, Y. Zhang, F. C. Sweep, N. Harbeck, A. Paradiso, T. Cufer, A. M. Sieuwerts, D. Talantov, P. N. Span, V. C. Tjan-Heijnen, A. F. Zito, K. Specht, H. Hioefler, R. Golouh, F. Schittulli, M. Schmitt, L. V. Beex, J. G. Klijn, and Y. Wang. Multicenter validation of a gene expression–based prognostic signature in lymph node–negative primary breast cancer. *Journal of Clinical Oncology*, 24(11), 2006.

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

B. Haibe-Kains, C. Desmedt, C. Sotiriou, and G. Bontempi. 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, 2008a. doi: 10.1093/bioinformatics/btn374. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/19/2200`.

Bibliography IV

Benjamin Haibe-Kains, Christine Desmedt, Fanny Piette, Marc Buyse, Fatima Cardoso, Laura van't Veer, Martine Piccart, Gianluca Bontempi, and Christos Sotiriou. Comparison of prognostic gene expression signatures for breast cancer. *BMC Genomics*, 9(1):394, 2008b. ISSN 1471-2164. doi: 10.1186/1471-2164-9-394. URL http://www.biomedcentral.com/1471-2164/9/394.

Amy Kapp, Stefanie Jeffrey, Anita Langerod, Anne-Lise Borresen-Dale, Wonshik Han, Dong-Young Noh, Ida Bukholm, Monica Nicolau, Patrick Brown, and Robert Tibshirani. Discovery and validation of breast cancer subtypes. *BMC Genomics*, 7(1): 231, 2006. ISSN 1471-2164. doi: 10.1186/1471-2164-7-231. URL http://www.biomedcentral.com/1471-2164/7/231.

MJ Pencina and RB D'Agostino. Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med*, 23(13):2109–2123, 2004. doi: 10.1002/sim.1802.

Charles M. Perou, Therese Sorlie, Michael B. Eisen, Matt van de Rijn, Stefanie S. Jeffrey, Christian A. Rees, Jonathan R. Pollack, Douglas T. Ross, Hilde Johnsen, Lars A. Akslen, Oystein Fluge, Alexander Pergamenschikov, Cheryl Williams, Shirley X. Zhu, Per E. Lonning, Anne-Lise Borresen-Dale, Patrick O. Brown, and David Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797): 747–752, 2000. URL http://dx.doi.org/10.1038/35021093.

Bibliography V

Lajos Pusztai, Chafika Mazouni, Keith Anderson, Yun Wu, and W. Fraser Symmans. Molecular Classification of Breast Cancer: Limitations and Potential. *Oncologist*, 11 (8):868–877, 2006. doi: 10.1634/theoncologist.11-8-868. URL http://theoncologist.alphamedpress.org/cgi/content/abstract/11/8/868.

T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisher, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese., P. O. Brown, D. Botstein, P. L. Eystein, and A. L. Borresen-Dale. Gene expression patterns breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Matl. Acad. Sci. USA*, 98(19):10869–10874, 2001.

T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geister, J. Demeter, C. Perou, P. E. Lonning, P. O. Brown, A. L. Borresen-Dale, and D. Botstein. Repeated observation of breast tumor subtypes in indepedent gene expression data sets. *Proc Natl Acad Sci USA*, 1(14):8418–8423, 2003.

C. Sotiriou, S. Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S.B. Fox, A. L. Harris, and E. T. Liu. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci.*, 100(18):10393–10398, 2003.

Bibliography VI

Christos Sotiriou, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren, Pierre Farmer, Viviane Praz, Benjamin Haibe-Kains, Christine Desmedt, Denis Larsimont, Fatima Cardoso, Hans Peterse, Dimitry Nuyten, Marc Buyse, Marc J. Van de Vijver, Jonas Bergh, Martine Piccart, and Mauro Delorenzi. Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis. *J. Natl. Cancer Inst.*, 98(4): 262–272, 2006. doi: 10.1093/jnci/djj052. URL http://jnci.oxfordjournals.org/cgi/content/abstract/jnci;98/4/262.

Andrew Teschendorff and Carlos Caldas. A robust classifier of high predictive value to identify good prognosis patients in er-negative breast cancer. *Breast Cancer Research*, 10(4):R73, 2008. ISSN 1465-5411. doi: 10.1186/bcr2138. URL http://breast-cancer-research.com/content/10/4/R73.

Andrew Teschendorff, Ahmad Miremadi, Sarah Pinder, Ian Ellis, and Carlos Caldas. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biology*, 8(8):R157, 2007. ISSN 1465-6906. doi: 10.1186/gb-2007-8-8-r157. URL http://genomebiology.com/2007/8/8/R157.

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

Bibliography VII

M. J. van de Vijver, Y. D. He, L. van't Veer, H. Dai, A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25): 1999–2009, 2002.

L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhiven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Forekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365:671–679, 2005.

Bibliography VIII

Pratyaksha Wirapati, Christos Sotiriou, Susanne Kunkel, Pierre Farmer, Sylvain Pradervand, Benjamin Haibe-Kains, Christine Desmedt, Michail Ignatiadis, Thierry Sengstag, Frederic Schutz, Darlene Goldstein, Martine Piccart, and Mauro Delorenzi. 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, 2008. ISSN 1465-5411. doi: 10.1186/bcr2124. URL `http://breast-cancer-research.com/content/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

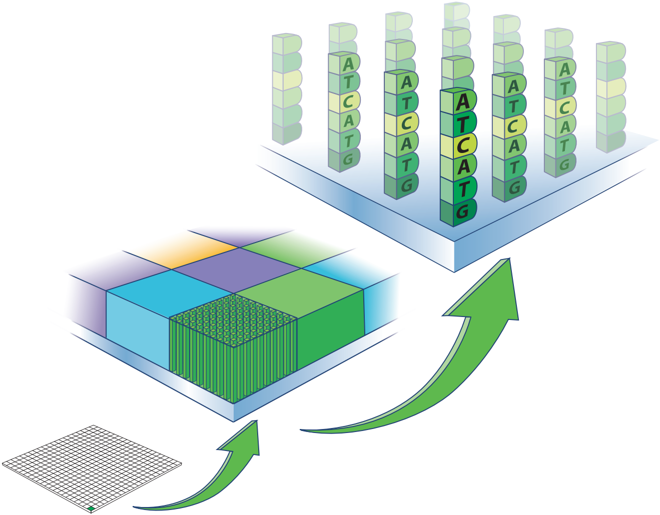
- A microarray is composed of
  - DNA fragments (*probes*) fixed on a solid support.
  - Ordered position of probes.
  - Principle of hybridization to a specific probe of complementary sequence.
  - Molecular labeling.

➡ Simultaneous detection of thousands of sequences in parallel.

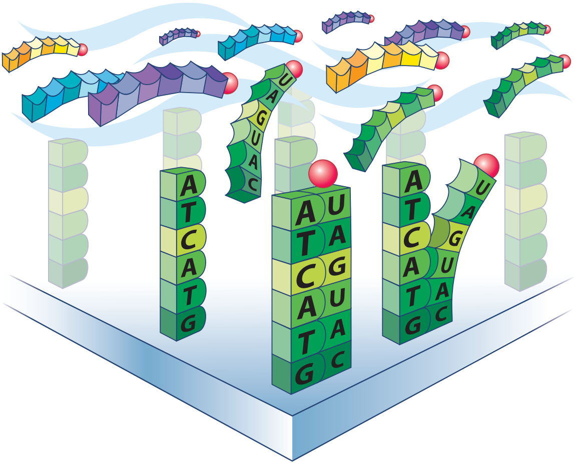
Affymetrix GeneChip

Affymetrix GeneChip

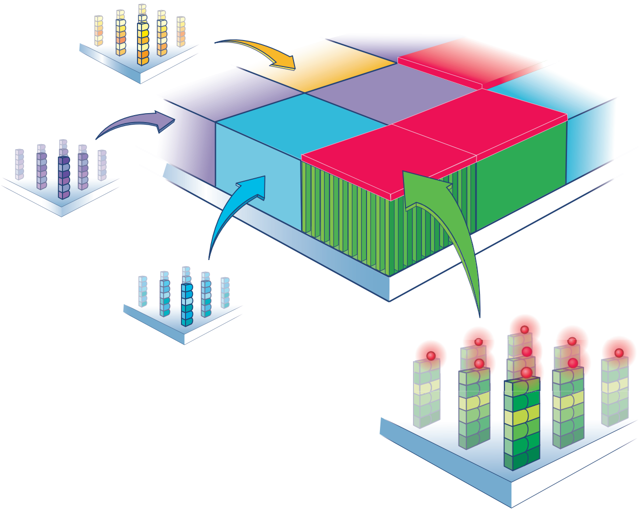
Probes



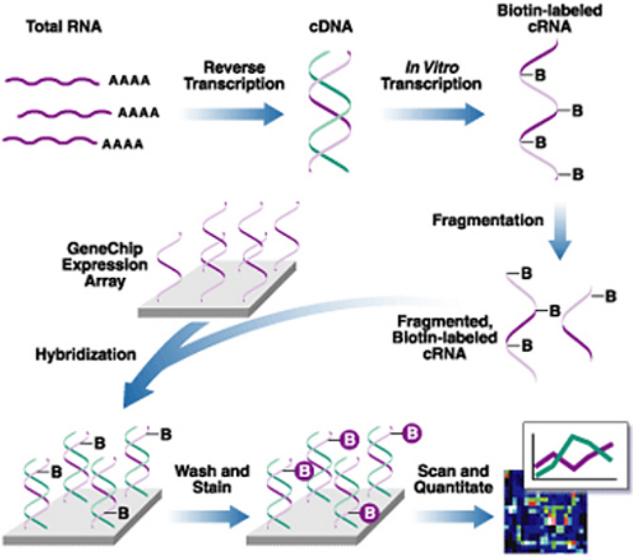
Hybridization



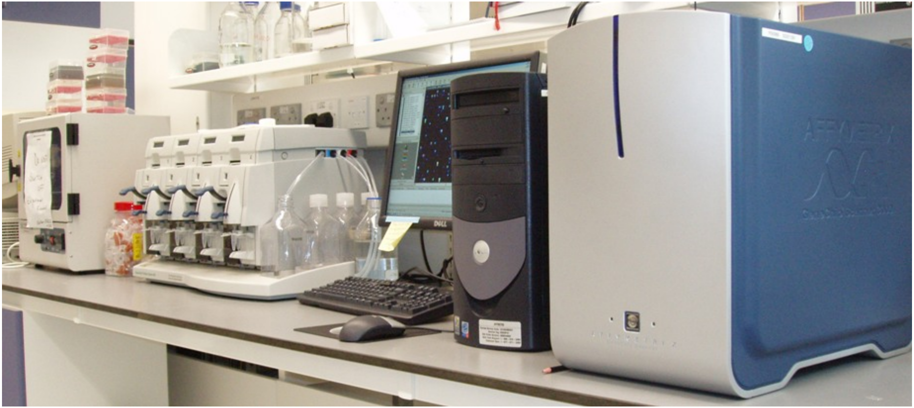
Detection



Affymetrix Design



Affymetrix Equipment



Prognostic Gene Signatures

A Single Gene?

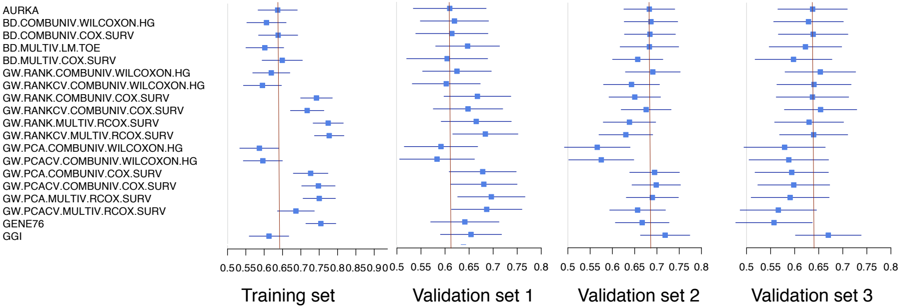
- From the validation studies, we learned that GGI yields similar (sometimes better) performance than other gene signatures [Haibe-Kains et al., 2008b].
- Since GGI is a very simple model from a statistical and a biological (proliferation genes) points of view, we challenged the use of complex statistical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

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

Prognostic Gene Signatures

A Single Gene? (cont.)

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



GENIUS

Identification of Subtypes

- The first step of GENIUS method is the identification of subtypes in the dataset.

- In BC, we applied the clustering model developed previously (training set: VDX).
- The model returns the probabilities $\Pr(s)$ for a patient to belong to each subtype $s \in S$.
  - S is composed of the ER-/HER2-, HER2+ and ER+/HER2- subtypes.

- We used a ranking-based gene selection method.
- The score (relevance) given to each gene is based on the significance of the concordance index.
- We introduced a weighted version of the concordance index in order to select genes relevant for a specific subtype;
- The weights were defined as the probability for a patient to belong to the subtype of interest.

➠ This feature selection allowed for using all the patients in the dataset.

- Survival data for the ith patient:
  - t_i stands for the event time
  - c_i for the censoring time
- C-index computes 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.

$$C\text{-index} = \frac{\sum_{i,j \in \Omega} 1\{r_i > r_j\}}{|\Omega|}$$

  - where r_i and r_j are the risk predictions of the patient i and j
  - Ω is the set of all the pairs of patients $\{i, j\}$ such that:
    - $r_i \neq r_j$ (no ties in r)
    - meet one of the following conditions: (i) both patients i and j experienced an event and time $t_i < t_j$ or (ii) only patient i experienced an event and $t_i < c_j$.

Weighted Concordance Index (cont.)

- We introduced a weighted version of the concordance

$$C\text{-index}_{wted} = \frac{\sum_{i,j \in \Omega} w_{ij} 1\{r_i > r_j\}}{\sum_{i,j \in \Omega} w_{ij}}$$

  - where $w_{ij} = w_i w_j$ is the weight for the pair of patients $\{i, j\} \in \Omega$.

- Significance of the C-index was computed by assuming asymptotic normality [Pencina and D'Agostino, 2004].

GENIUS

Signature Stability

- Once the genes were ranked, the only hyperparameter to tune was the signature size k (number of selected genes in the signature).
- We assessed the stability with respect to the signature size by resampling the training set.
- The stability criterion was inspired from [Davis et al., 2006]:
  - Let X be the set of features and $freq(x_j)$ be the number of sampling steps in which a feature $x_j \in X$ has been selected out of m sampling steps.
  - The set X is sorted by frequency into the set $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ where $freq(x_{(i)}) \geq freq(x_{(j)})$ if $i < j$ where $i, j \in \{1, 2, \ldots, n\}$.
  - A first measure of stability for a given signature size k is returned by

$$Stab(k) = \frac{\sum_{i=1}^{k} freq(x_{(i)})}{km}$$

GENIUS

Signature Stability (cont.)

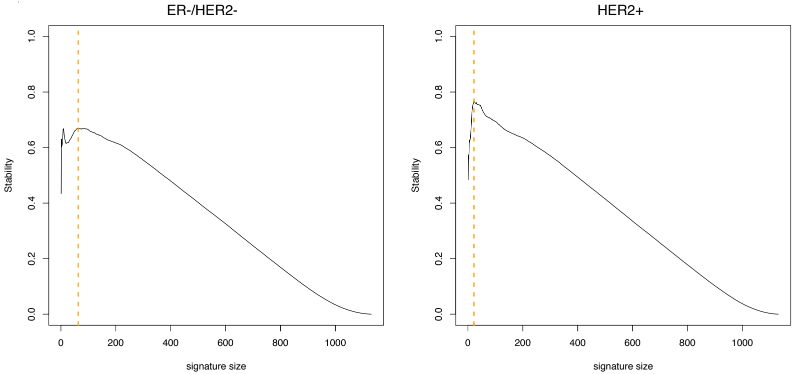
- Since the *Stab* statistic can be made artificially high by simply increasing k, we formulated an adjusted statistic

$$Stab_{adj}(k) = \max\left\{0, Stab(k) - \alpha\frac{k}{n}\right\}$$

  - where α is a penalty factor depending on the number of selected features (usually $\alpha = 1$).

GENIUS

Signature Stability (cont.)



- In the training set (VDX), the most stable signatures were composed of 63 and 22 genes for the ER-/HER2- and HER2+ subtypes.

- The risk score predictions for the subtype s is defined as

$$R(s) = \frac{\sum_{i \in Q} w_i x_i}{n_Q}$$

  - where Q is the set of of genes in the signature for subtype s
  - x_i is the expression of gene i
  - $w_i \in \{-1, +1\}$ depending on the concordance index (> 0.5 or ≤ 0.5)
  - n_Q is the signature size.

- The global risk score is defined as

$$R = \sum_{s \in S} \Pr(s) R(s)$$

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/