You've reproduced the Figure

BUT can you validate it in an independent study… maybe not

Problem 1.

- Poor analysis methods
- Reviewed in Detail by Richard Simon in JCNI and other reviews

Simon R: Proper Cross validation

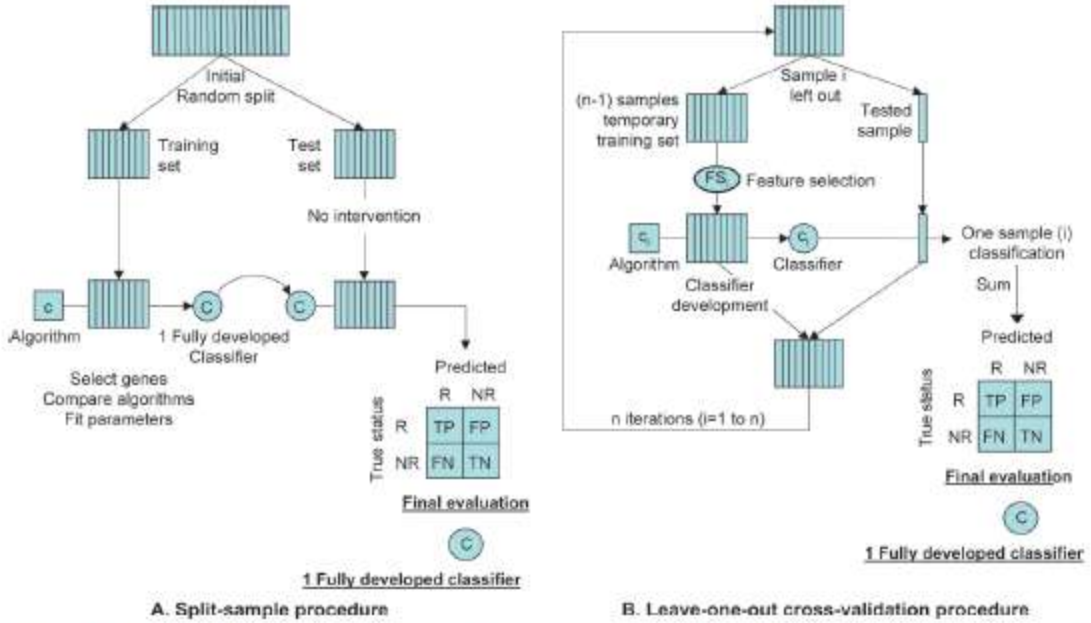


Fig. 2. Developing and validating a classifier. The classifier is for a binary outcome such as response (R) or nonresponse (NR) to a treatment. The final result is composed of two parts: a contingency table presenting the results of the validation (TP = true positive; TN = true negative; FP = false positive; FN = false negative) and the fully parametrized classifier that can be used to predict the outcome in new samples with unknown status. **A**) Split-sample procedure. A random split divides the initial dataset into a training set and a test set. The classifier is developed on the training set. Once the classifier has been fully specified, the test set is accessed once and only for estimating the prediction accuracy of the classifier. The two steps of developing (on the training set) and evaluating (on the test set) are physically and temporally distinct. The information from the test set samples has never been used in any prior step of data handling. **B**) Leave-one-out cross-validation procedure. From the initial n-sample dataset, one sample is withdrawn, leaving a temporary (n − 1)-sample training set and one left-out test sample. On the training set, a group of outcome-related genes is selected (FS = feature selection). The expression data values from these selected genes are used to parametrize the classification algorithm (c in square = classification algorithm before parametrization; C in circle = parametrized classifier). The parametrized classifier is then used to classify the previously left-out test sample as "responder" or "nonresponder." These steps are reiterated n times, until each sample has in turn been left out once for testing. Subscripts i denote that the genes selected (feature selection, FSi) or the parameters of the classifier (Ci) are different at each iteration of the cross-validation.

Case Study: Re-Analysis of the Van 't Veer Breast Cancer Study

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‡

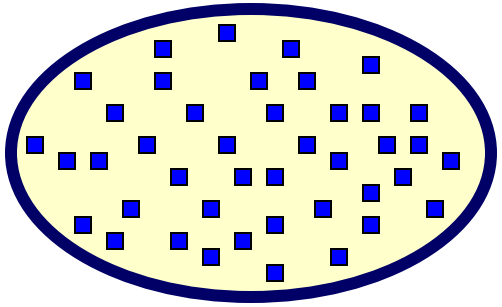
** Divisions of Diagnostic Oncology, Radiotherapy and Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands*
‡ Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034, USA
† These authors contributed equally to this work

- 97 young patients
- Sporadic breast tumours
- patients <55 years
- tumour size <5 cm
- All lymph node negative
- Either ER+/-

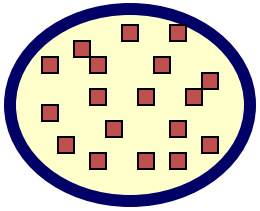
Examined gene expression with respect to: "Interval (5 year) to distant metastases"

Van ‘tVeer Data

78 Training Samples



19 Test Samples



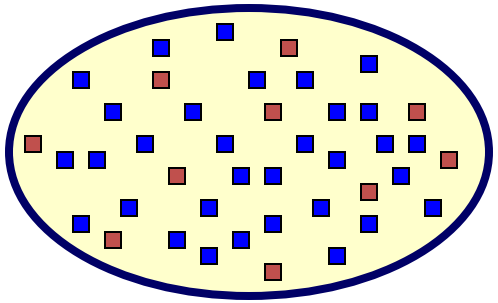
Dataset divided into:

- Training
  - 34 bad prognosis
  - 44 good prognosis
- Test
  - 12 bad prognosis
  - 7 good prognosis

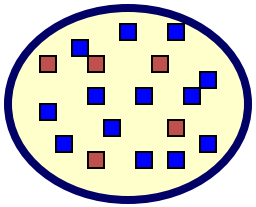
- 24k genes filtered according to Van ‘t Veer criteria to 5k genes

Repeating BGA with Different subsets of Training/Test Data

69 Training Samples



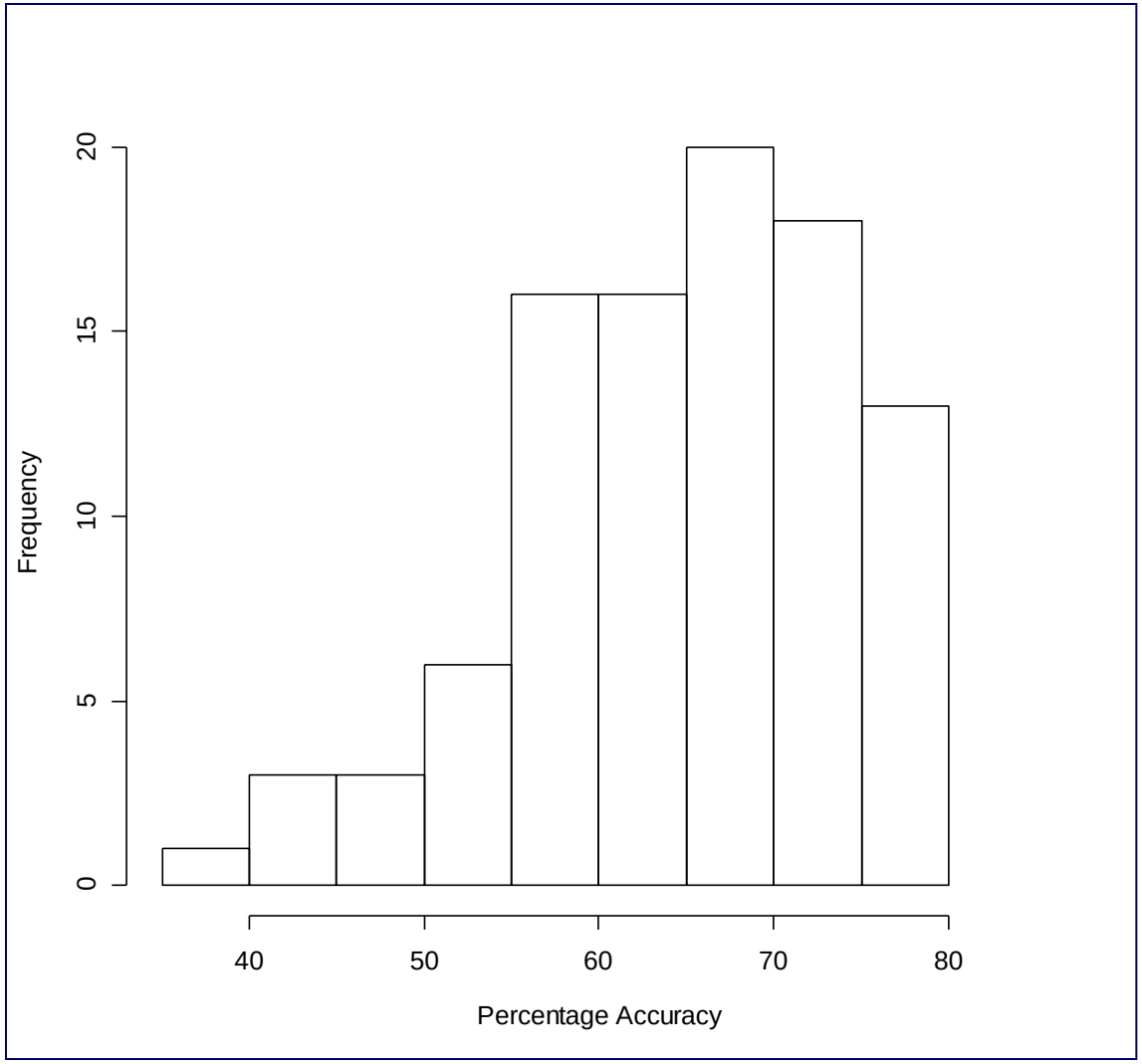
17 Test Samples



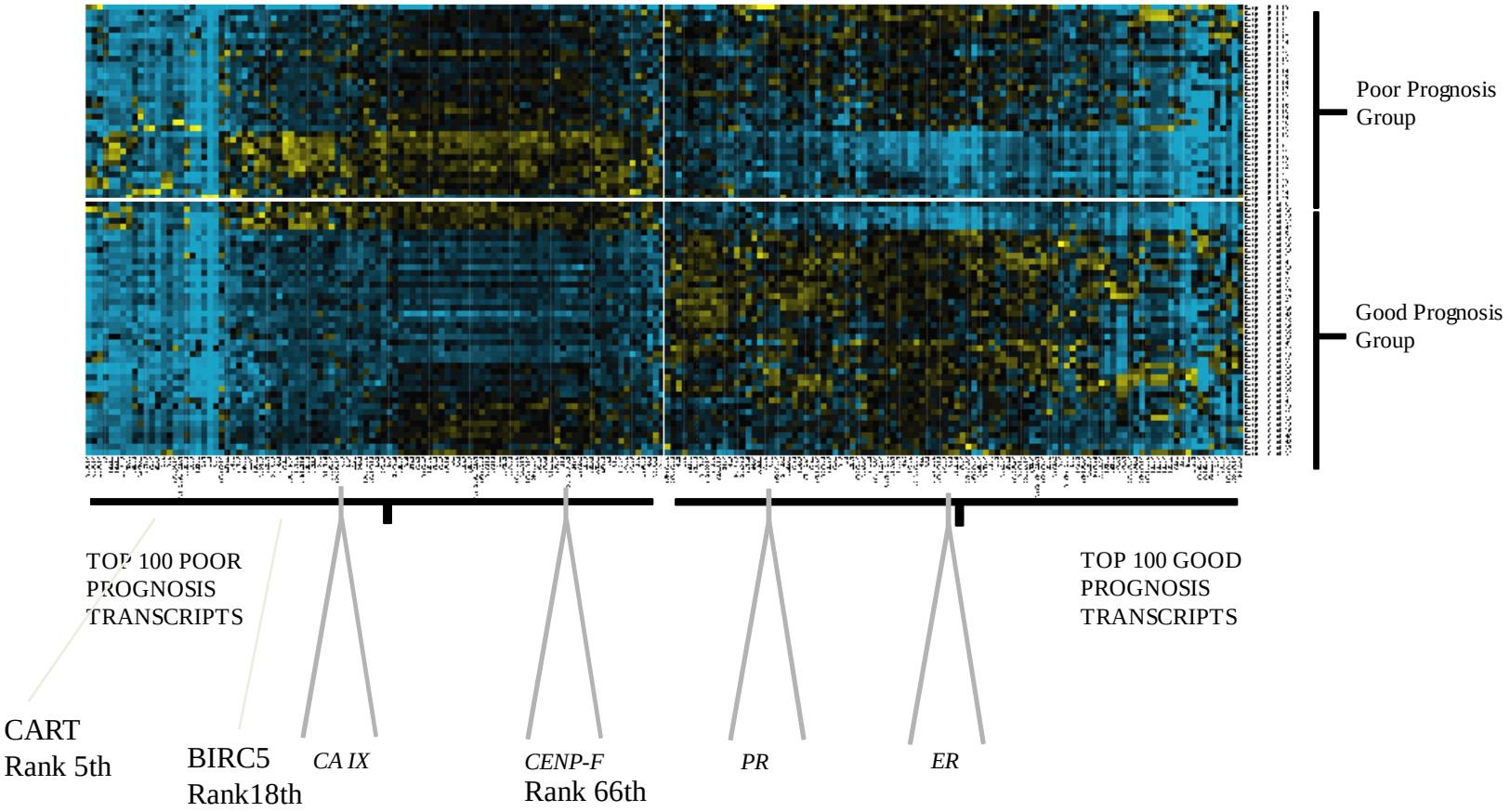
- Randomised the training and test (in case of bias)
- Applied Classifier (BGA*)
- Should get same/improved result (83% accuracy)?

* Culhane et al., Bioinformatics 2002.

Histogram of BGA accuracy – Randomised Data (n=100)

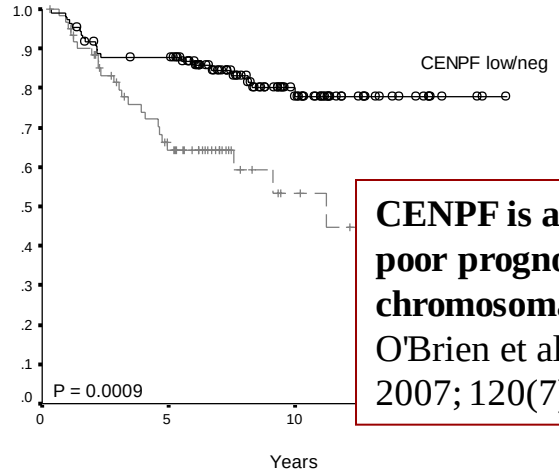


Discriminating Genes

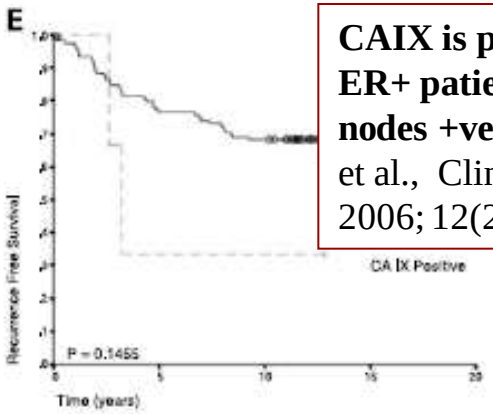


O'Brien et al., . Int J Cancer. 2007; 120(7):1434-43.

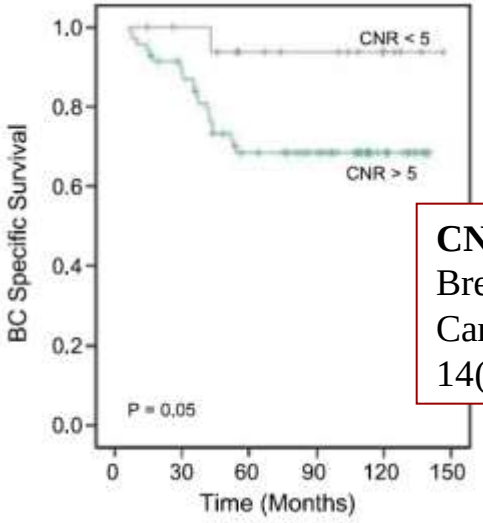
Validated Protein Expression on Tissue Microarrays



CENPF is associated with poor prognosis and chromosomal instability. O'Brien et al., . Int J Cancer. 2007; 120(7):1434-43.

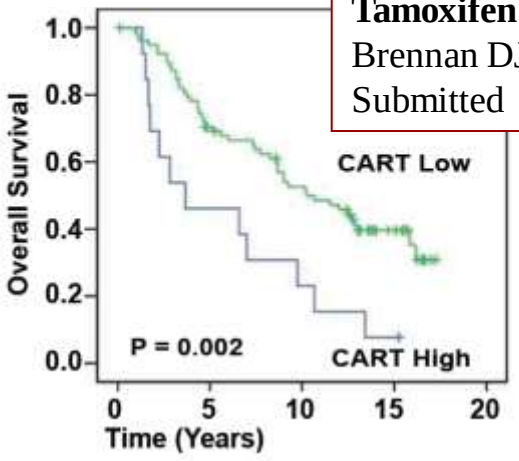


CAIX is predictive in ER+ patients with 1-3 nodes +ve. Brennan DJ, et al., Clin Cancer Res. 2006; 12(21):6421-3



CNR of Survivin (BIRC5) Brennan DJ, et al., Clin Cancer Res. 2008; 14(9):2681-9

CART is prognostic in Node-ve and Predicts Tamoxifen Response. Brennan DJ et al., JNCI. Submitted



Problem 2

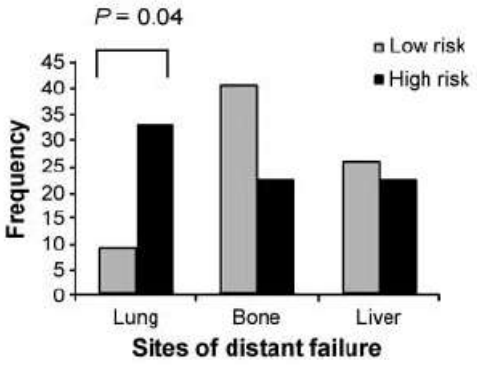
- Poor experimental Design
- Confounding Covariates

Case Study 2:
A 6 gene signature of lung metastasis

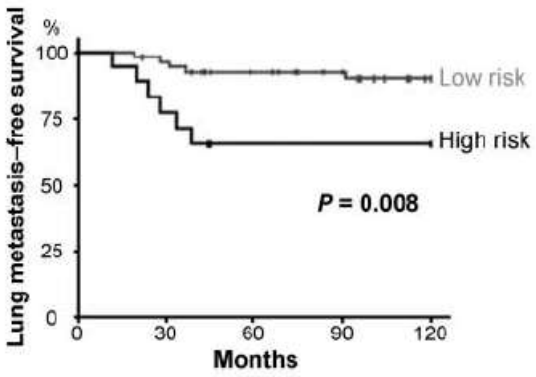
A

Characteristics	All patients (*n*=72)	High risk (*n*=18)	Low risk (*n*=54)	*P*
Metastasis	38 (53%)	10 (56%)	28 (52%)	-
Lung metastasis	11 (15%)	6 (33%)	5 (9%)	0.04
Postmenopause	40 (69%)	9 (60%)	31 (72%)	-
Macroscopic tumor size (>20 mm)	47 (67%)	13 (72%)	34 (65%)	-
Grade 3 (SBR)	18 (29%)	9 (56%)	9 (20%)	0.01
Estrogen receptor negative	28 (39%)	15 (83%)	13 (24%)	<0.001
Progesterone receptor negative	35 (49%)	14 (78%)	21 (39%)	0.004

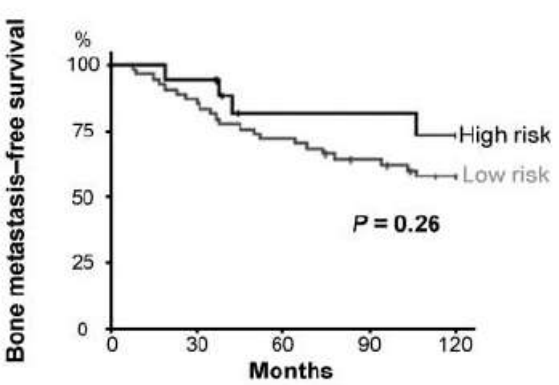
B



C



D



Landemaine T et al., Cancer Res. 2008 Aug 1;68(15):6092-9.

But metastatic profile of breast cancer differs by tumor subtype

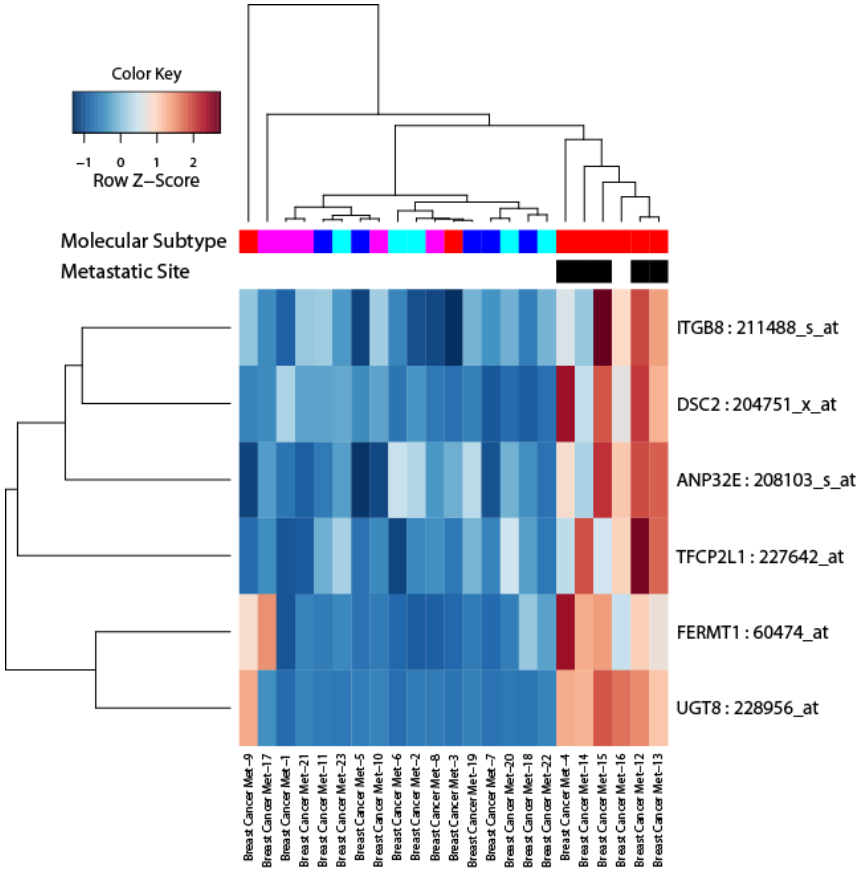
Table 1. Frequencies of site of relapse in the molecular subtypes

Subtype	Site of relapse					
	Bone	Lung	Liver	Brain	Pleura	Total
Luminal B	26 (36.6)	11 (36.7)	2 (11.1)	1 (7.1)	5 (41.7)	45
Luminal A	22 (31.0)	2 (6.7)	4 (22.2)	1 (7.1)	5 (41.7)	34
Erbb2	14 (19.7)	4 (13.3)	6 (33.3)	3 (21.4)	0 (0.0)	27
Normal	4 (5.6)	1 (3.3)	2 (11.1)	1 (7.1)	1 (8.3)	9
Basal	5 (7.0)	12 (40.0)	4 (22.2)	8 (57.1)	1 (8.3)	30
Total	71	30	18	14	12	145

NOTE: Numbers between parentheses are column percentages, e.g., 36.6% of bone relapses are in the luminal B subtype.

Smid et al., 2008 Cancer Res 68(9):3108–14

Confounding Covariates



Confounding Covariates

Supplementary Table 4. Results of Analysis of Global Test and GlobalAncova analysis of MSK dataset (p-value)

Method	globaltest	globaltest	GlobalAncova	GlobalAnova
Number of Probesets tested *	4	10	4	10
Q1: Are the genes associated with metastases status?	0.048	0.100	0.015	0.023
Q2: Are the genes associated with molecular subtype ?	<0.00000001	<0.0000001	0	0
Q3: Is metastases status significant independent of molecular subtype?	0.720	0.694	0.630	0.696
Q4: Is molecular subtype significant independent of metastases status ?	<0.000001	<0.000001	0	0
Q5: Are the genes associated with metastases status in the basal-like tumors?	0.514	0.168	0.380	0.190

Culhane AC & Quackenbush J. 2009 Cancer Research (In Press)

CASE STUDY 2:

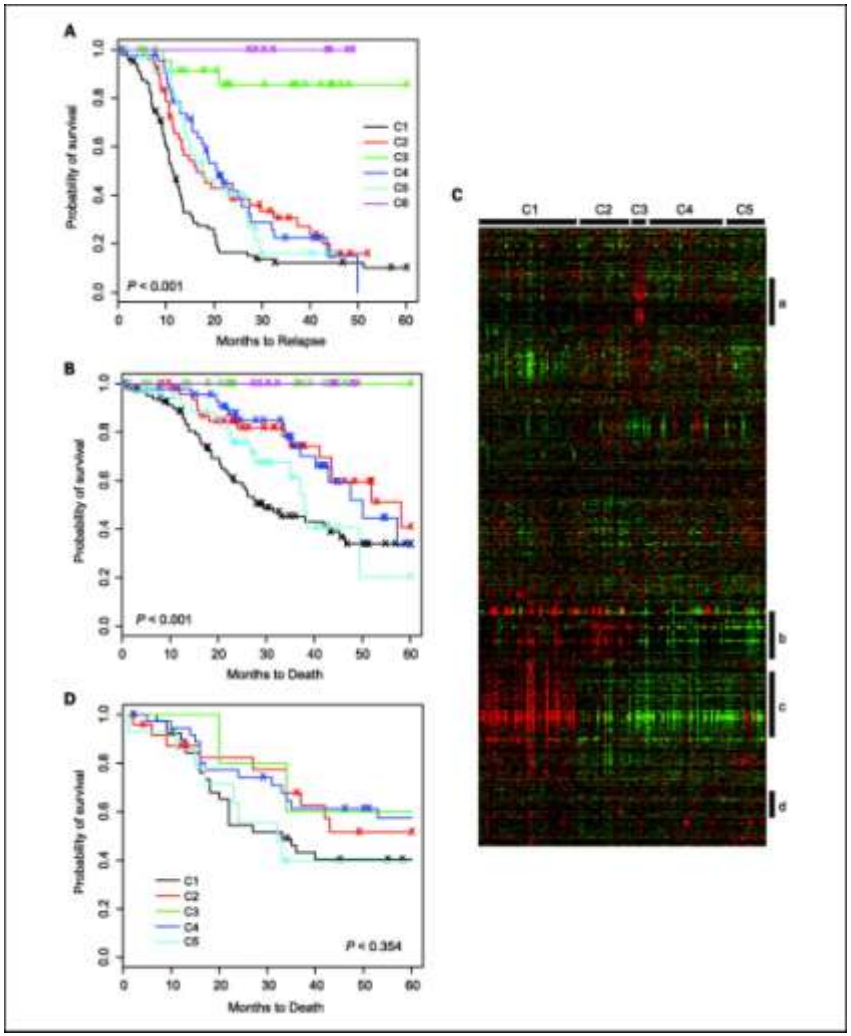
Non-tumor cell contamination confounds prognostic subtype discovery in ovarian cancer

Matthew Schwede, David Harrington, Melissa Merritt, John Quackenbush, Aedín C. Culhane

Molecular Subtypes of Ovarian Cancer

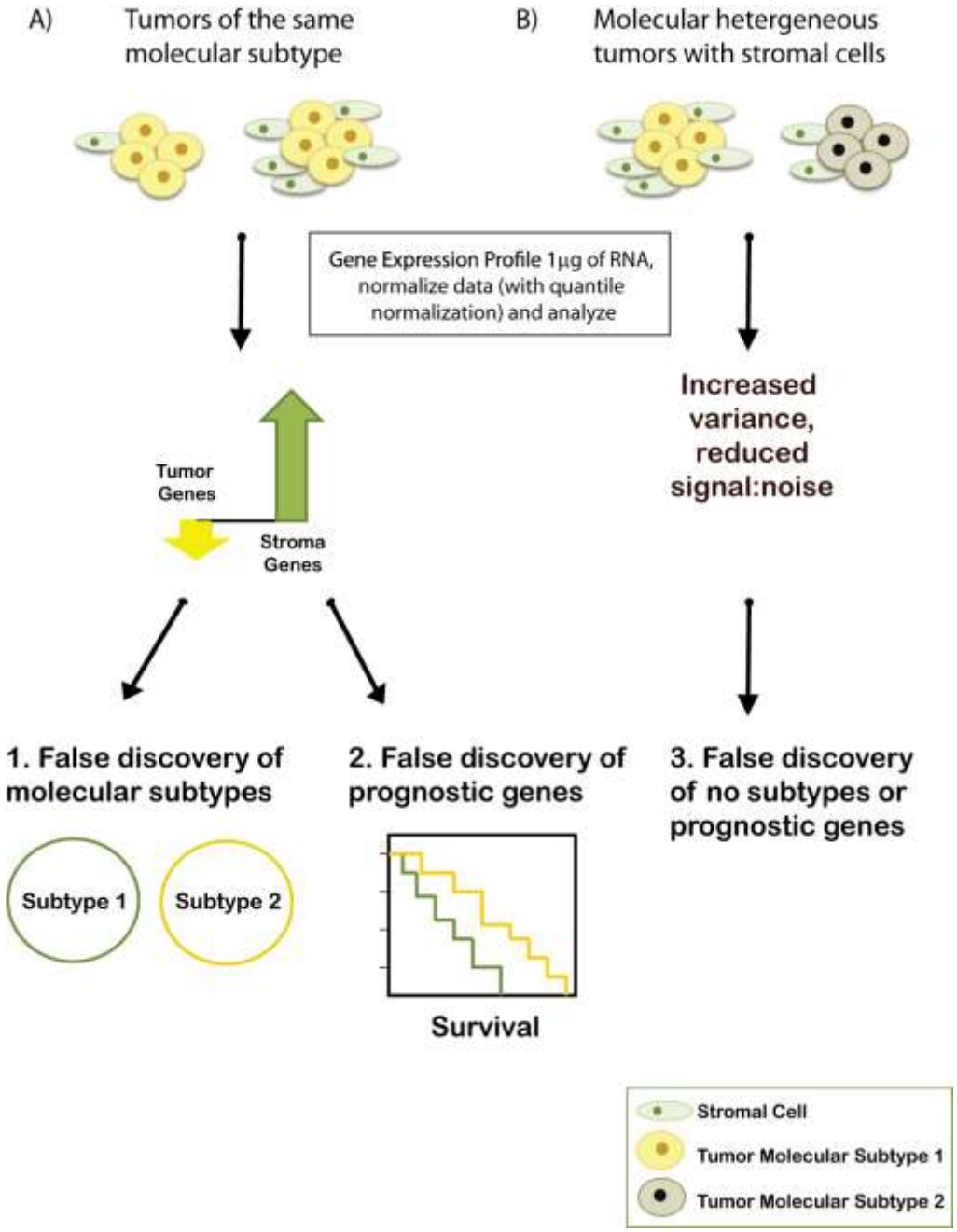
6 molecular subtypes

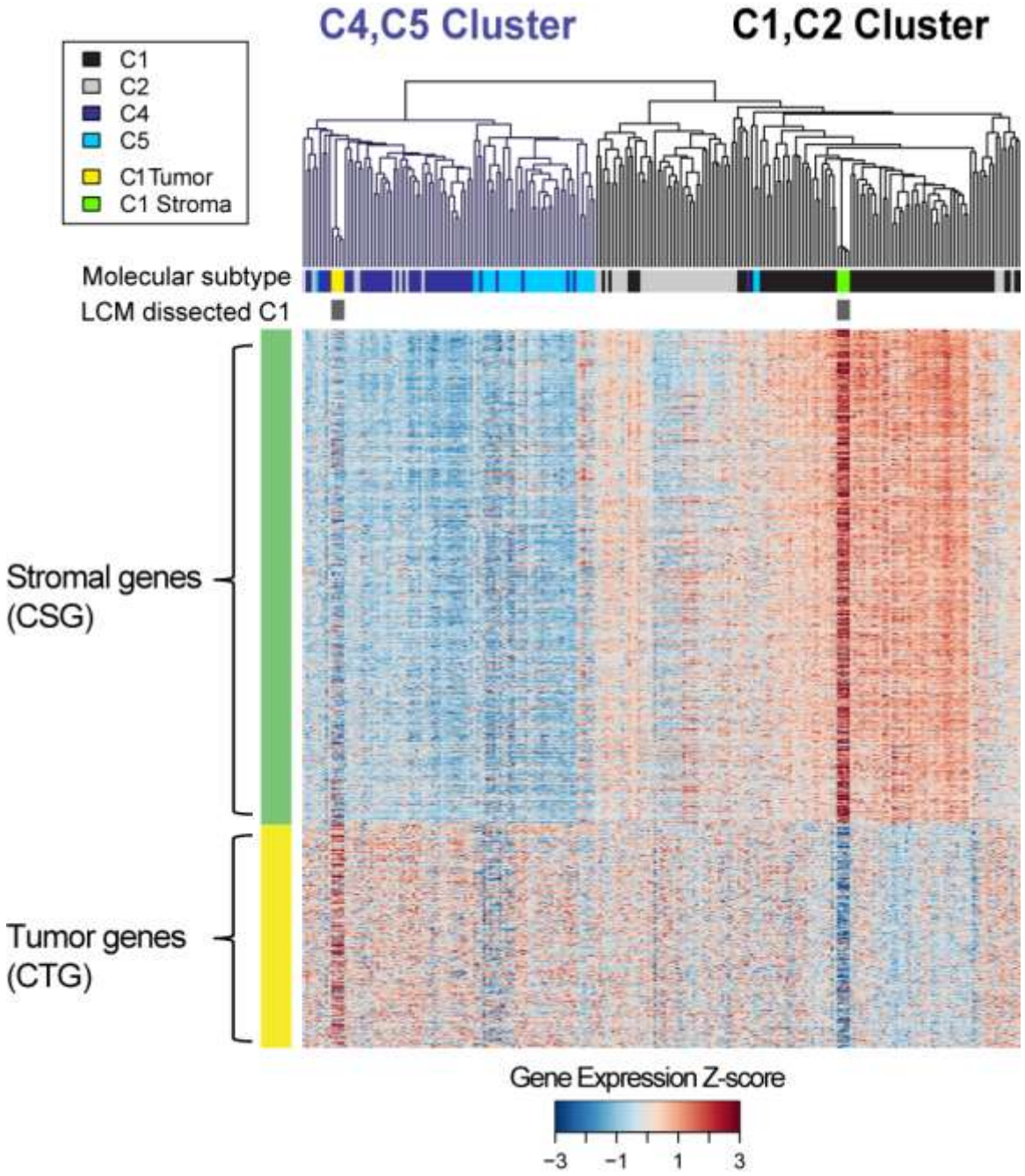
C1 – most prognostic in high grade serous ovarian cancer



Tothill R W et al. Clin Cancer Res 2008;14:5198-5208

Clinical Cancer Research


A)
Tumors of the same molecular subtype
B)
Molecular hetergeneous tumors with stromal cells
Gene Expression Profile 1μg of RNA, normalize data (with quantile normalization) and analyze
Increased variance, reduced signal:noise
Tumor Genes
Stroma Genes
1. False discovery of molecular subtypes
2. False discovery of prognostic genes
3. False discovery of no subtypes or prognostic genes
Subtype 1
Subtype 2
Survival
Stromal Cell
Tumor Molecular Subtype 1
Tumor Molecular Subtype 2


C4,C5 Cluster
C1,C2 Cluster
C1
C2
C4
C5
C1 Tumor
C1 Stroma
Molecular subtype
LCM dissected C1
Stromal genes
(CSG)
Tumor genes
(CTG)
Gene Expression Z-score
−3
−1
1
3

Published gene signature		Total # genes	Overlap with CSGs and CTGs		Ovarian % stroma	Breast stroma vs. epithelium		Prostate % non-tumor content	
Study	Signature		# CTGs (%)	# CSGs (%)	TCGA (n = 518)	Boersma (n = 34)	Casey (n = 28)	Wang (n = 109)	Wang (n = 136)
Current study – CTG, CSG	C1 tumor genes C1 stromal genes	227 461	227 0	0 461	5.6×10^{-13} (+)	4.3×10^{-8} (+)	7.1×10^{-8} (+)	1.5×10^{-9} (+)	4.9×10^{-17} (+)
AOCS (Tothill et al.)	Down in C1 Up in C1	147 287	2 (1) 0	2 (1) 176 (61)*	3.8×10^{-11} (+)	2.8×10^{-8} (+)	5.1×10^{-7} (+)	4.9×10^{-8} (+)	1.1×10^{-14} (+)
	Good PFS Poor PFS	143 135	16 (11)* 0	0 103 (76)*	8.6×10^{-13} (+)	4.0×10^{-9} (+)	4.1×10^{-8} (+)	1.2×10^{-9} (+)	2.0×10^{-17} (+)
	Good OS Poor OS	146 147	23 (16)* 0	0 101 (69)*	5.3×10^{-11} (+)	4.6×10^{-9} (+)	9.7×10^{-8} (+)	1.7×10^{-5} (+)	6.5×10^{-13} (+)
Bentink et al.	Non-angiogenic Angiogenic	19 74	0 0	1 (5) 35 (47)*	1.6×10^{-9} (+)	1.5×10^{-7} (+)	0.93 (+)	1.1×10^{-6} (+)	6.8×10^{-14} (+)
Bignotti	Up in primary tumor Up in metastasis	36 89	1 (3) 0	0 54 (61)*	1.2×10^{-9} (+)	4.5×10^{-7} (+)	5.2×10^{-5} (+)	7.2×10^{-4} (+)	7.2×10^{-8} (+)
Spentzos et al.	Favorable prognosis Unfavorable prognosis	43 73	4 (9)* 0	0 19 (26)*	1.7×10^{-10} (+)	1.8×10^{-5} (+)	0.0074 (-)	4.1×10^{-14} (+)	1.0×10^{-7} (+)
Bonome et al.	Good prognosis Poor prognosis	272 288	5 (2) 0	1 37 (13)*	9.4×10^{-11} (+)	3.5×10^{-6} (+)	1.5×10^{-6} (+)	7.7×10^{-5} (+)	3.0×10^{-11} (+)
Biade et al.	Benign cluster Malignant cluster	21 15	1 4 (27)*	6 (29)* 0	1.1×10^{-12} (-)	9.8×10^{-8} (-)	5.2×10^{-6} (-)	6.5×10^{-10} (-)	7.1×10^{-13} (-)
Konstantino-poulos et al.	BRCA-like Non-BRCA-like	32 27	0 0	2 (6) 2 (7)	0.729 (-)	0.0043 (-)	7.6×10^{-6} (-)	0.34 (+)	0.19 (+)

Published gene signature		TCGA dataset multivariate analysis p-values					
Study	Signature	OS	OS, adjust stroma	OS, adjust stroma + stage	RFS	RFS, adjust stroma	RFS, adjust stroma + stage
Current study	CSGs vs. CTGs	0.087	0.27	0.48	0.41	0.61	0.88
AOCS (Tothill et al.)	C1	0.018	0.076	0.28	0.13	0.2	0.45
	PFS	0.0045	0.035	0.15	0.094	0.15	0.34
	OS	0.00099	0.0076	0.054	0.081	0.16	0.41
Bentink et al.	Angiogenic vs. non-angiogenic	0.013	0.054	0.17	0.053	0.1	0.27
Bignotti et al.	Metastasis vs. primary	0.017	0.061	0.31	0.096	0.14	0.4
Spentzos et al.	Prognosis	0.043	0.13	0.21	0.43	0.51	0.49
Bonome et al.	Prognosis	0.031	0.11	0.37	0.12	0.17	0.39
Biade et al.	Malignant vs. benign	0.31	0.59	0.58	0.22	0.39	0.31
Konstantino-poulos et al.	BRCAness	0.018	0.029	0.037	0.013	0.013	0.011

Experimental Design Pop Quiz

Example: benzopyrene toxicity

- Study: toxic effect of Benzo(a)pyrene on rats
- 8 rats are to be treated with BP and 8 rats with a control compound
- Each array will be hybridised against a reference sample
- 16 arrays in experiment

Experimental Design

- There are 2 batches of 8 slides, from 2 different print runs
- The hybridisation will be done by 2 different researchers, Alison and Brian
- What is the best way to arrange the experiment?

Design 1

- Alison prepares all 8 BP samples and hybridises them to the arrays of print run 1

- Brian prepares all 8 control samples and hybridises them to the arrays of print run 2

Design 2

- Alison chooses 8 rats and treats 4 with BP and 4 with control substance
- She prepares and hybridises 2 BP samples to arrays from print run 1 and 2 BP samples to arrays from print run 2
- She prepares and hybridises 2 control samples to arrays from print run 1 and 2
- Brian does the same with the other 8 rats

Design 3

- 8 rats are randomly assigned to Alison, along with 4 BP preps and 4 control preps - she is not told which preps are which
- She prepares and hybridises samples to randomly prearranged arrays so that 2 BP samples and 2 control samples are hybridised to 4 arrays from each of print runs 1 and 2
- Brian does the same with the other 8 rats

What is wrong with design 1?

- Treatment, researcher and print run are CONFOUNDED variables
- We cannot tell whether differences between the two groups of rats result from treatment, researcher or print run
- Use blocking, in designs 2 and 3 to deconfound the variability of interest (treatment) from the extraneous variables
- Designs 2 and 3 are also BALANCED, which increases the power of the analysis

What is wrong with Design 2?

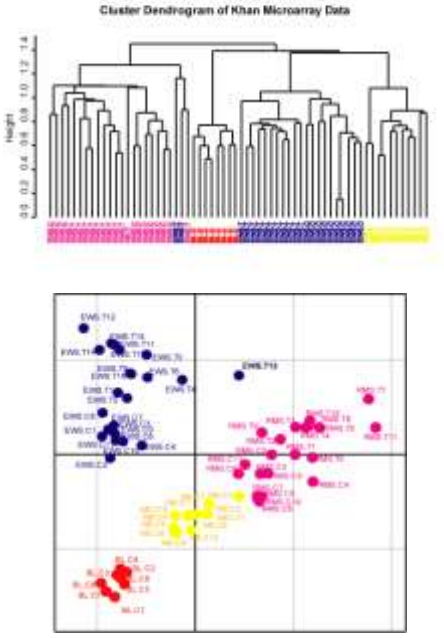
- Alison's choice of rats may be BIASED
- e.g. she may choose the healthiest rats, confounding potential treatment effects with researcher variability
- Use randomisation and blinding in design 3 to avoid bias

Blocking, Randomization and Blinding

- Arrangement of experimental design that minimises problems from extraneous sources of variability
- Use blocking to avoid CONFOUNDING (extraneous variables)
- Use randomisation and blinding to avoid BIAS

Exploration of Data is Critical

- Detect unpredicted patterns in data
- Decide what questions to ask
- Can also **help detect cofounding covariates**



Good Experimental Design & Sample Processing is Critical

Dr. Frederick Frankenstein: Igor, would you mind telling me whose brain I did put in?

Igor: And you won't be angry?

Dr. Frederick Frankenstein: I will NOT be angry.

Igor: Abby someone.

Dr. Frederick Frankenstein: Abby someone. Abby who?

Igor: Abby Normal.

Dr. Frederick Frankenstein: Abby Normal?

Igor: I'm almost sure that was the name.

Dr. Frederick Frankenstein: Are you saying that I put an abnormal brain into a seven and a half foot long, fifty-four inch wide GORILLA? IS THAT WHAT YOU'RE TELLING ME?

From the film
Young Frankenstein, 1974

http://www.youtube.com/watch?v=NOe_4mgmyyA

Please feel free to contact me

aedin@jimmy.harvard.edu