

Deriving and evaluating prognostic gene signatures from functional genomics data

Rory Stark

17 July 2019

Translating DEG for clinical utility

- How do we go from DEGs to something clinically useful?
- Marker genes
 - Normal/disease
 - Risk assessment
 - Molecular classification
- Gene signatures
 - Prognostic
 - Risk
 - Survival
 - Response
 - Diagnostic

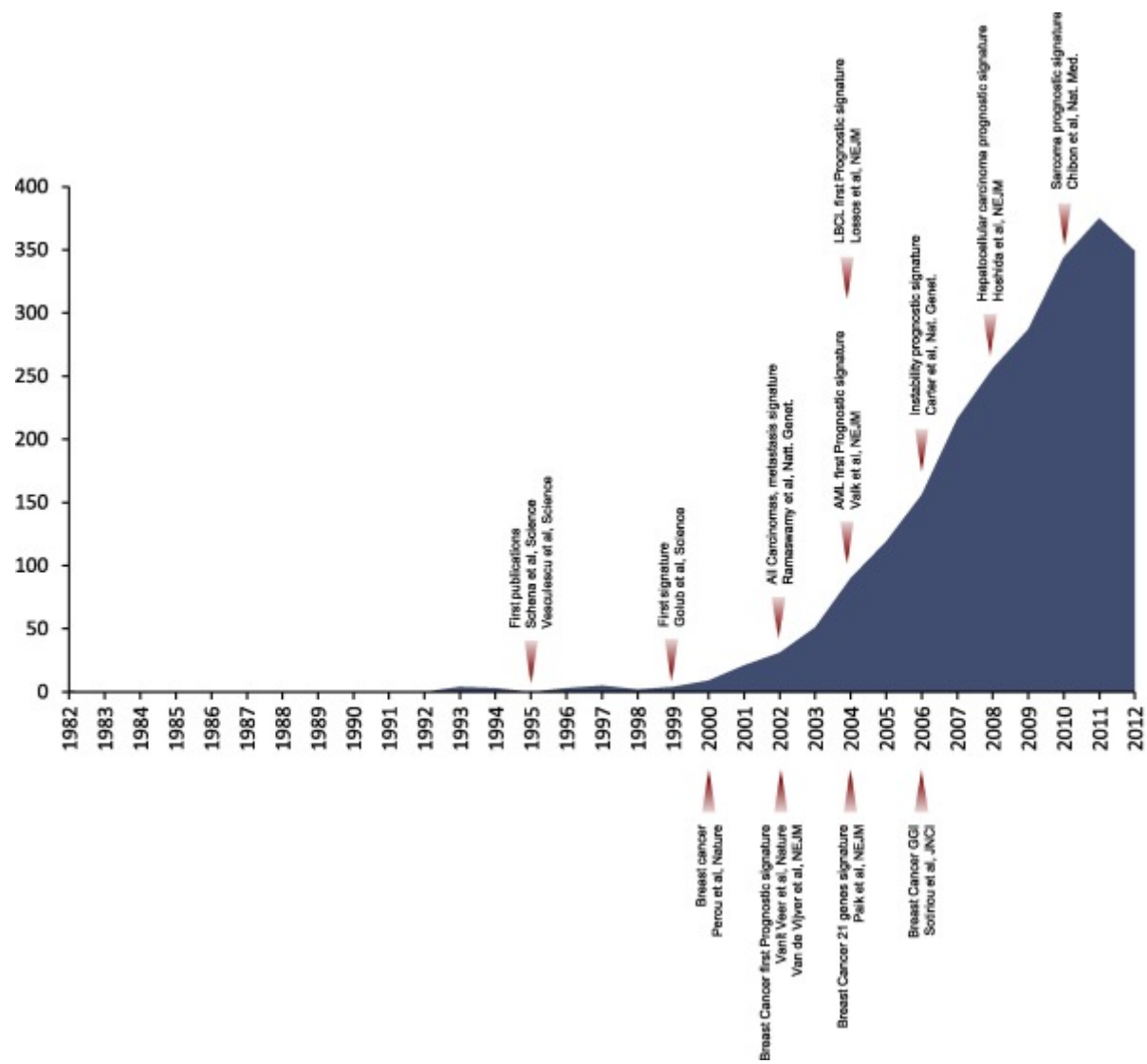


Fig. 1. Number of results (Y-axis) for the PubMed query: 'cancer gene expression signature' per year (X-axis). Red arrows indicate the year of some of the main publications in the field.

Steps in generating and evaluating a prognostic signature

- Samples
- Data
- Derivation
- Assessment
- Validation

Samples and Metadata

- Clinical (tumour) samples
 - Alternatives: signatures in “normal” tissue
- Metadata
 - Clinical
 - Tumour characteristics (size etc)
 - Grading
 - Outcome
 - Etc. Survival (alive/dead)
 - Recurrence
 - Metastasis
 - Demographic
 - Sex
 - Age
 - Environmental factors
 - Genotypes?
 - Technical
 - Collection info
 - Extraction info
 - When, who, where, how

Genomics Data

- mRNA Expression profiles
- Other genomic data:
 - Genotypes
 - Mutational signatures
 - CNV
 - TF Binding
 - Epigenomic Marks
 - Open chromatin
 - Histone marks

Signature derivation

- Unsupervised clustering
 - Identify clusters
 - Determine genes unique to each cluster
- Supervised Classification
 - Classify as high/low risk
 - Filter for variable genes
 - Possibly restrict to DE genes
 - Narrow down to “optimal” signature

Clustering (unsupervised)

- Molecular subtypes identified by clustering
- Subtypes may correlate with outcome/response

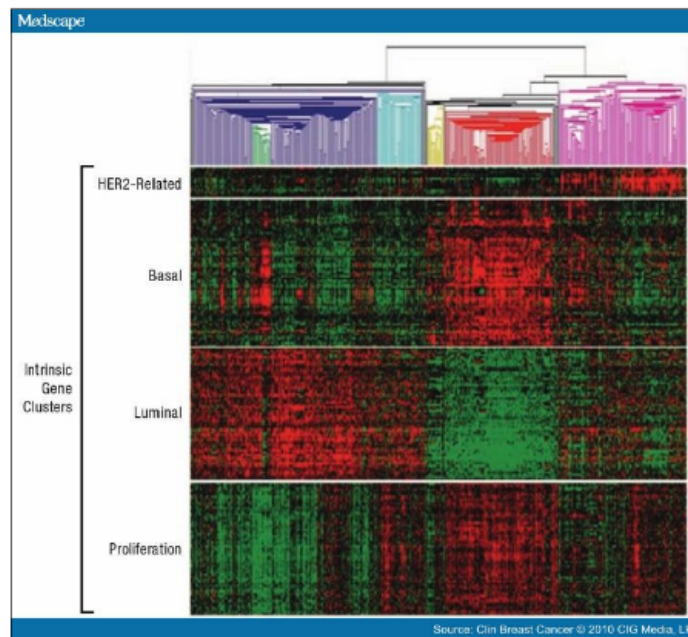
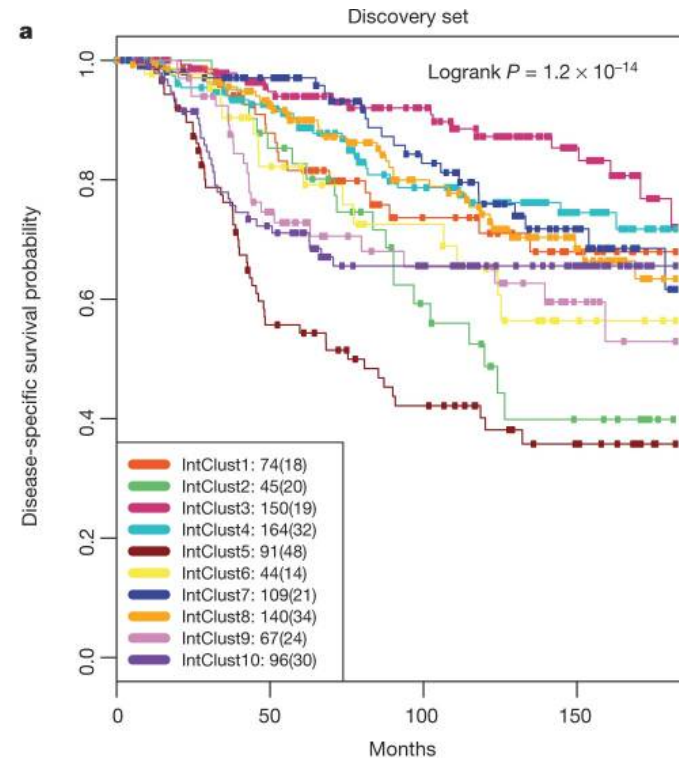


Figure 1.

Semi-Unsupervised Gene Expression Array Analysis of a Cohort of Breast Cancers Identifies Several Intrinsic Subtypes

Shown are luminal A (outlined in dark blue), luminal B (pale blue), HER2-enriched (pink), basal-like (red), claudin-low (yellow), and normal-like (green) tumors. Heat map courtesy of CM Perou.

Sorlie *et al* 2001




Curtis *et al* 2012


Supervised Classification

1. Assign each sample to high/low risk group
2. Split samples into training set / test set
3. Identify potentially informative genes in training set (e.g. DE analysis high vs low risk)
4. Find “optimal” set of DE genes that enable maximal separation of high/low risk samples
5. Validate in test set samples

Letter | Published: 31 January 2002

Gene expression profiling predicts clinical outcome of breast cancer

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

Nature **415**, 530–536 (2002) | [Download Citation](#) 

The New England Journal of Medicine

Copyright © 2002 by the Massachusetts Medical Society

VOLUME 347

DECEMBER 19, 2002

NUMBER 25



A GENE-EXPRESSION SIGNATURE AS A PREDICTOR OF SURVIVAL IN BREAST CANCER

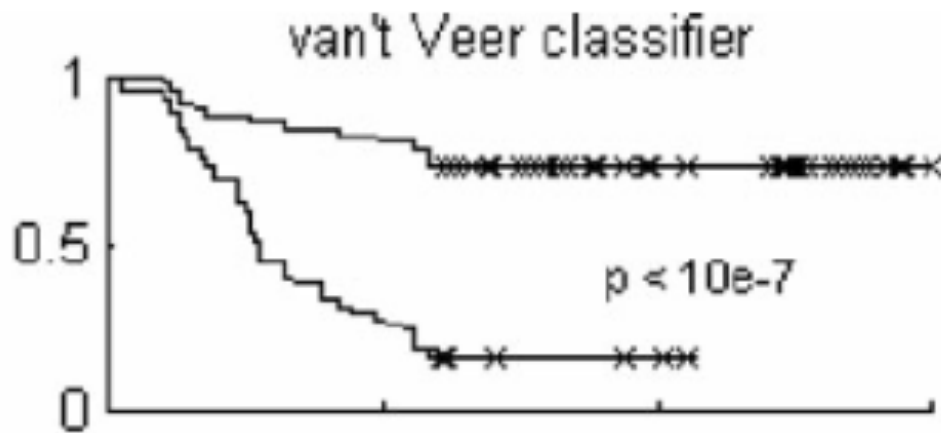
MARC J. VAN DE VIJVER, M.D., PH.D., YUDONG D. HE, PH.D., LAURA J. VAN 'T VEER, PH.D., HONGYUE DAI, PH.D.,
AUGUSTINUS A.M. HART, M.Sc., DORIEN W. VOSKUIL, PH.D., GEORGE J. SCHREIBER, M.Sc., JOHANNES L. PETERSE, M.D.,
CHRIS ROBERTS, PH.D., MATTHEW J. MARTON, PH.D., MARK PARRISH, DOUWE AT SMA, ANKE WITTEVEEN,
ANNUSKA GLAS, PH.D., LEONIE DELAHAYE, TONY VAN DER VELDE, HARRY BARTELINK, M.D., PH.D.,
SJOERD RODENHUIS, M.D., PH.D., EMIEL T. RUTGERS, M.D., PH.D., STEPHEN H. FRIEND, M.D., PH.D.,
AND RENÉ BERNARDS, PH.D.

van't Veer method

1. Assign samples to risk group based on distant metastases within five years (34 high/44 low)
2. Use all samples as training set (78 samples)
3. Identify ≈ 5000 “significantly regulated” genes
4. Rank genes by correlation with risk group; retain genes with $R > 0.3$ (231 genes)
5. Leave-one-out-cross-validation (78 fold) on progressively larger gene sets (rank order)
6. Classifier: correlation with low risk template (average expression)
7. Best performance on top 70 genes (MammaPrint)

Signature assessment

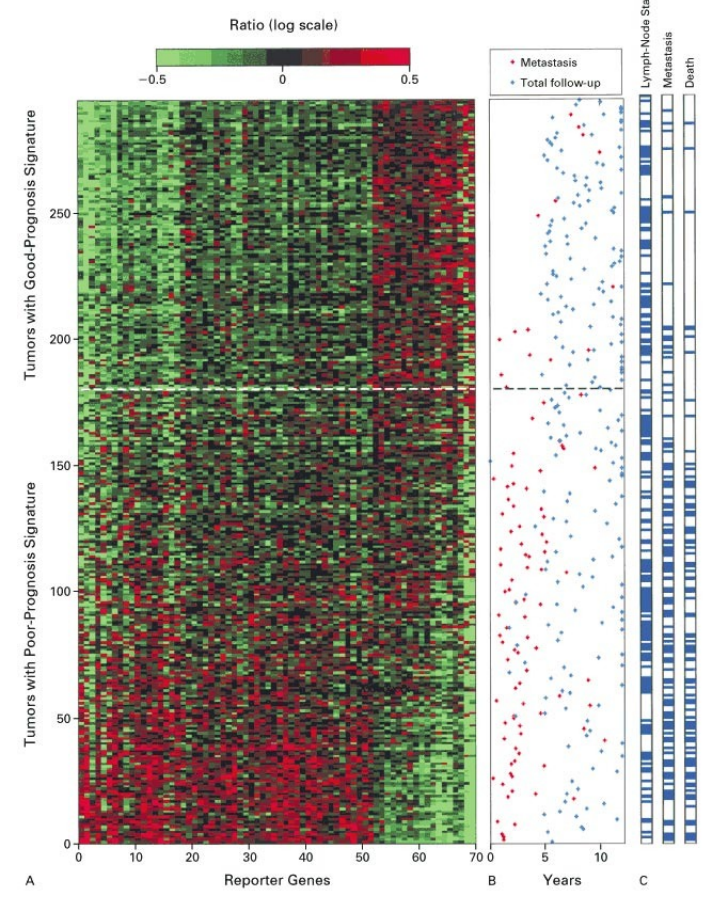
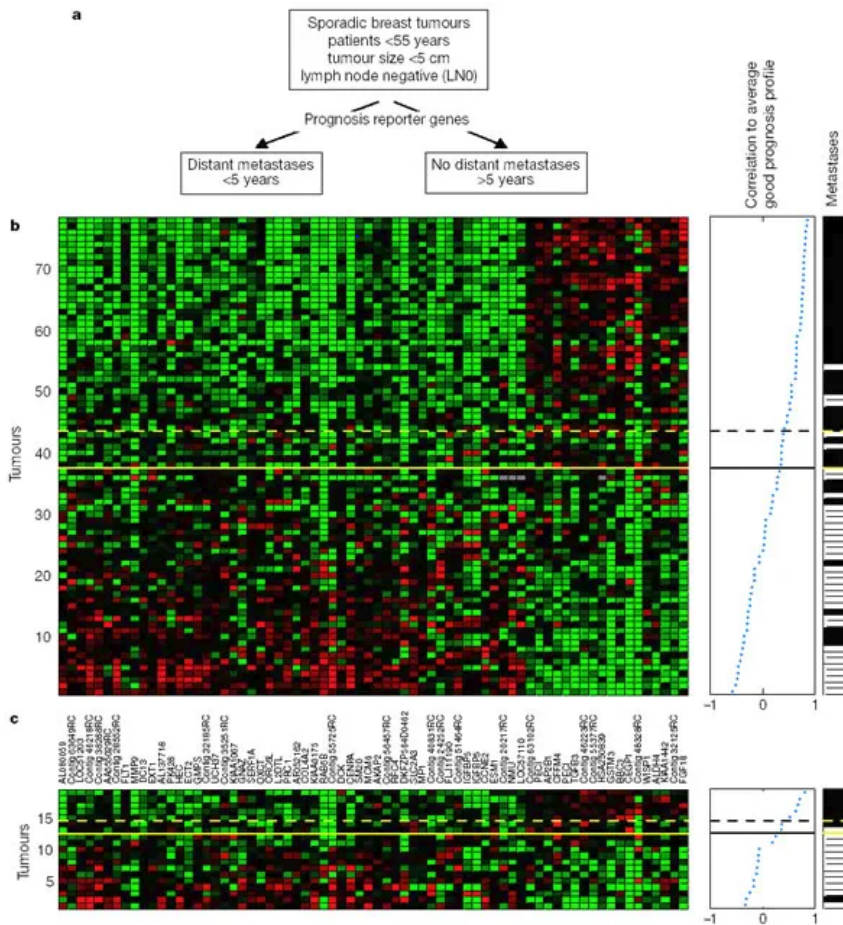
- Survival Analysis
- Kaplan-Meier Plots
- p-value (Cox proportional hazard, log-rank test)



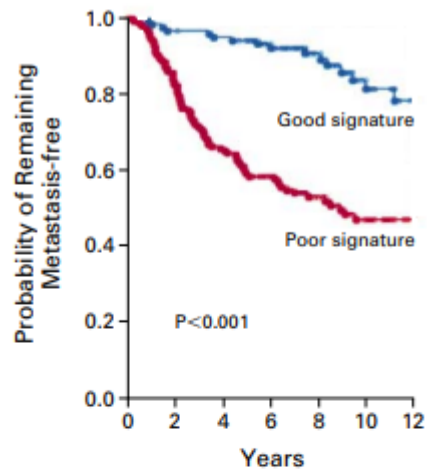
- cf Dom's talk next

Signature validation

- Evaluate in unseen samples
- Evaluate in external data
 - Different locations
 - Different experimenters
 - Different platforms
- Evaluate in new patients



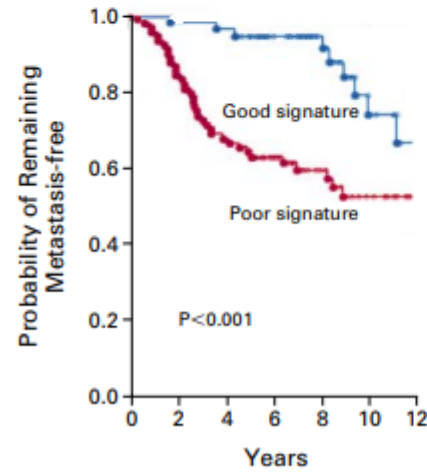
A All Patients



No. AT RISK

Good signature	115	111	107	87	59	36	19
Poor signature	180	146	111	84	52	33	17

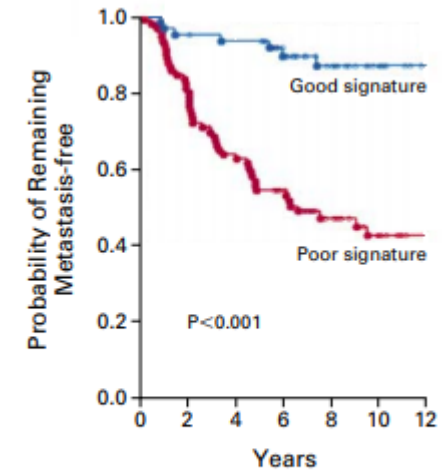
E Lymph-Node-Positive Patients



No. AT RISK

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

C Lymph-Node-Negative Patients



No. AT RISK

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

Methods for signature derivation

- Narrow gene based on differential expression
- Train ML classifiers using cross-validation on different gene sets
- Integration of other data types
- Pathway or other higher level biological functions