# Gene expression predictors for breast cancer outcome reflect programs related to proliferation

Vincent Detours

IRIBHM, Université Libre de Bruxelles (U.L.B.)
vdetours@ulb.ac.be

Part I

# How failed attempts to understand thyroid cancer agressiveness led to the super PCNA signature

# Papillary vs. anaplasic thyroid cancers

Papillary thyroid cancers (PTC)

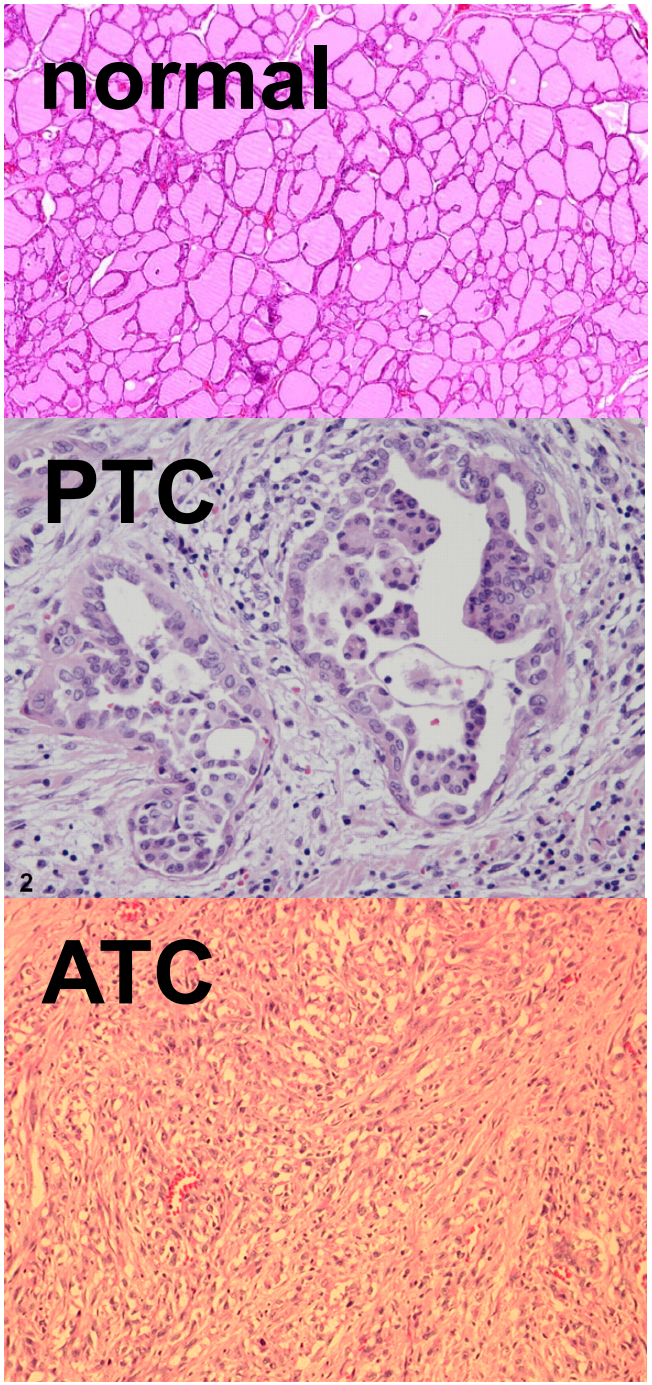
- Survival rate at 20 years is >90%

Anaplasic thyroid cancers (ATC)

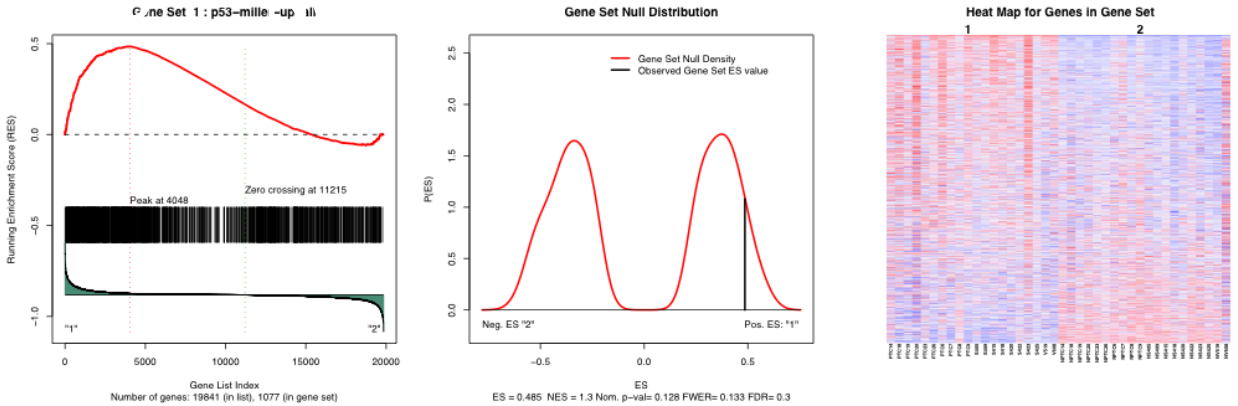
- Survival rate at 5 year is 1-5%

What explains this agressivness difference?

We profiled mRNA of 9 ATCs, 20 PTC, and 20 heathy thyroid tissues

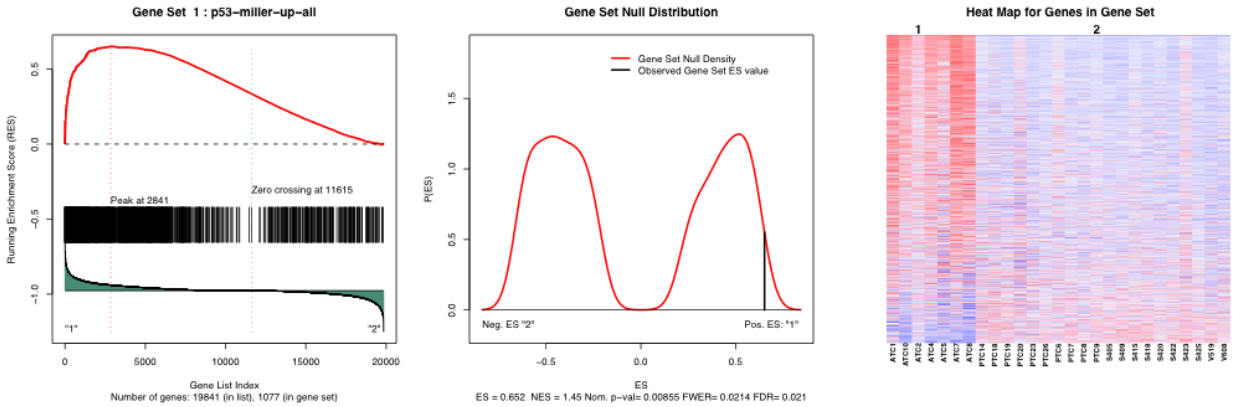


# p53 genes are over-expressed in PTC compared to healthy tissues...




- p53-regulated genes were derived by comparing $p53^{mut}$ $p53^{wt}$ breast tumors (Miller *et al.*, *PNAS*, 2005) and removing breast-specific genes

# ...but are even more over-expressed in ATCs compared to PTCs!




The same result was obtained with other p53 transcriptional signatures.

So expression data contradicts our hypothesis

# Another (unsupported) hypothesis: scenecence controls PTCs, not ATCs

1. Derive scenecence signature from published *in vitro* profiling of scenecent cells (Hardy *et al*, *Mol. Cell Biol.*, 2005)
2. assess its expression in thyroid data

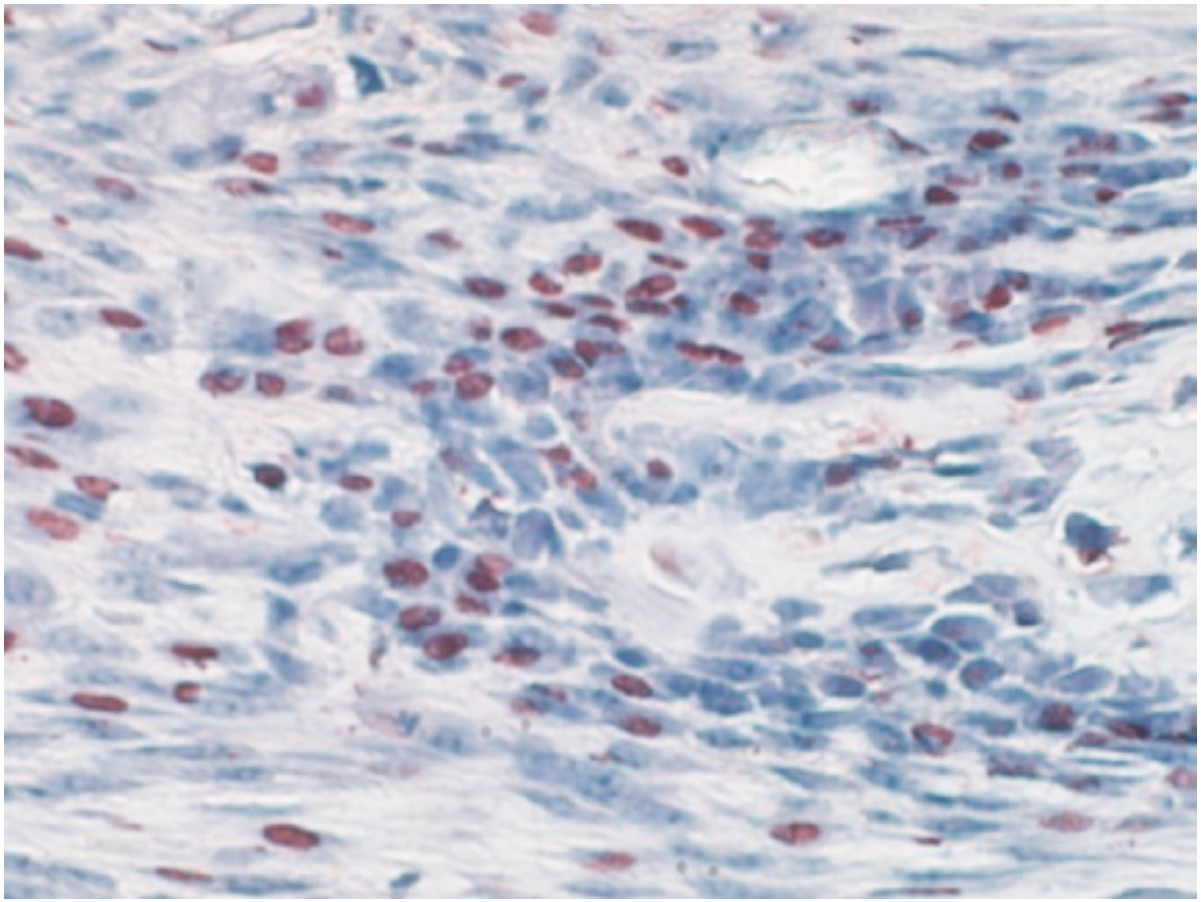
Again, it fails to support our hypothesis! But,

- signatures devived by comparing proliferating and non proliferating cells seems to parallel the trend seen for the scenescence signature in thyroid data

- Thus, proliferation is likely a confounder in the senescence and p53 signatures, hence their higher ATC expression

> **It might be possible to assess the proliferative status of a tissue from its expression profile**

# PCNA



- PCNA in among the most widely used proliferation markers in biology

⇐ Here is a PCNA immunostaining of a developing mouse bone tissue (cycling cell are stained in brown)

# Super PCNA,
# a gene expression signature of proliferation in normal, healthy human tissues

- We define the *super PCNA* signature as the 1% genes the most positively correlated with PCNA expression in a panel of normal human organ expression profiles
- Thus, super PCNA genes are expressed when PCNA is expressed (i.e. many cells are cycling) and not expressed when PCNA is not expressed (i.e. few cells are cycling)
- Normal organs were obtained from subject who died in traffic accidents (Ge *et al*., *Genomics*, 2005)

Super PCNA genes tend to be involved in DNA metabolism and the cell cycle, and to be expressed in cell's nuclei:

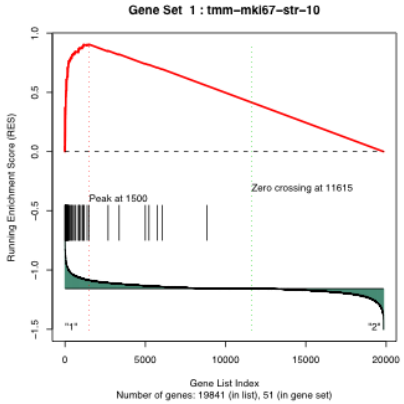
**Gene Ontology biological processes**

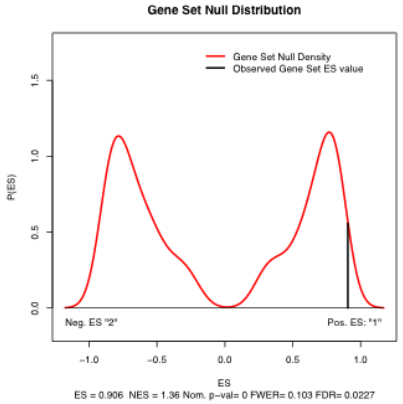
| Category | % of super PCNA genes |
|---|---|
| biopolymer metabolism | 43 |
| mitotic cell cycle | 13 |
| regulation of cell cycle | 18 |
| DNA repair | 11 |
| nucleobase, nucleoside, nucleotide and nucleic acid metabolism | 41 |
| M phase | 10 |

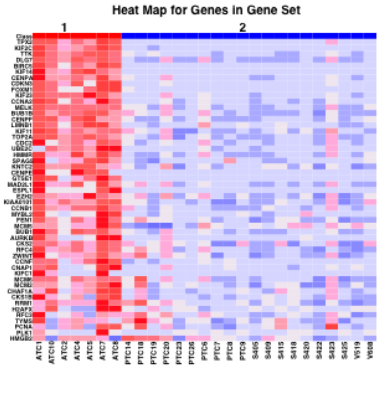
**Gene Ontology cellular localization**

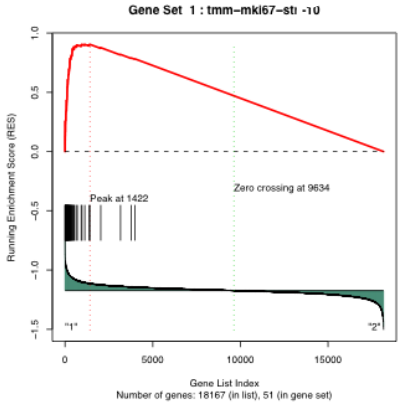
| Category | % of super PCNA genes |
|---|---|
| chromosome | 17 |
| intracellular non-membrane-bound organelle | 32 |
| nucleus | 51 |
| microtubule cytoskeleton | 11 |
| Spindle | 6 |
| replication fork | 4 |
| chromosome, pericentric region | 5 |

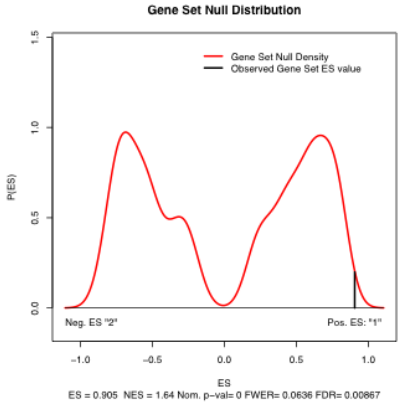
# Super PCNA in breast and thyroid cancers

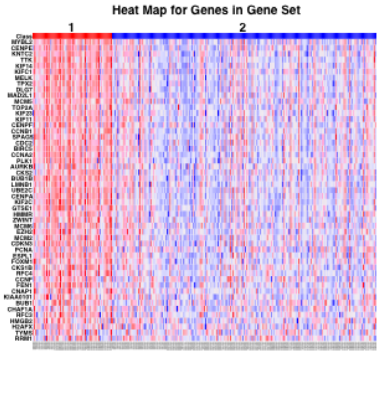


# Super PCNA and super Ki-67 seem to be universal markers of agressivness

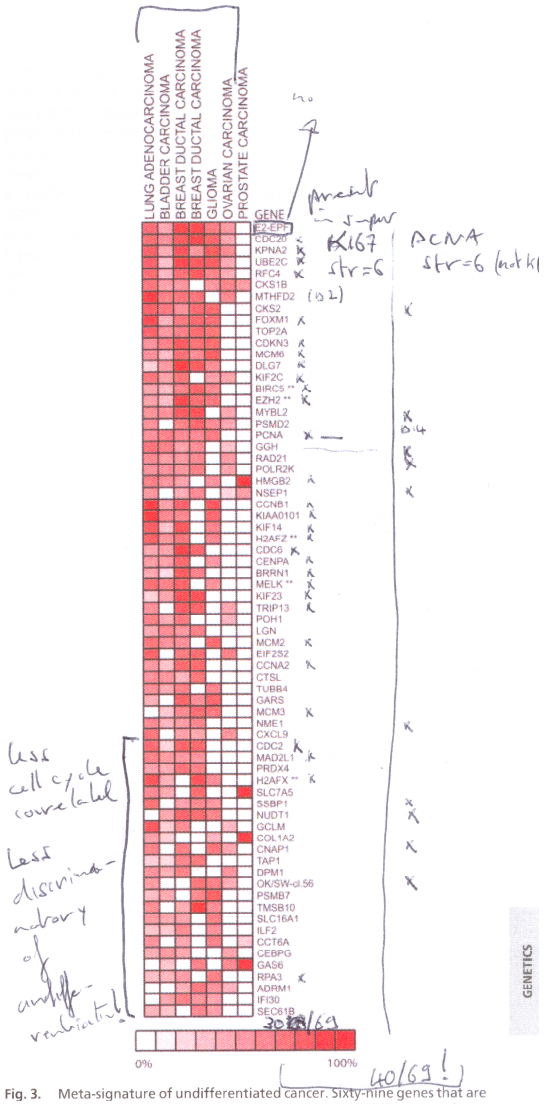


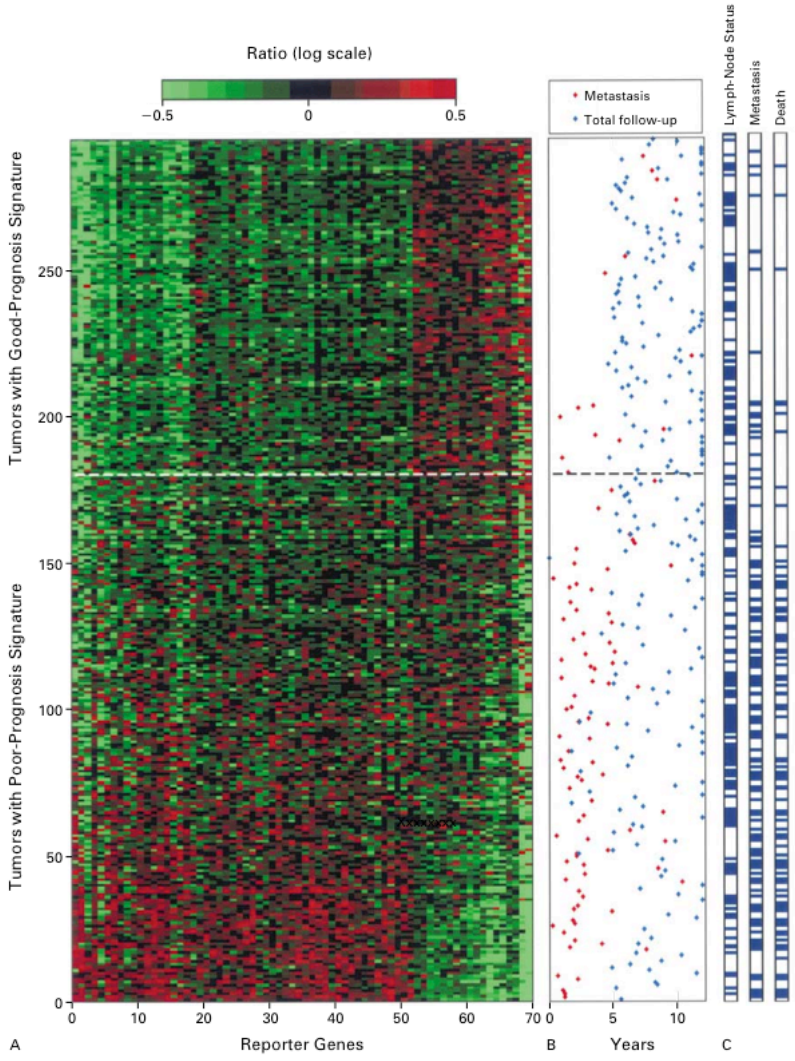

Fig. 3. Meta-signature of undifferentiated cancer. Sixty-nine genes that are overexpressed in undifferentiated cancer relative to well differentiated cancer ($Q < 0.10$) in at least four of seven signatures representing six types of cancer. See Fig. 2 legend for description.

- Rhodes et al. (PNAS, 2004) compiled expression data from 40 cancer studies
- They proposed a 69 genes 'meta-signature of undifferentiated cancers' that encompass lung, bladder, breast, glioma, ovarian tumors
- This signatures overlaps at 40/69 with super PCNA
- and 30/69 with super MKI67
- The probability of observing such overlaps by chance is nearly 0. The 3 signatures are essentially the *same* signature
- Our normal tissues proliferation signatures are makers of aggressiveness in several cancers

Part II

# Proliferation and breast cancer outcome predictors

# Microarray expression profiles predict survival in early breast cancer patients




- Genome-wide expression profiles of 295 early breast cancers
- Supervised search for genes associated with survival

$\Rightarrow$ 70 genes outcome predictor

(van de Vijver *et al*. *NEJM*, 2002)

# A vexing question

**What biological processes are the outcome predictors involved in?**

Beside sheer curiosity, this question has a bearing on devising treatments to control aggressive cancers

# But do these signatures have different predictive abilities?

- Four signatures tested in van de Vijver data showed nearly identical predictive value (Fan *et al*., *NEJM*, 2006)
- The same patients tended to be misclassified by all four signatures
- This suggested that although they have different biological rationale, the signature predictive ability rely on the *same* biological parameter
- **What biological parameter underlies outcome predictions?**
- **Is it possible to predict significantly better than Mammaprint's 70 genes signature?**

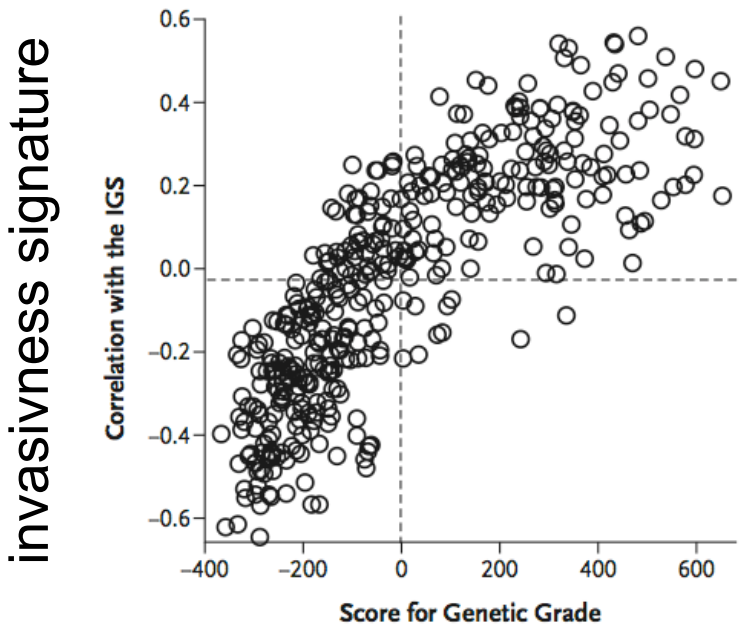
# Can proliferation be ruled out?

- Lui *et al*. (*NEJM*, 2006) reported, the IGS, a 'invasivness' signature that predicts breast cancer outcome

- Wennmalm *et al. (NEJM*, 2007*)* challenged it, noticing it has a 0.81 correlation with their own grade signature: '*ruling out dependency on proliferation now seems to be important in demonstrating the novelty of the IGS*'...

- **But can proliferation be ruled out?**

Wennmalm *et al. (NEJM*, 2007*)*



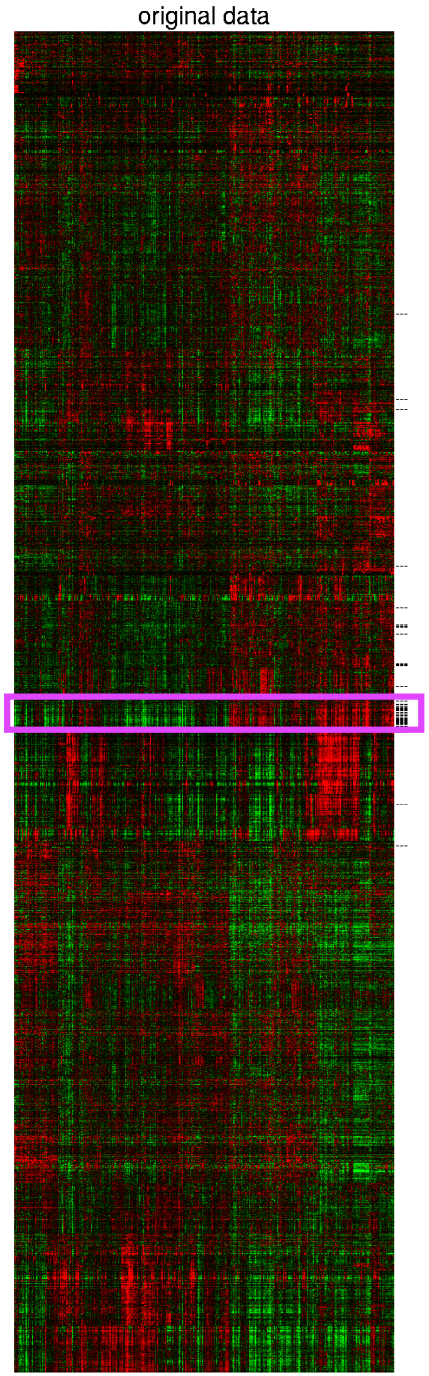

Grade signature

# Method outline

1. Remove proliferation-related signals from cancer expression data

2. See if the 46 published outcome predictors still predict anything once the proliferation signals are gone

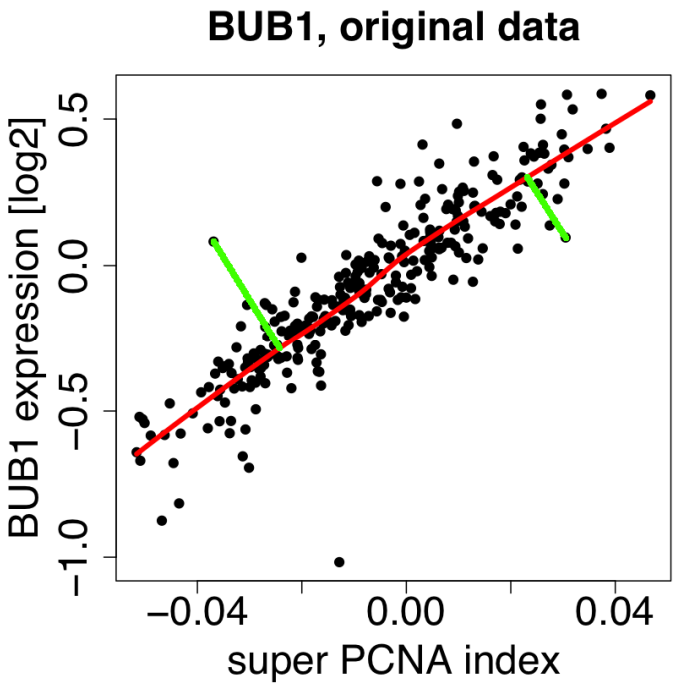
3. Do the same on 10,000 randomly generated predictors to get a comprehensive view of the process of signature searching and of the transcriptome properties




# Super PCNA genes show a coherent expression in breast cancers

- Super PCNA genes are correlated by definition in normal tissues
- The fact that they are also correlated in cancers is nontrivial
- It make sense to summarize their expression in a tumor to a single number, the meta PCNA index, defined as the median expression of super PCNA genes

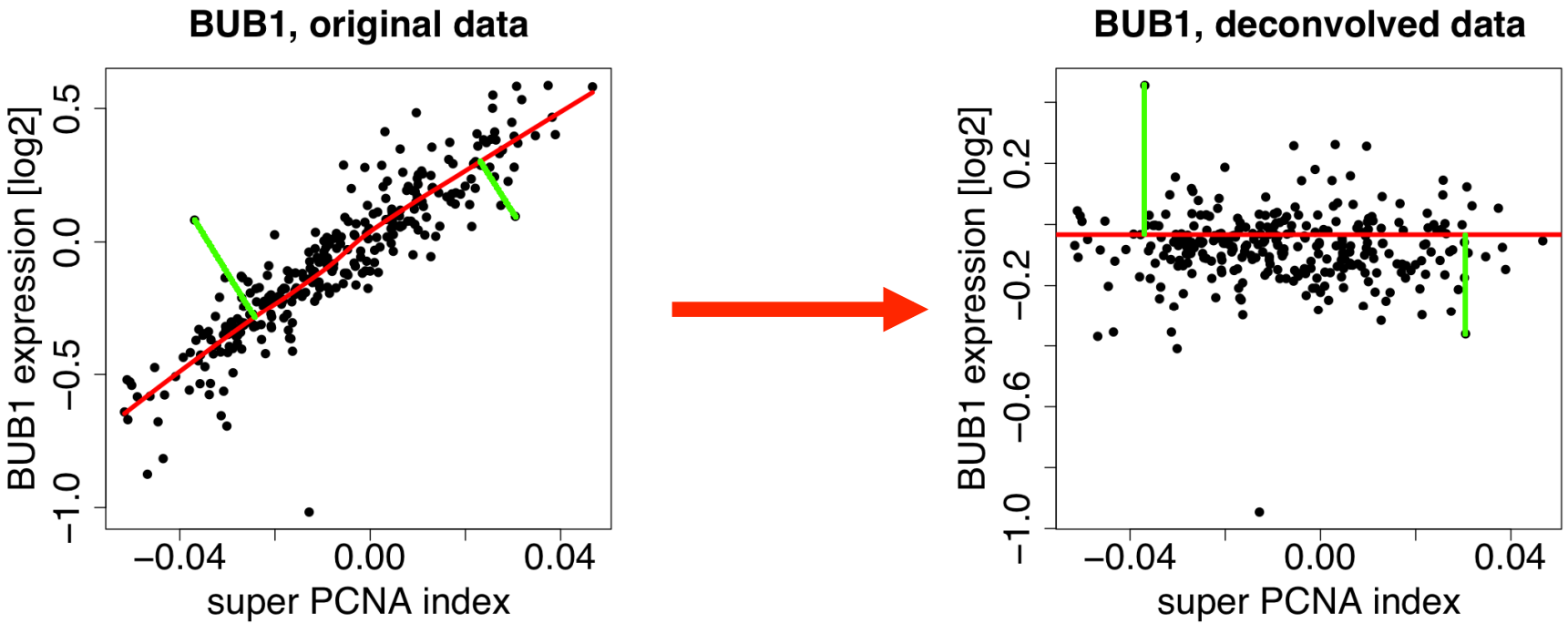
# The super PCNA deconvolution: removing proliferation-related signals out of any genome-wide expression data set




BUB1 vs. super PCNA index

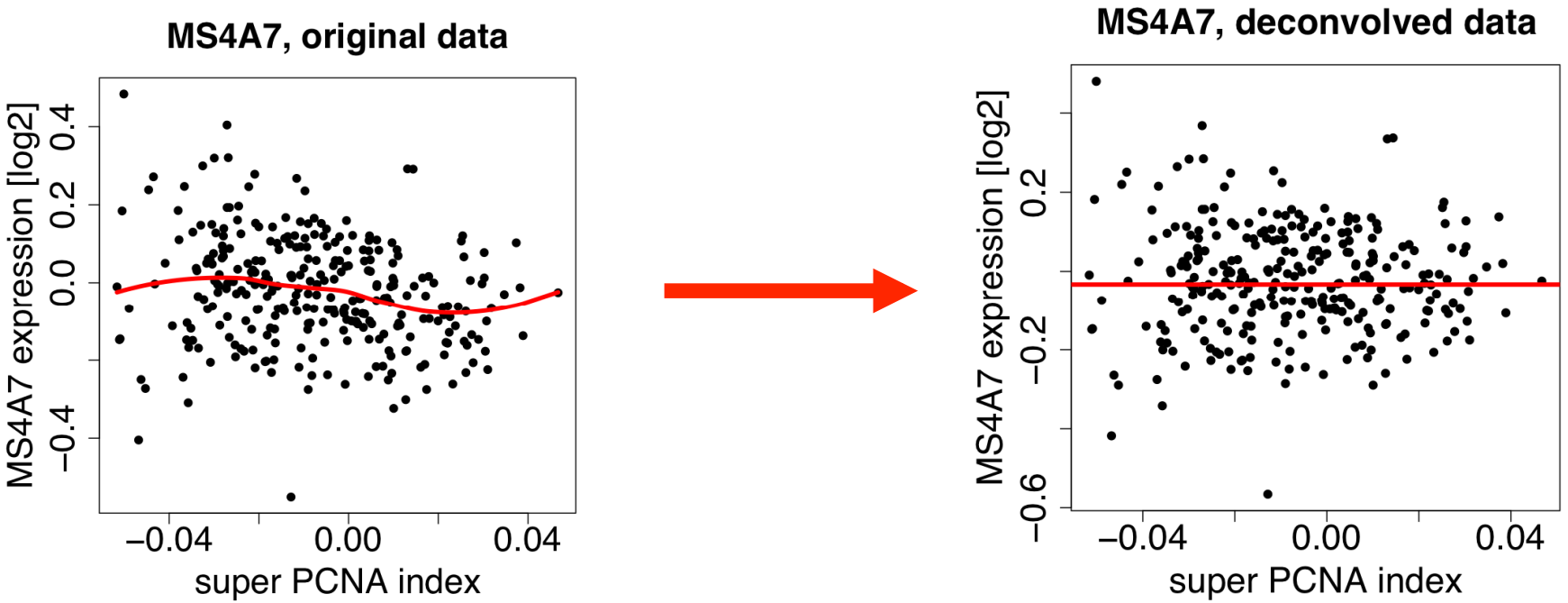
(BUB1 is a gene involved in mitotic spindle check point)

# The super PCNA deconvolution: removing proliferation-related signals out of any genome-wide expression data set



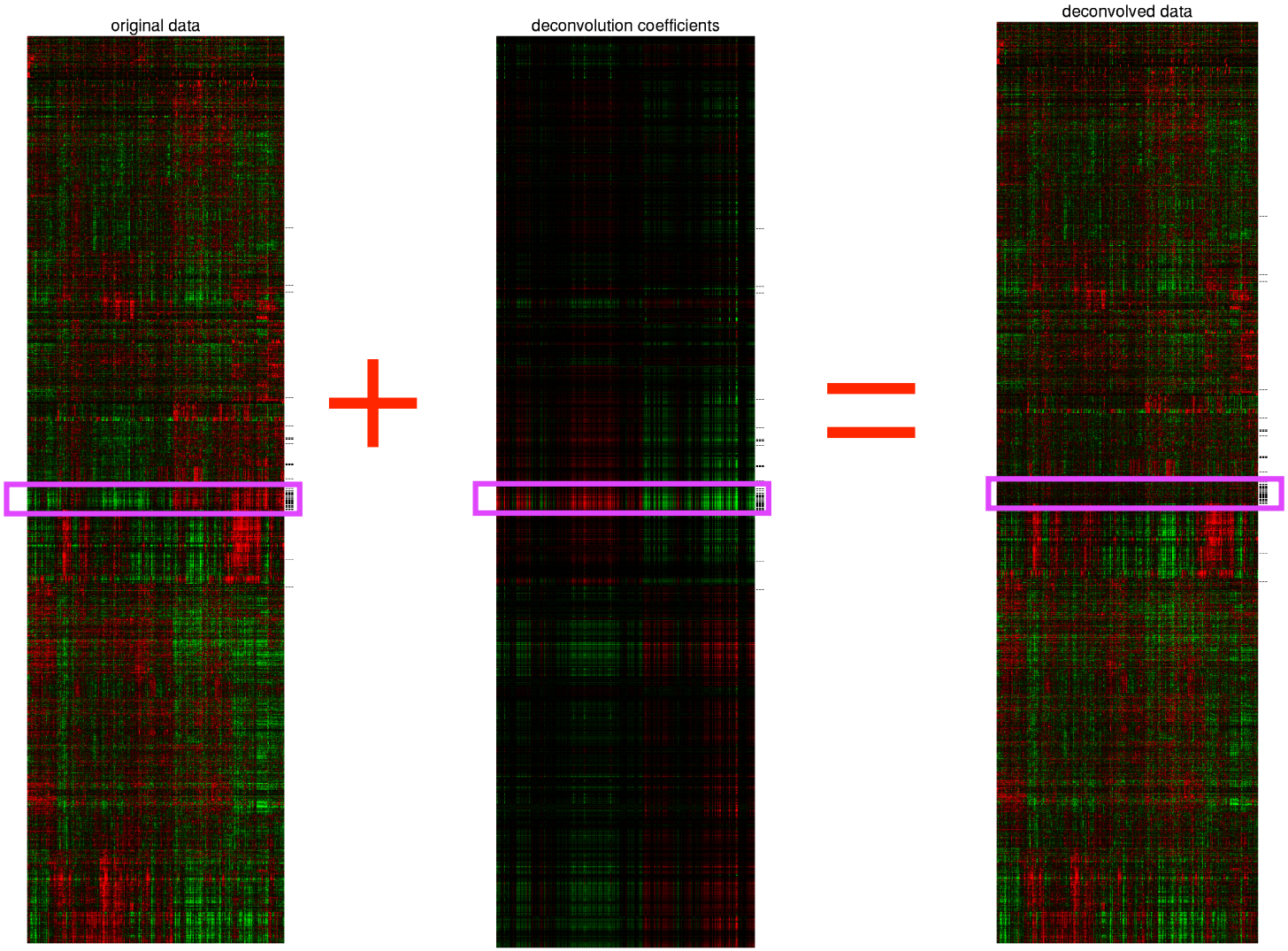

- Deconvolution is applies to all genes, *no gene is being removed*
- Deconvolution has the potential to reveal signals obscured by proliferation in the original data

# The meta PCNA deconvolution: removing proliferation-related signals out of genome-wide expression data sets




Deconvolution has no effect on genes unrelated to the meta PCNA index

# The super PCNA deconvolution: removing proliferation-related signals out of genome-wide expression data sets

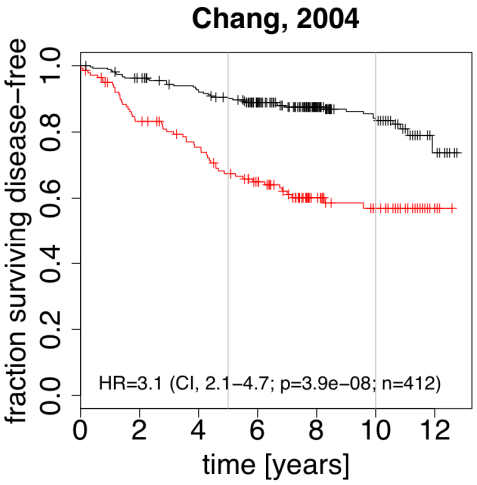


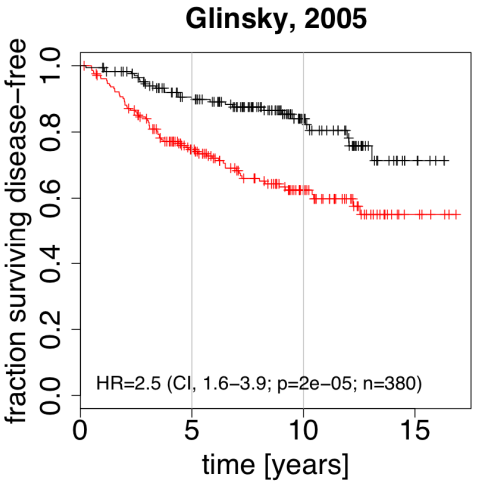
# Now we investigate outcome predictors in the original and deconvolved versions of 3 breast cancer expression data sets

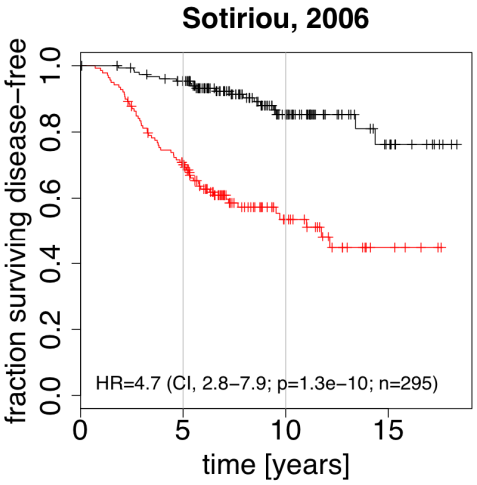
- 412 tumors from Calza *et al*. (*Breast Cancer Res.*, 2006)
- 380 tumors from Loi *et al*. (*J. Clin. Oncol*., 2006)
- 295 tumors from van de Vijver *et al*. (*NEJM*, 2002)

# The super PCNA deconvolution drastically reduces the predictive abilities of published and random signatures

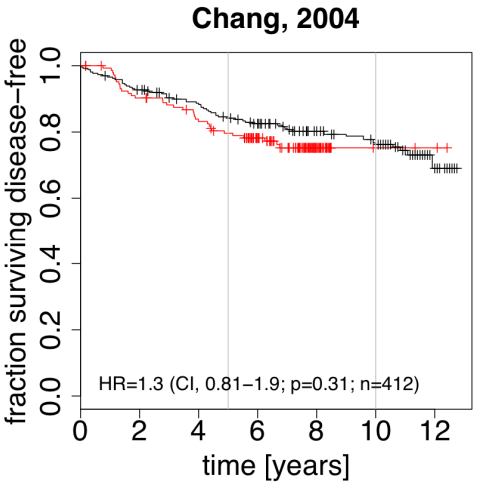
**Original data**

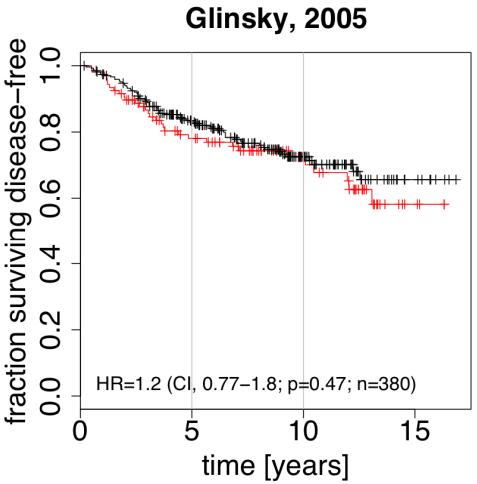


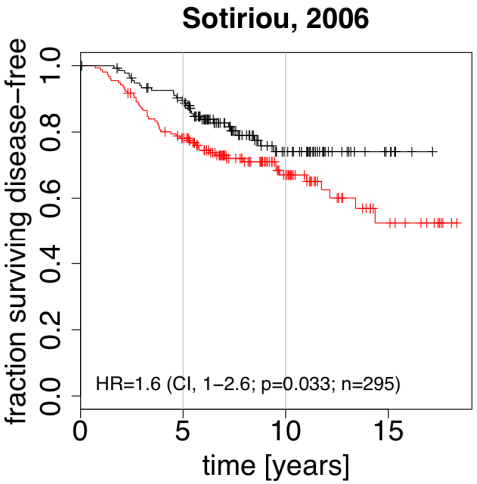



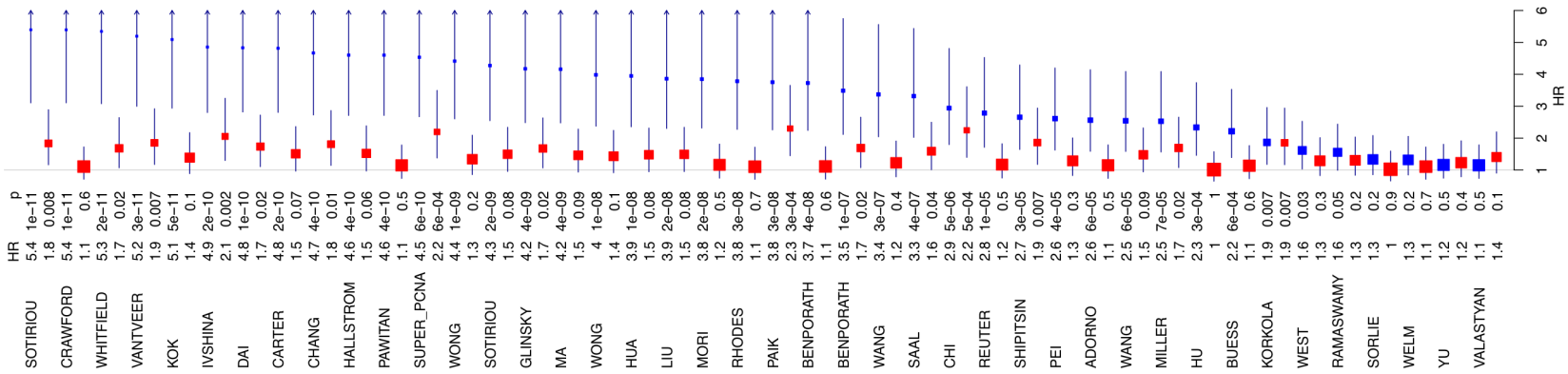

**Deconvolved data**



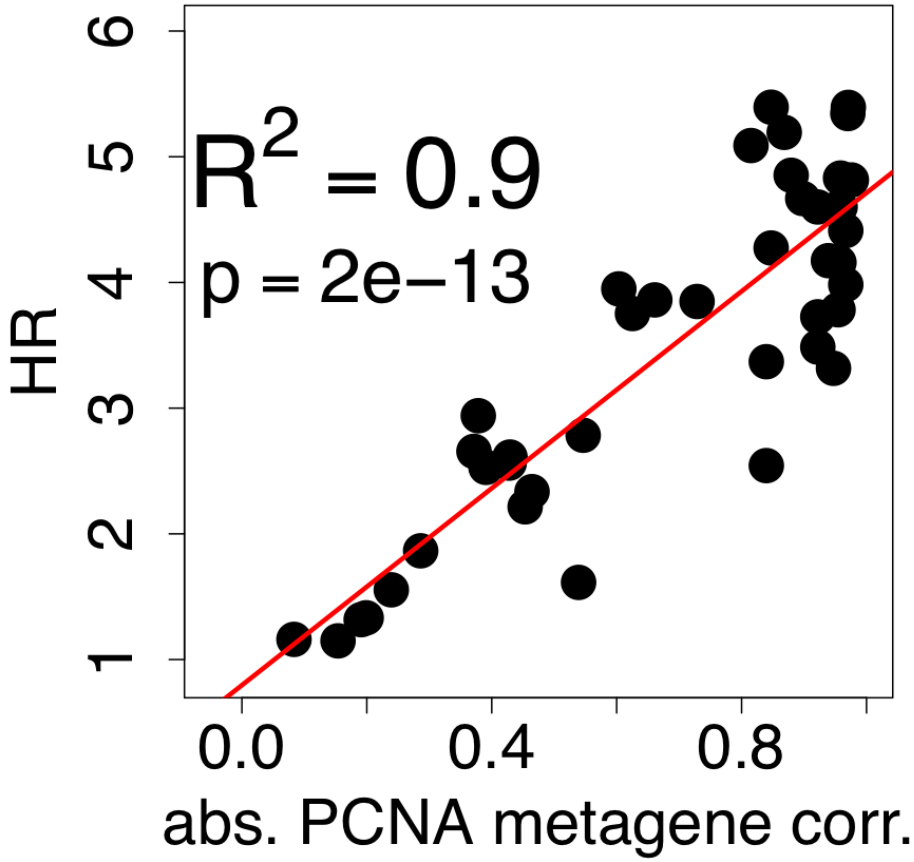



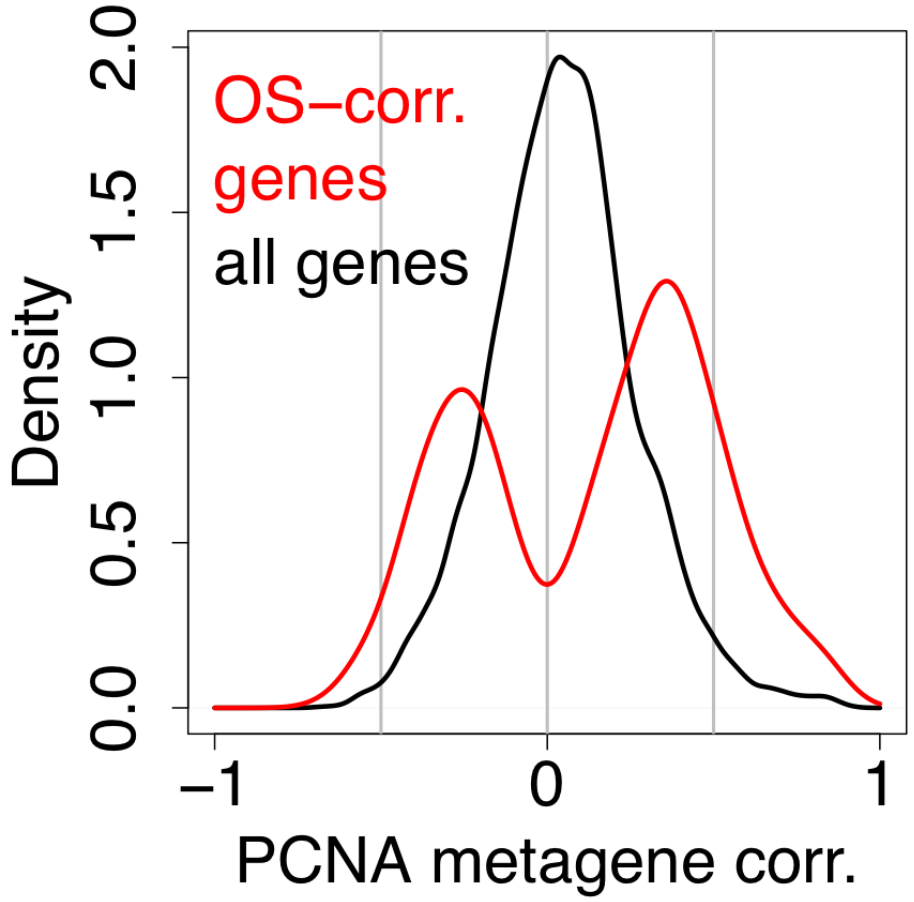
# The meta PCNA deconvolution drastically reduces the predictive abilities of published and random signatures



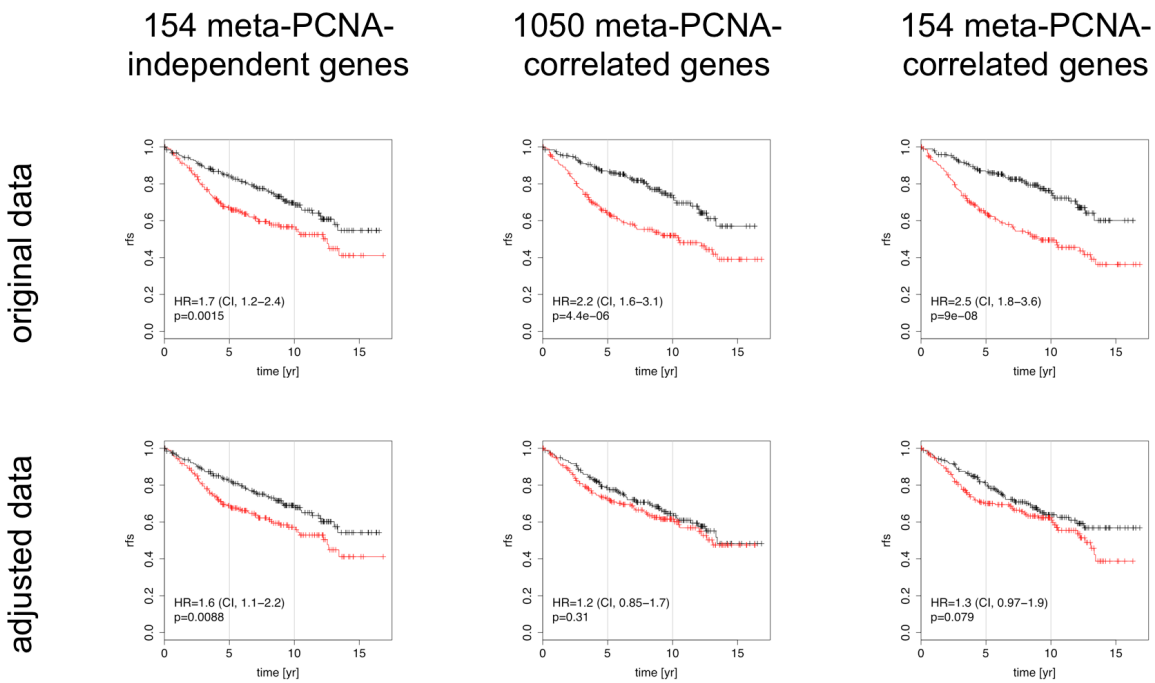
# The more outcome-related a signature the stronger correlation with proliferation

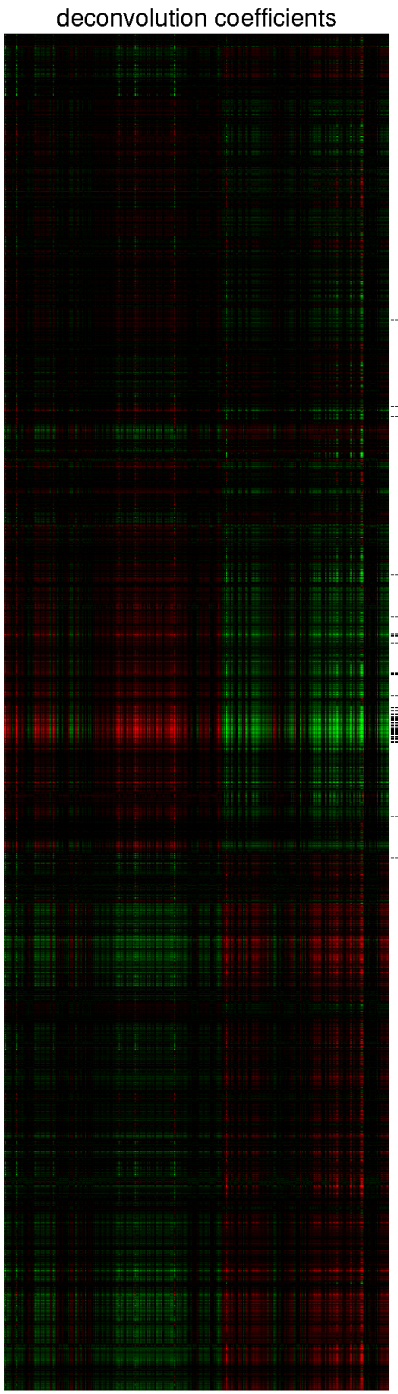


# Most outcome-related genes are correlated with meta PCNA



# 1% of genes are prognostic, yet not proliferation-related






# Proliferation-related signals are ubiquitous in the breast cancer transcriptome

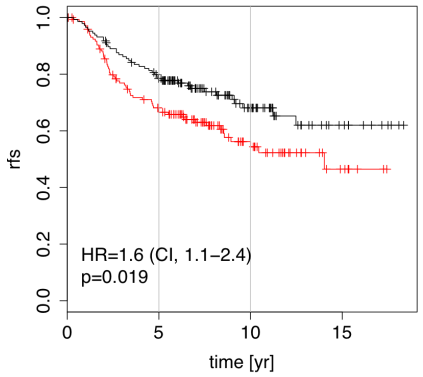
- The meta PCNA deconvolution coefficients reveal otherwise invisible proliferation-related signals
- 3-4% percent of the transcriptome is strongly proliferation-related
- Much of the transcriptome is weakly, but clearly, related to proliferation

# So, does *any* signature work?

**Effect of post-prandial laughter on PBMC** **(Hayashi et al., 2006, Psychother. Psychosom 75 p66)**

# So, does *any* signature work?

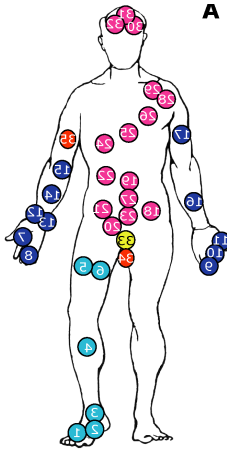
**Effect of post-prandial laughter on PBMC** (Hayashi et al., 2006, Psychother. Psychosom 75 p66)

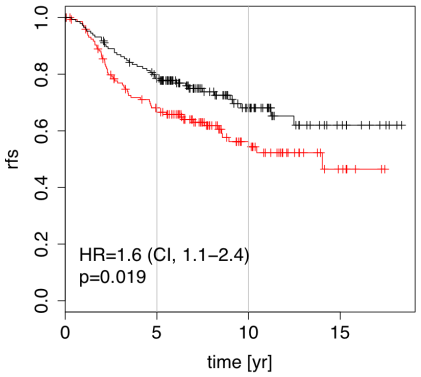


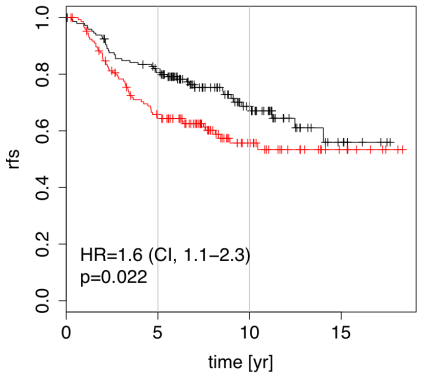
# So, might *any* signature work?

**Effect of post-prandial laughter on PBMC** (Hayashi et al., 2006, Psychother. Psychosom 75 p66)

**Signature of human skin fibroblast localization** (Rinn et al., 2006, PLoS Genet 2, e119)

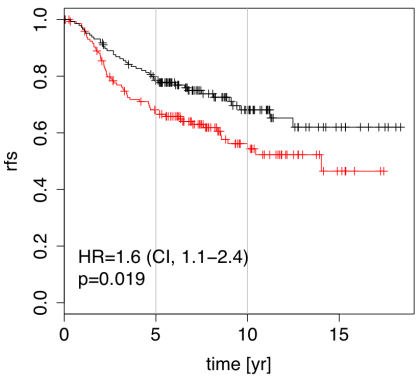


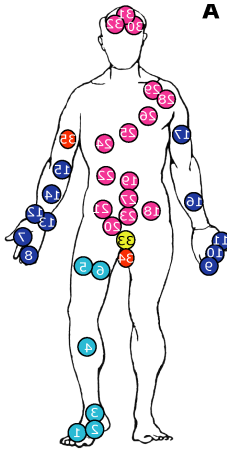



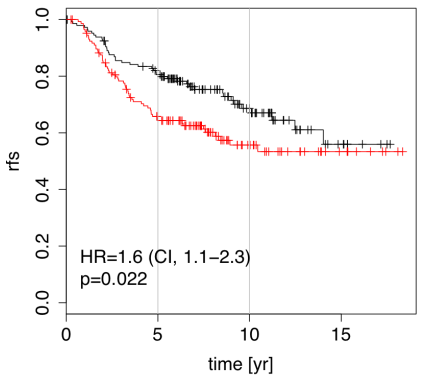
# So, might *any* signature work?

**Effect of post-prandial laughter on PBMC** (Hayashi et al., 2006, Psychother. Psychosom 75 p66)

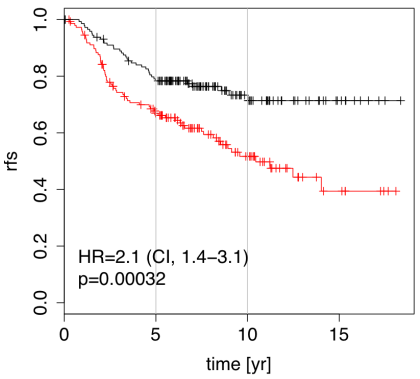



**Signature of human skin fibroblast localization** (Rinn et al., 2006, PLoS Genet 2, e119)



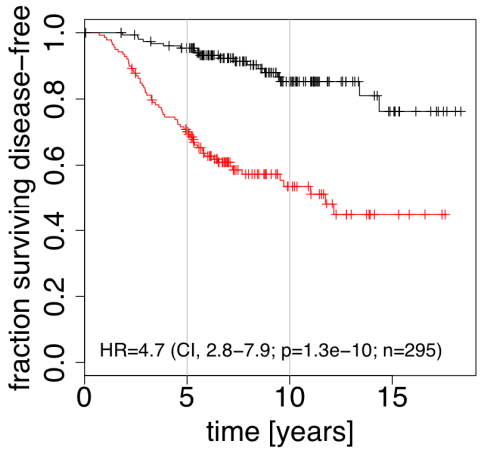




**Effect of social defeat susceptibility on mice brains** (Krishnan et al., 2007, Cell 131, p391).

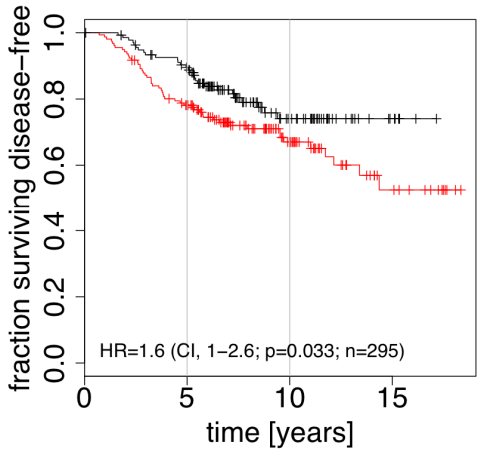


# Technical parenthesis: the predictive value of a signature may be summarized as a log-rank *p*-value

- We are going to study 10,000 signatures, so we need a simple way to describe their predictive value
- The log rank *p*-value is the probability that the observed survival difference between two patient groups is explained by chance alone
- The smaller it is the more robust the prediction is



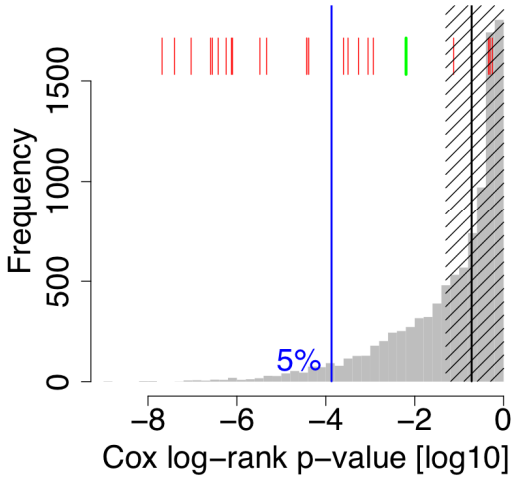

This is a robust predictor!
$p=1.3*10^{-10}$



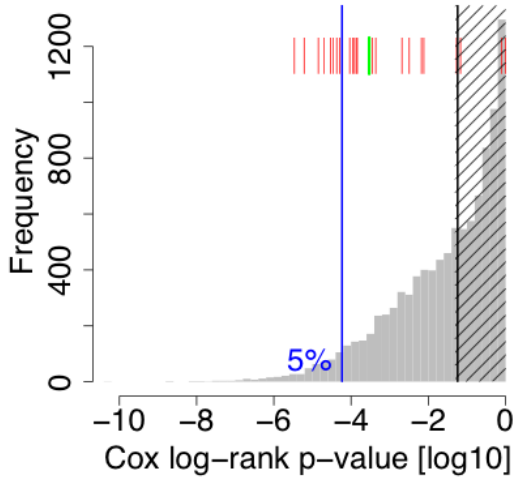

This one is less robust,
$p=0.03$,
yet significant according to typical statistical standards

# Most randomly generated signatures are significant outcome predictors

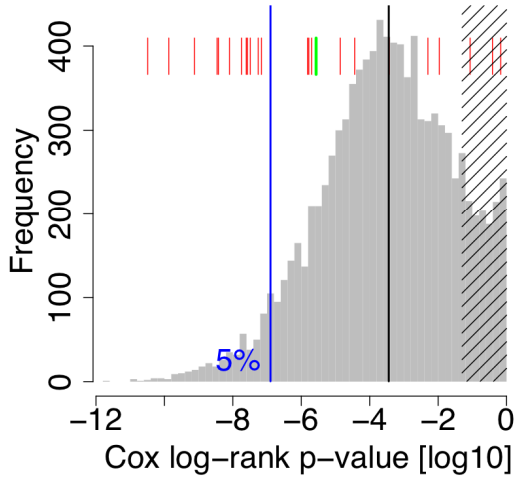
Calza *et al.* (412 tumors)




Loi *et al.* (380 tumors)



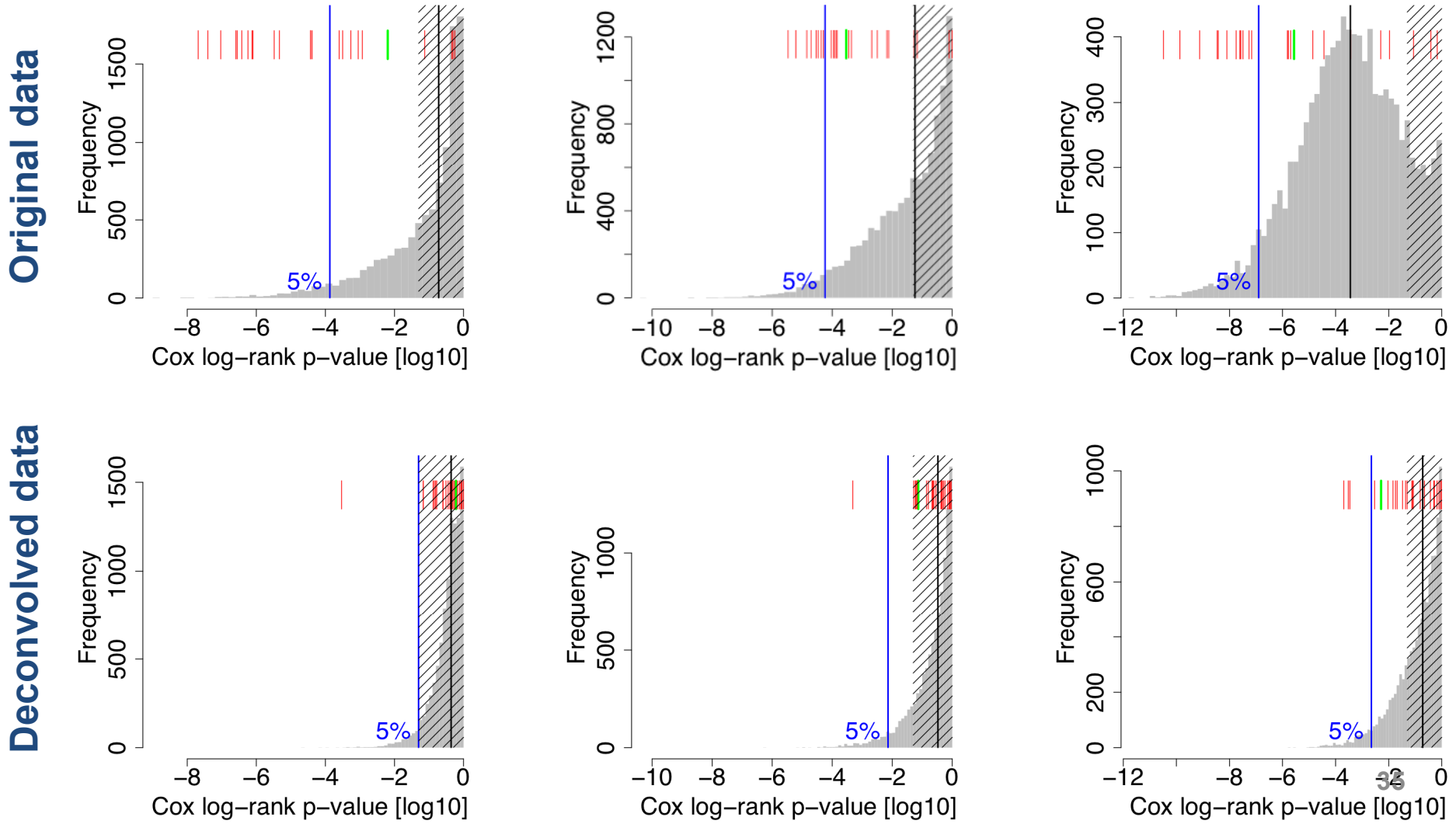

van de Vijver *et al.* (295 tumors)

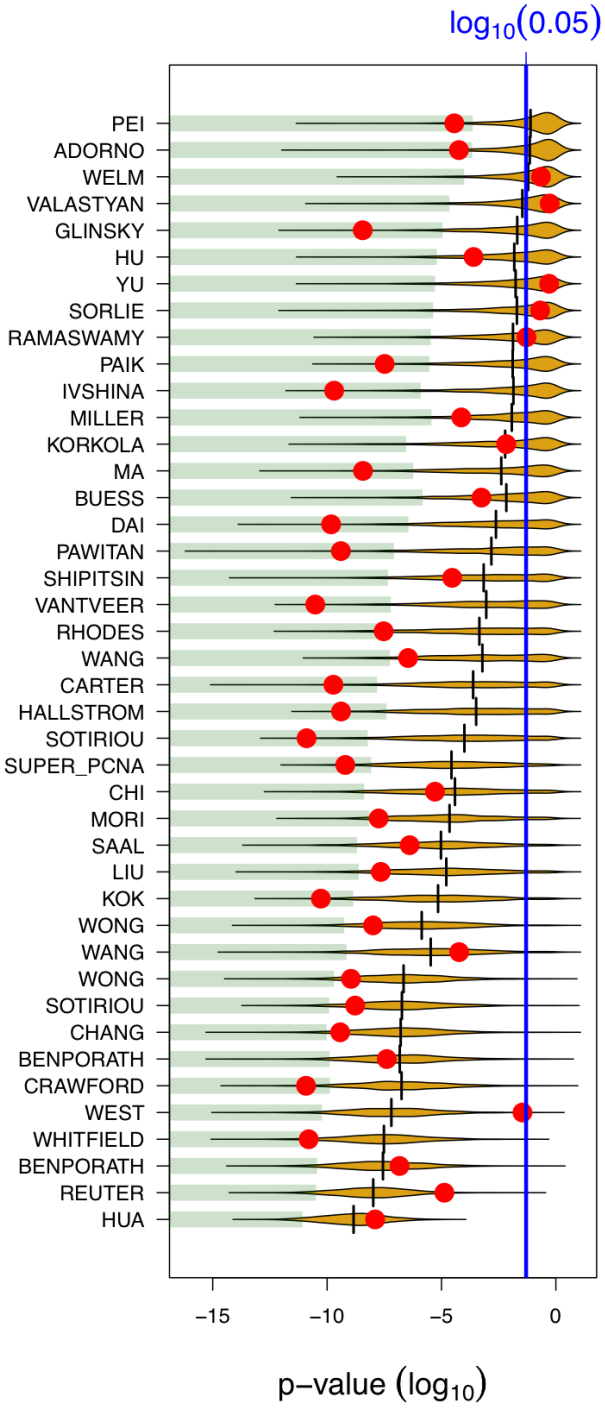



• >50% random signatures predict outcome in Loi *et al*. and van de Vijver *et al*. data sets

• 34-64% of published signatures are not significantly better than random signatures, i.e. *no* statistical evidence supports their clinical significance

# Meta PCNA deconvolution drastically reduces the predictive abilities of published and random signatures






Random signatures are likeky to predict outcome and many published signature are not significantly better

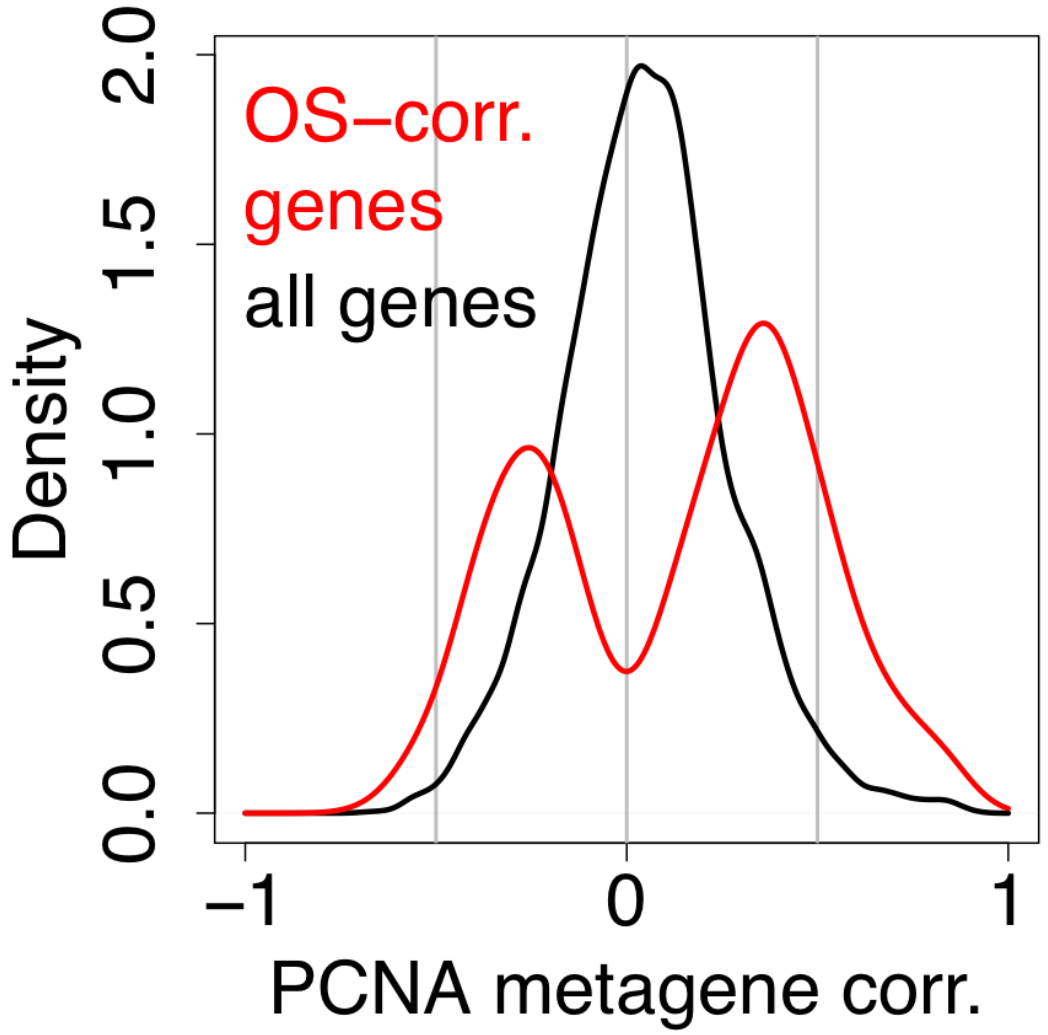
**published signature**

**distribution for 1,000 random signatures**

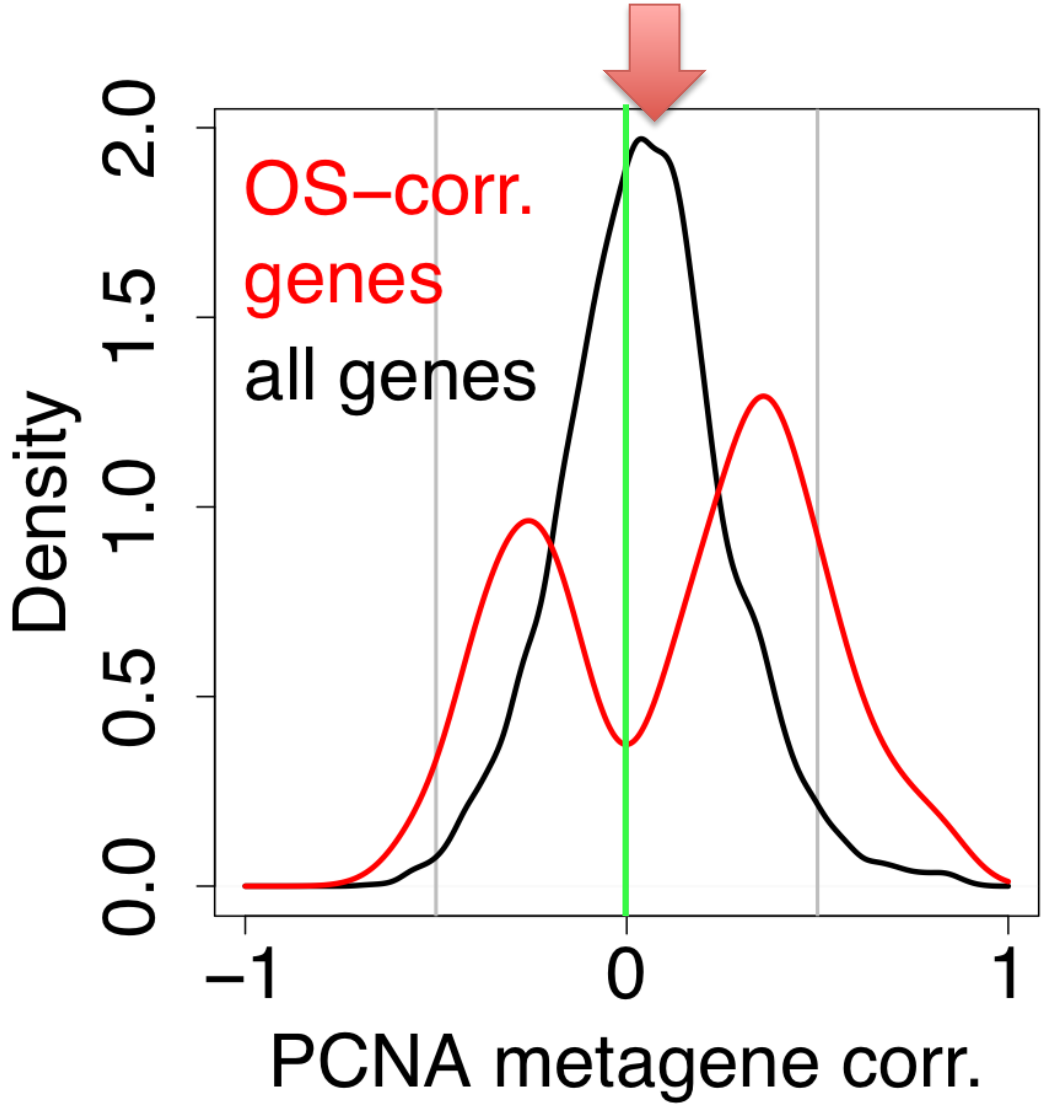
**5% best same-size random signatures**

- **22 signatures in 41 are not better that the 5% best same-size random signatures**
- **10 are worst than the median random signature**

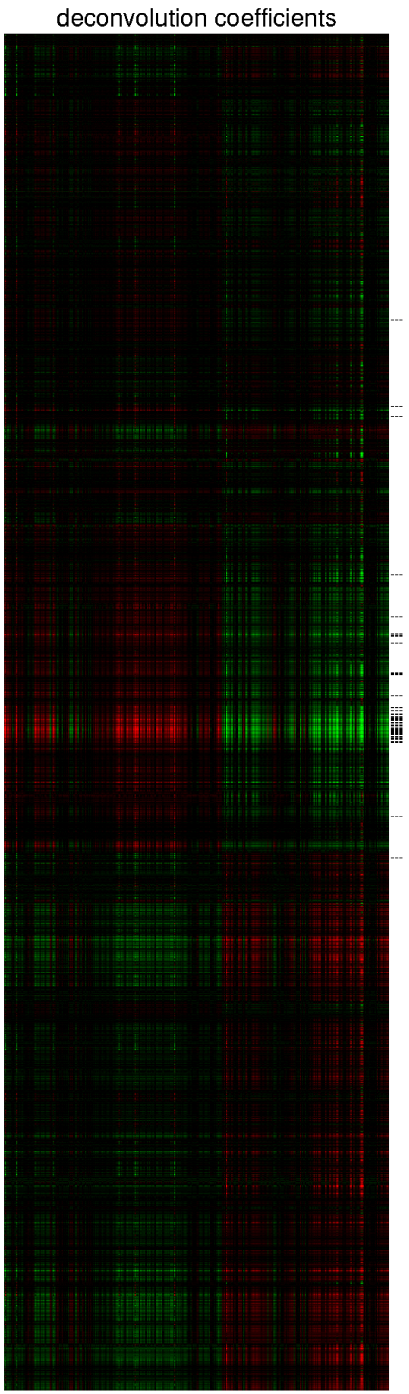
Any set of genes associated with outcome is almost certainly proliferation-related



Any set of genes associated with outcome is almost certainly proliferation-related



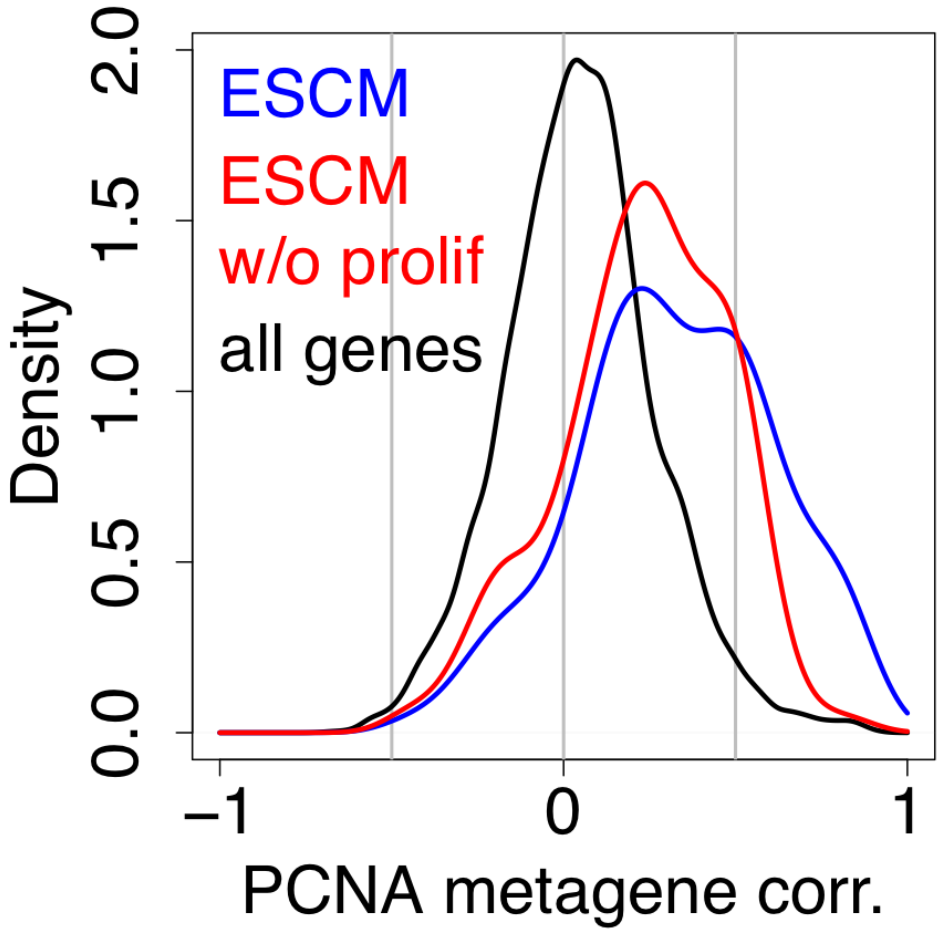

**~50% of the breast cancer transcriptome is correlated with proliferation**




# Proliferation-related signals are ubiquitous in the breast cancer transcriptome

- The meta PCNA deconvolution coefficients reveal otherwise invisible proliferation-related signals
- 3-4% percent of the transcriptome is strongly proliferation-related
- Much of the transcriptome is weakly, but clearly, related to proliferation

# Removing known ‘proliferation genes’ does *not* rule out proliferation




**(Counter) example:**

**The ‘embryonic stem cells module’ of Wong et al. (2008, Cell Stem Cell 2, p333), is still massively correlated with meta PCNA after removal of known proliferation genes**

# Meta PCNA deconvolution drastically reduces the predictive abilities of published and random signatures

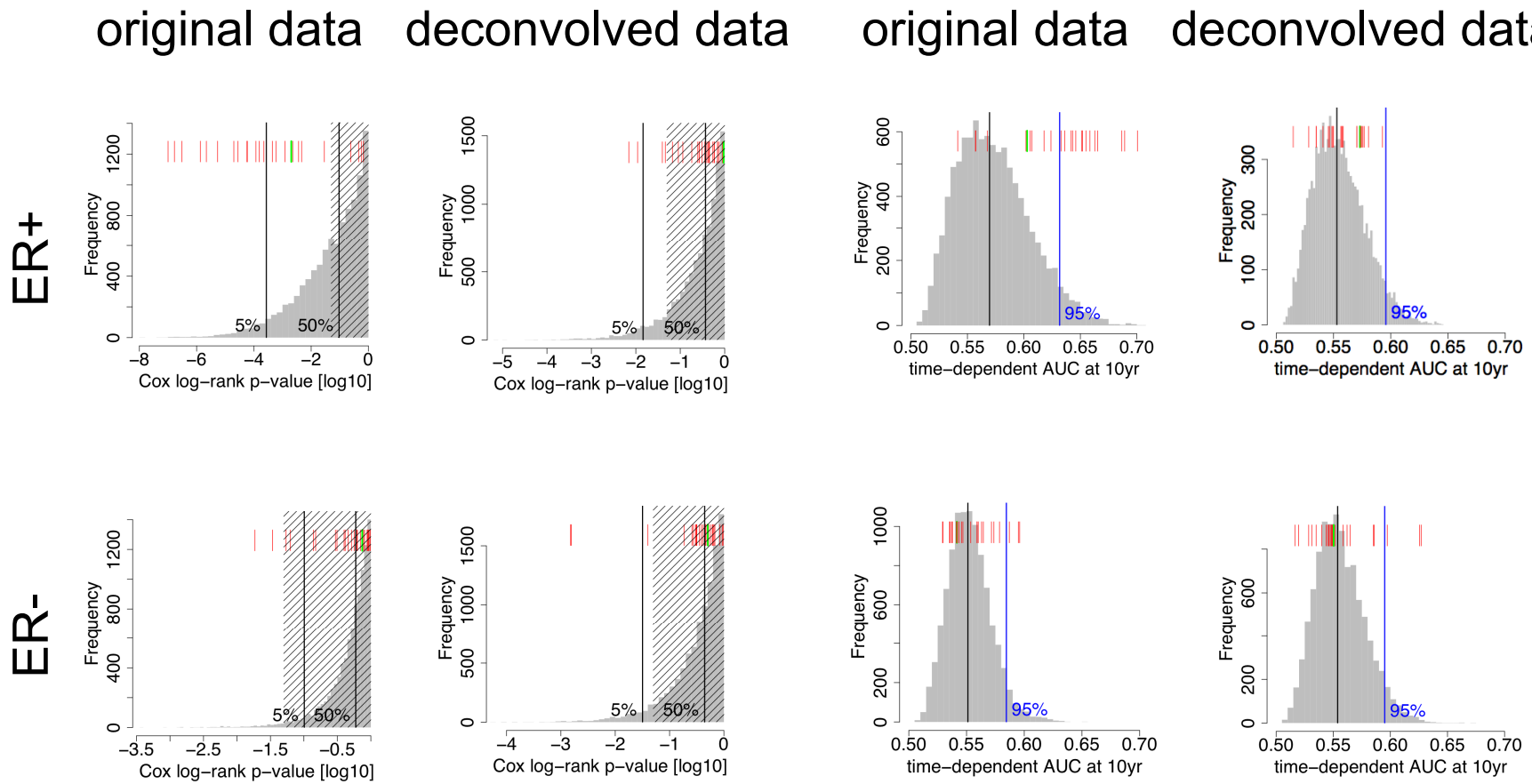
**1- Loss of predictive power could result from a faulty deconvolution that damages the data**

**Control**: deconvolving permuted (i.e. identically distributed, but biologically meaningless) super PCNA indices does *not* affect predictions

**2- The observed effect may not be specific of super PCNA**

**Control**: the entire procedure was rerun with 200 randomly selected gene instead of PCNA. 2-4% of these genes lead to equal or stronger loss of predictive power

# Effect of ER status

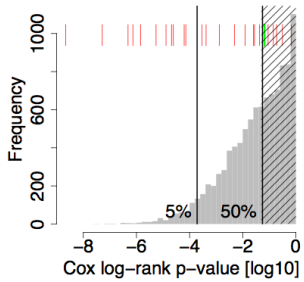


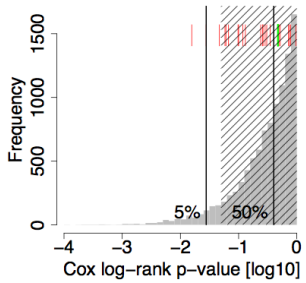

reanalysis of van de Vijver *et al*. (*NEJM*, 2002) data

# Effect of nodal status

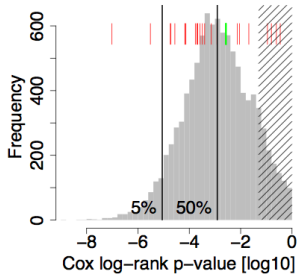
original data deconvolved data

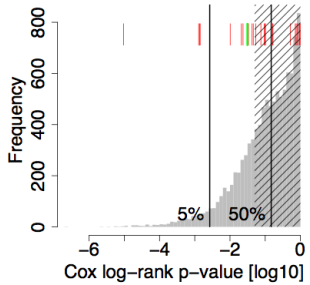
node +




node -



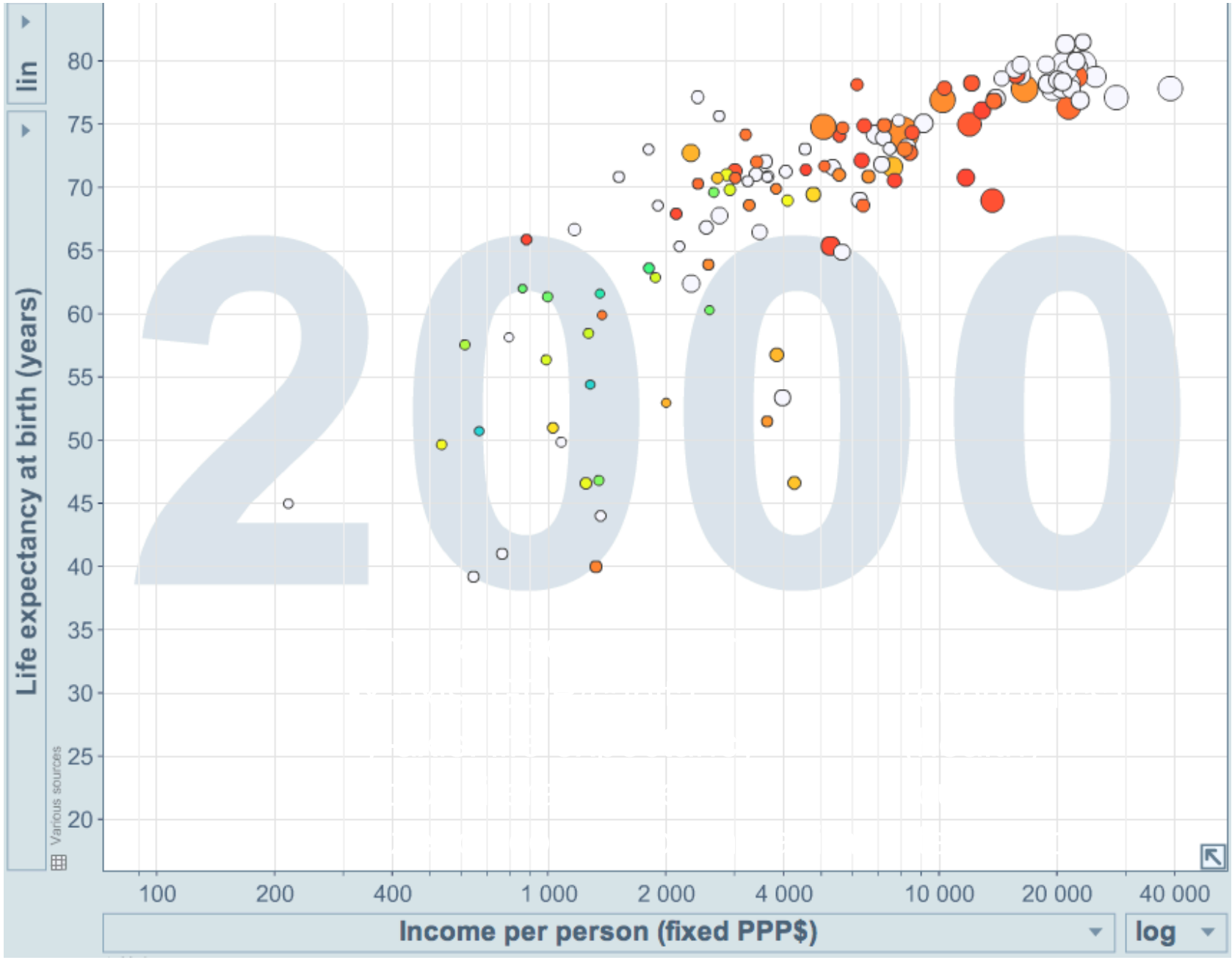




reanalysis of van de Vijver *et al*. (*NEJM*, 2002) data

# Conclusions

- Much more stringent statistical standards are needed to establish the biological/clinical importance of gene expression signatures (and the of the underlying biological rationale)

- Signatures' predictive abilities lye in their ability to capture the proliferation-related signals that are ubiquitous in cancer expression data




(from www.gapminder.org)

# Limits

- We and others addressed a relatevely wide patient population. Subgroups of patients may have proliferation-independent outcome signatures

- The data in this and other studies rely on bulk tumors and may miss the biology of small tumor compartments (e.g. tumor stem cells, invasive edges, etc.)