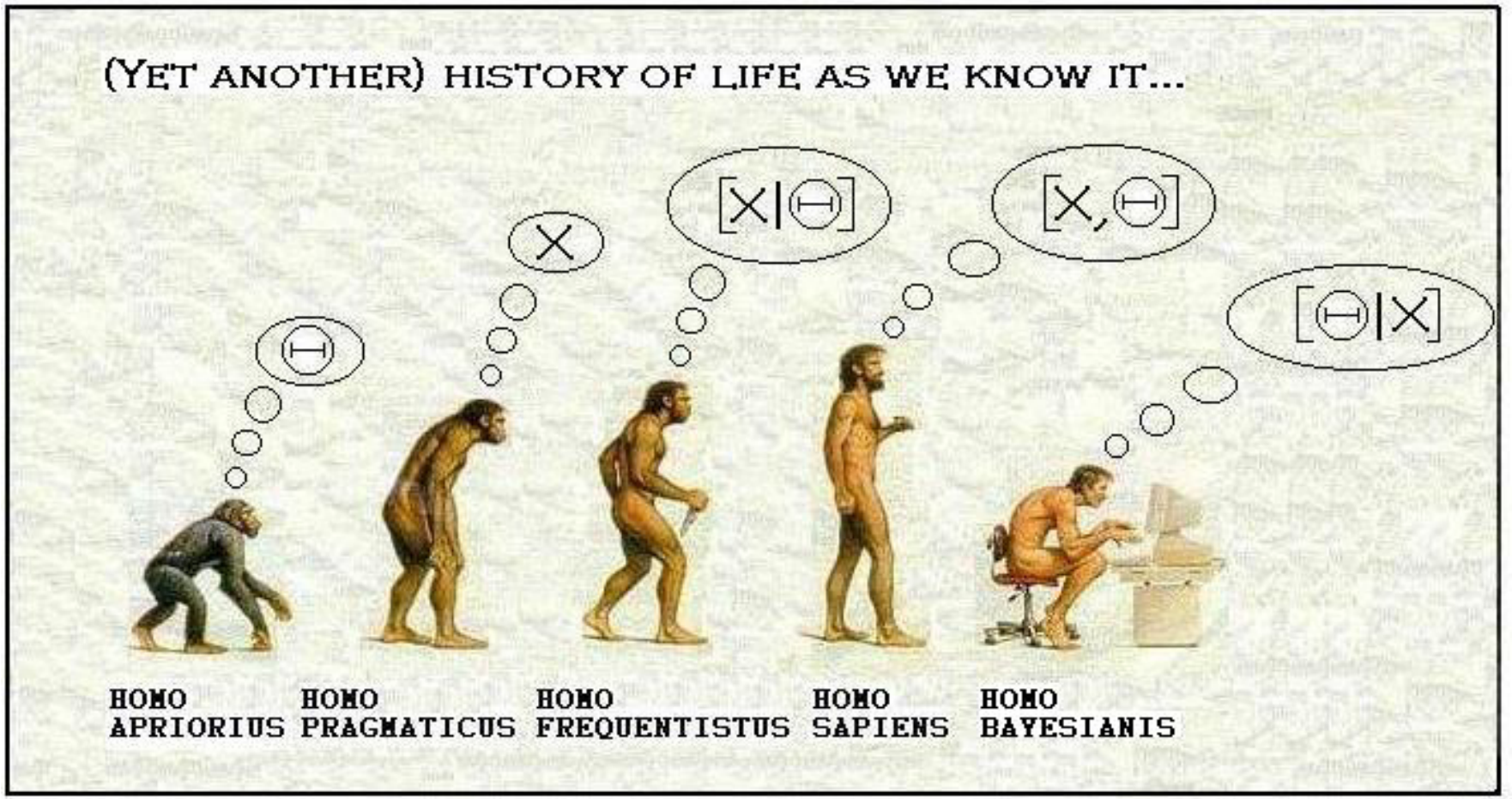
Binary Regressions & Molecular Phenotyping


(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
Θ
X
$[X|\Theta]$
$[X,\Theta]$
$[\Theta|X]$
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS

Binary Regression & Molecular Phenotyping

Physiological & clinical states: 0/1
Initial context of much molecular phenotyping

Prediction: Prognosis, diagnosis
Gene discovery, relationships : Pathways?

Breast cancer:
ER status, nodal status, treatment response

Cardiovascular: State of atherosclerosis
Lung, ovarian cancer:
Tumour vs.Normal? Treatment response
…

Regression & Molecular Phenotyping

Gene expression as predictors/covariates

"Large *p*, Small *n*" paradigm

Factor regression:
Reduce dimension, metagene predictors

Shrinkage priors, but with *p=000s*?

Variable selection issues

Later:

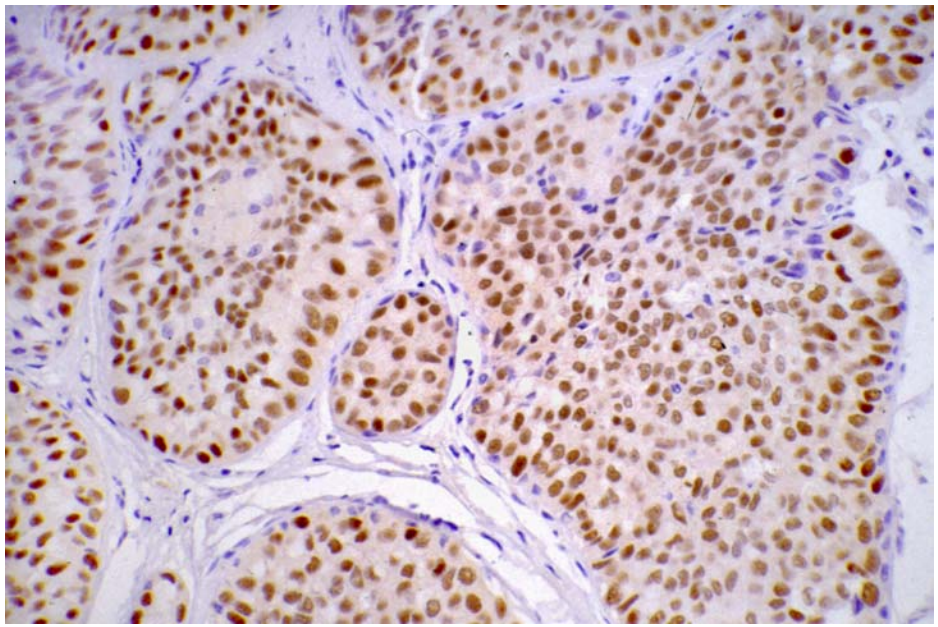
- Bayesian variable selection on large-scale
- Prior probs on variable "In" model, number of included variables

Example: Breast Cancer Data

ER – Hormone Receptor Positivity: IHC Test

Z=0/1 (ER -/+)
Protein assay
Immunohistochemical staining
0/1 (0-3)

Frozen tumour: Gene expression



ER positive tumour

IHC for Estrogen Receptor
(~60x magnification)

nuclei of breast epithelial cells
cytoplasm of breast epithelial cells

brown-red & pink ~ ER+

nucleii of stromal cells; collagen

Biomarkers of Human Cancer site

Duke PNAS 2001 Breast Cancer Data

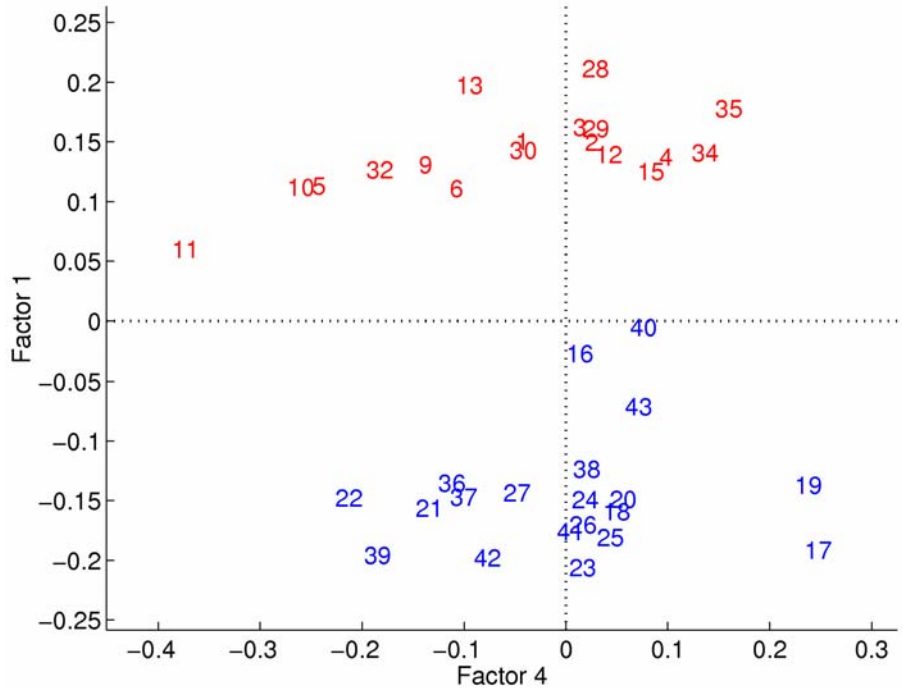
ER 'proof of principle' example

p=100 top genes: *n*=38

ER "Big" pathway – many genes

Discriminatory patterns
– ER pathway genes? –

Heterogeneity?
Predictive validity?
"Interesting" tumours?



Data separation: no MLE - need for priors/shrinkage

Binary Regression Models

Generalised linear models

$$\Pr(z_j = 1) = \pi(\mathbf{h}'_j \boldsymbol{\beta})$$

Probit: $\pi(\mu) = \Phi(\mu)$ Normal CDF

Logit: $\pi(\mu) = (1 + \exp(-\mu))^{-1}$ Logistic CDF

$$\mu_j = \mathbf{h}_j' \boldsymbol{\beta}$$

$$\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\beta}$$

Latent variable genesis and interpretation

$$z_j = I(y_j > 0)$$

$$y_j \sim \pi(y_j - \mu_j)$$

Each observation results from thresholding a latent underlying continuous variable … with linear regression structure

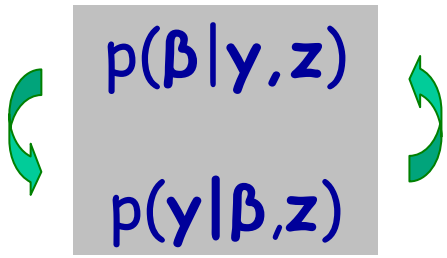
Probit:

$$y_j \sim N(\mu_j, 1)$$

$$\mathbf{y} \sim N(\mathbf{H}\boldsymbol{\beta}, \mathbf{I})$$

MCMC in Probit Models

Iteratively simulate : impute latent *y*



$$z_j = I(y_j > 0)$$

$$\mathbf{y} = \mathbf{H}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I})$$

Prior $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{C}^{-1})$

Posterior $\boldsymbol{\beta} \mid \mathbf{y} \sim N(\mathbf{b}, \mathbf{B}^{-1})$

$p(y_j \mid \beta, z_j)$ independent

z_j truncates normal distribution for $y_j \mid \beta$

Inverse CDF to sample new y_j

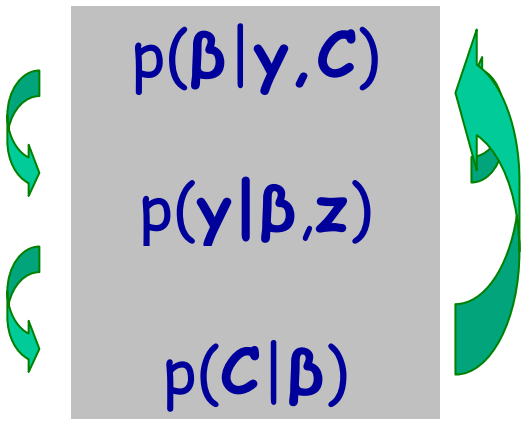
(see stats notes)

nb. Extends to logit, T link, etc

MCMC with Shrinkage Priors in Probit Models

Exactly as in linear model:
Couple in simulation of shrinkage parameters

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{C}^{-1})$$
$$\mathbf{C}^{-1} = diag(\tau_1, ..., \tau_n)$$



$$\tau_i^{-1} \sim Ga(k/2, h/2)$$

$$\tau_i^{-1} \,|\, \boldsymbol{\beta} \sim Ga((k+1)/2, (h+\beta_i^2)/2)$$

Binary Regression Models

Factor regression: Special case, as in linear model

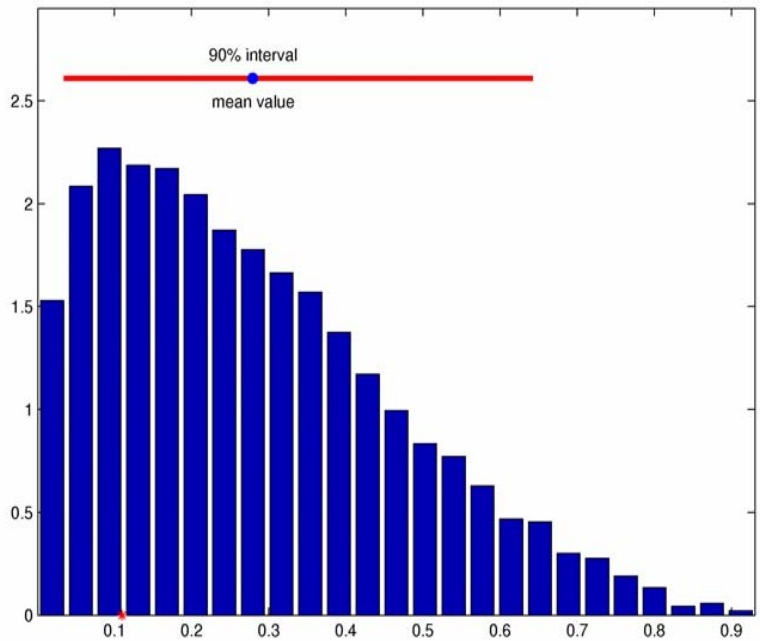
Inferences:

posteriors (samples, histograms), point estimates. intervals …

- regression parameters (factors, genes)
- probabilities

$$\Pr(z_j = 1) = \pi(\mathbf{h}'_j \boldsymbol{\beta})$$

- fitted model
- out-of-sample predictions



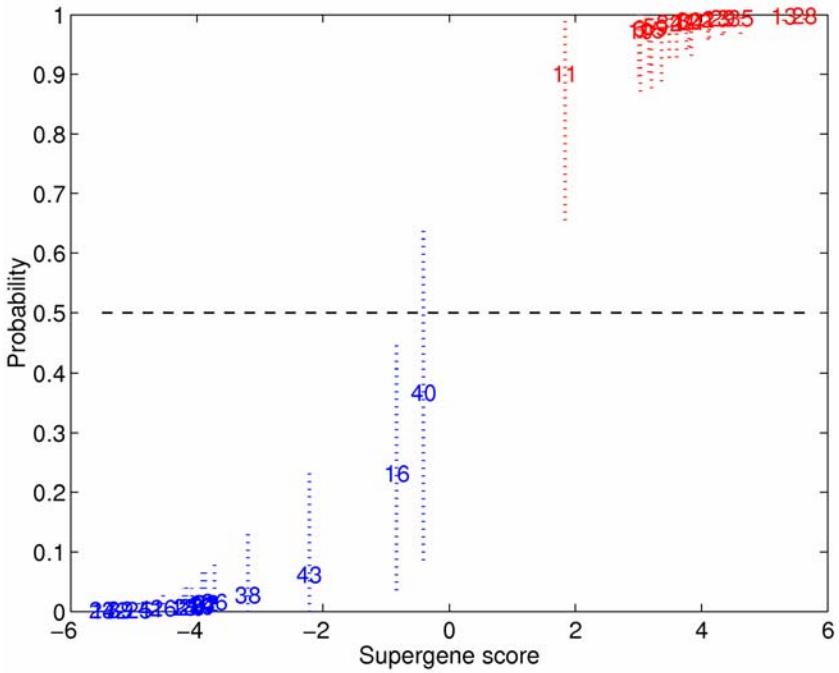
Duke PNAS 2001 Breast Cancer Data

ER example

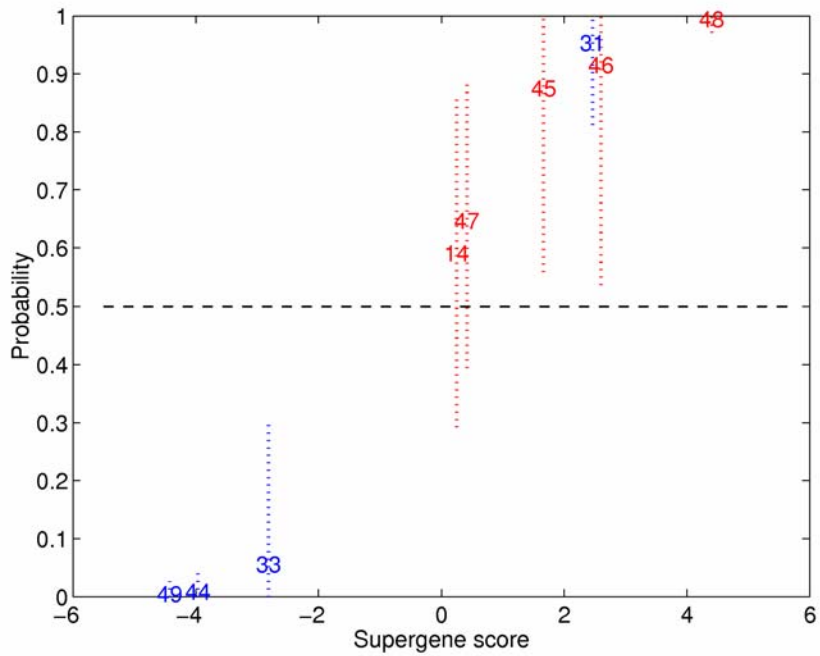
Factor regression: 5 main factors in 100 'top' genes

Gene selection/screening: Reduce "noise" in key factor(s)

Fitted probs: Training data



Predictions: Hold-out cases



Cross-Validation Assessment

Leave-one-out analysis:

repeat n times, predict hold-out

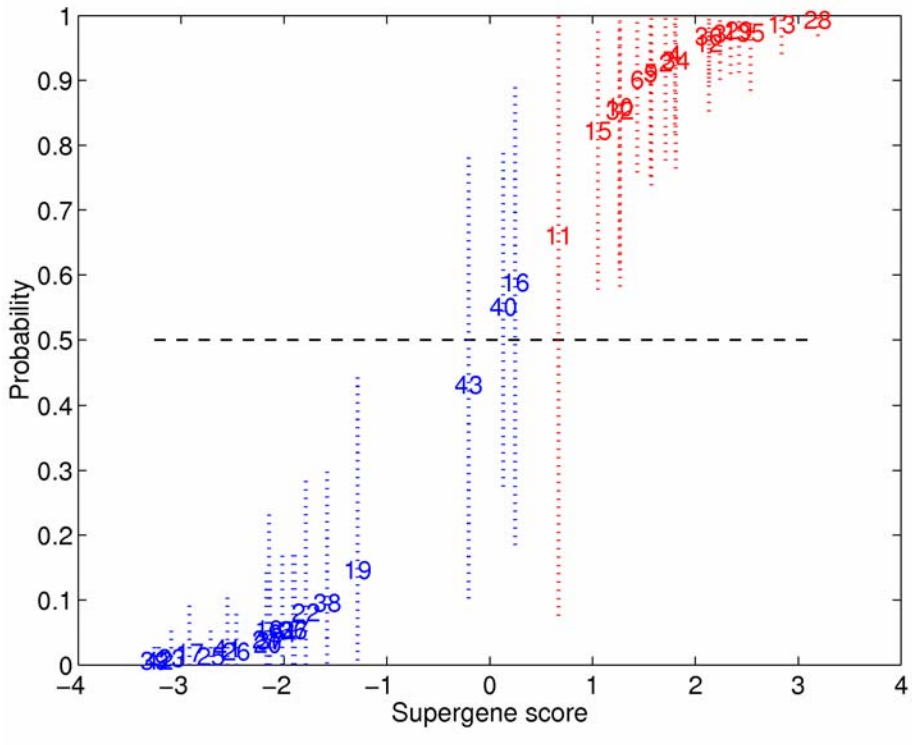
"Honest" assessment of precision

Feature/Variable selection

Repeat CV analysis – varying (overlapping) subsets selected

Heterogeneity, small samples

Critical (dominant) component of predictive assessment



Duke PNAS 2004 Breast Cancer Data
Extended ER example

Examples (see Matlab code explorations)

Explore ER (0/1) regression using a small set of genes related to ER gene by expression, and also Erb-B2/Her-2-nu oncogene, also ER related

Explore factor regression using factors on 50,100.. etc ER 'top' genes

Explore out-of-sample prediction (ERlevel=1,2 based on model fit to ERlevel=0 versus ERlevel=1 cases

Explore leave-one-out cross-validation predictions for real model assessment

Duke PNAS 2001 Breast Cancer Data

"Interesting" cases: 16,40,43

ER gene 'down', other ER+ genes 'up

Conflicting info increases prediction uncertainty

Why conflict?

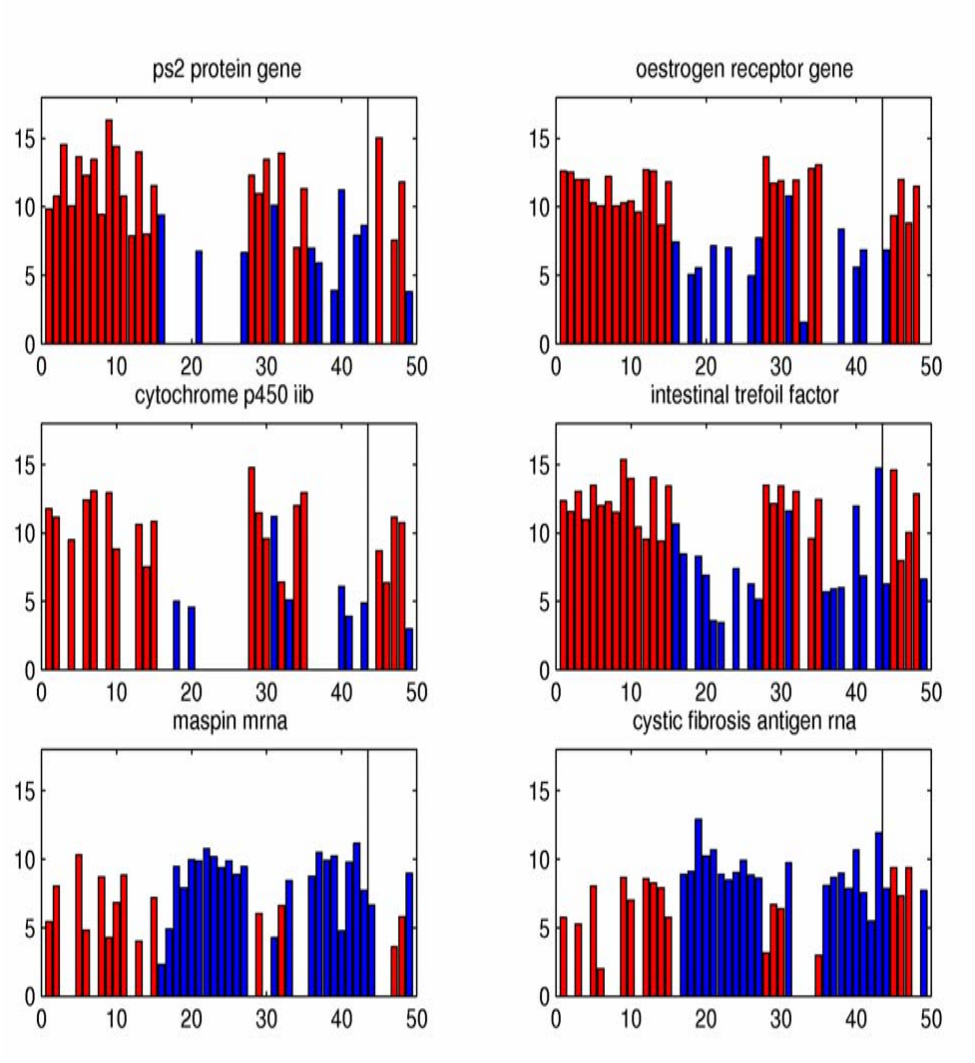
ER protein assay errors

Noise in expression

Temporal changes?

....

Error in visual inspection of antibody staining for ER



More on "Conflicting" expression patterns

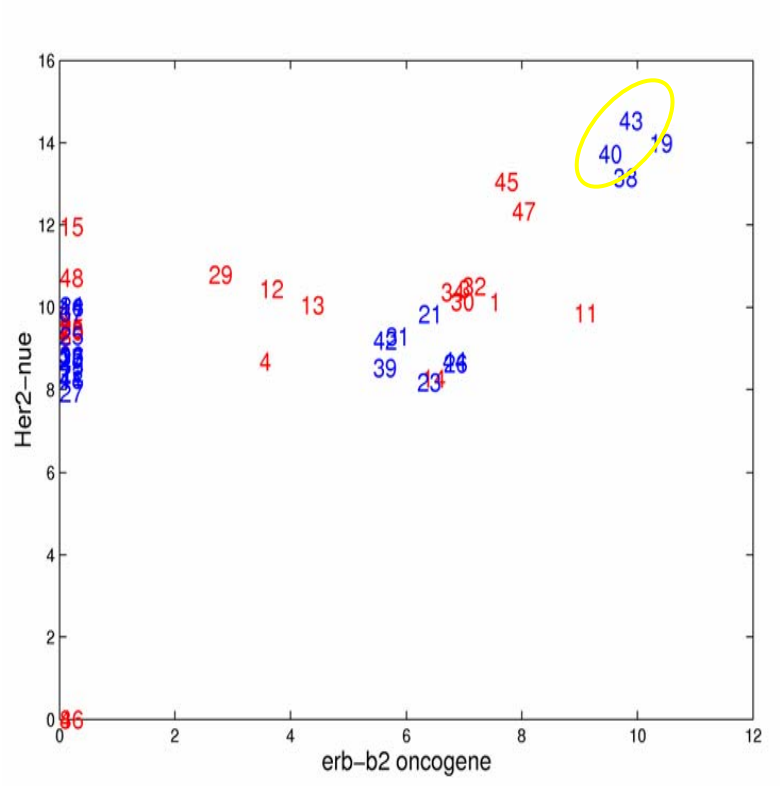
Natural heterogeneity?

Interacting Pathways

induce mixed signals – heterogeneity in expression related to a single outcome

* Other regulators of Ps2, Liv-1, …,
* Subtypes of ER- tumours:
  - i) Basal
  - ii) 20-30% up-express oncogene Erb-B2/Her2-nu

(Perou, Botstein Nature 2000)



Cases 40, 43 here

More on "Interesting" Samples

6 initial validation (hold-out) samples

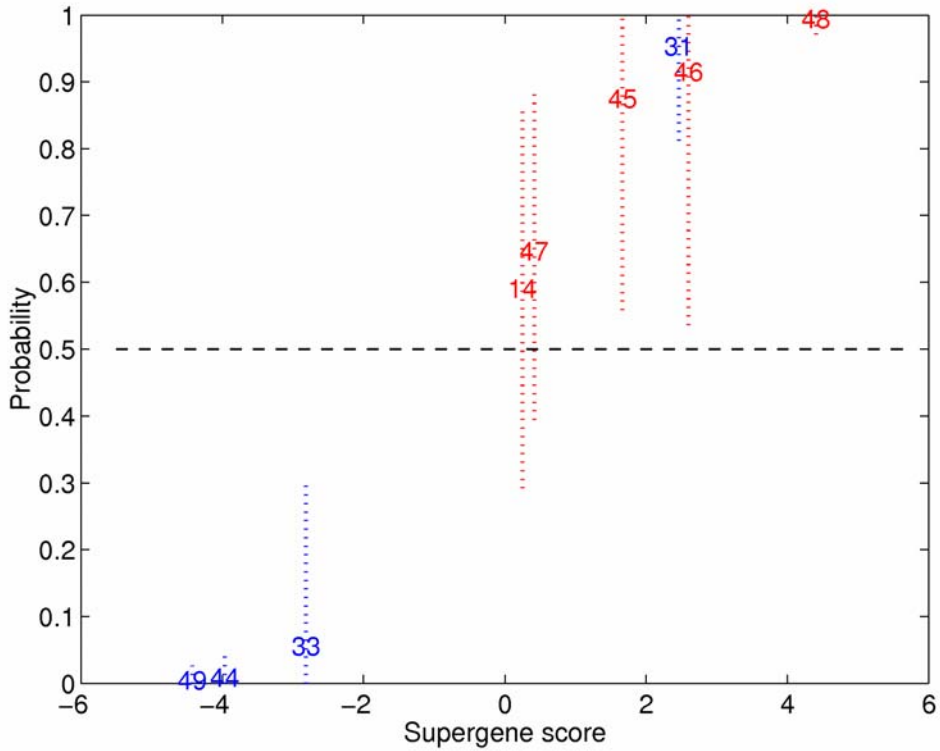
… plus 3 cases: #14,31,33

ER protein IHC assay checked following initial statistical analysis

Western blot test

Cases 14,31,33 reversed – uncertainties

Add to validation set



Duke PNAS 2001 Breast Cancer Data

Axillary Lymph Node Metastatis

Tumours metastasized

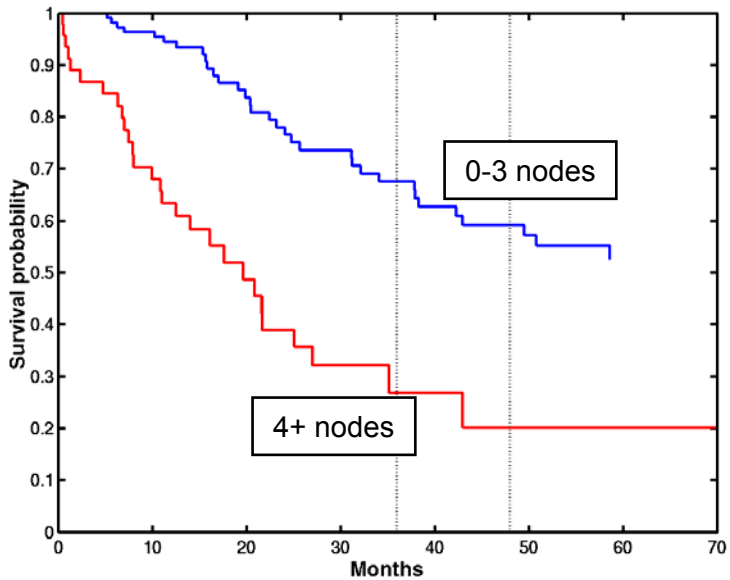
Key clinical risk factor

- recurrence predictor
- treatment decisions

Gene expression in primary tumour:

- predict aggressive cases
- replace nodal surgery

avoid morbidity, cost



Early assessment: Tumours "poised" to metastasize?

Missed diagnoses: False negatives

Duke PNAS 2001 Breast Cancer Data

Axillary Lymph Node Metastatis

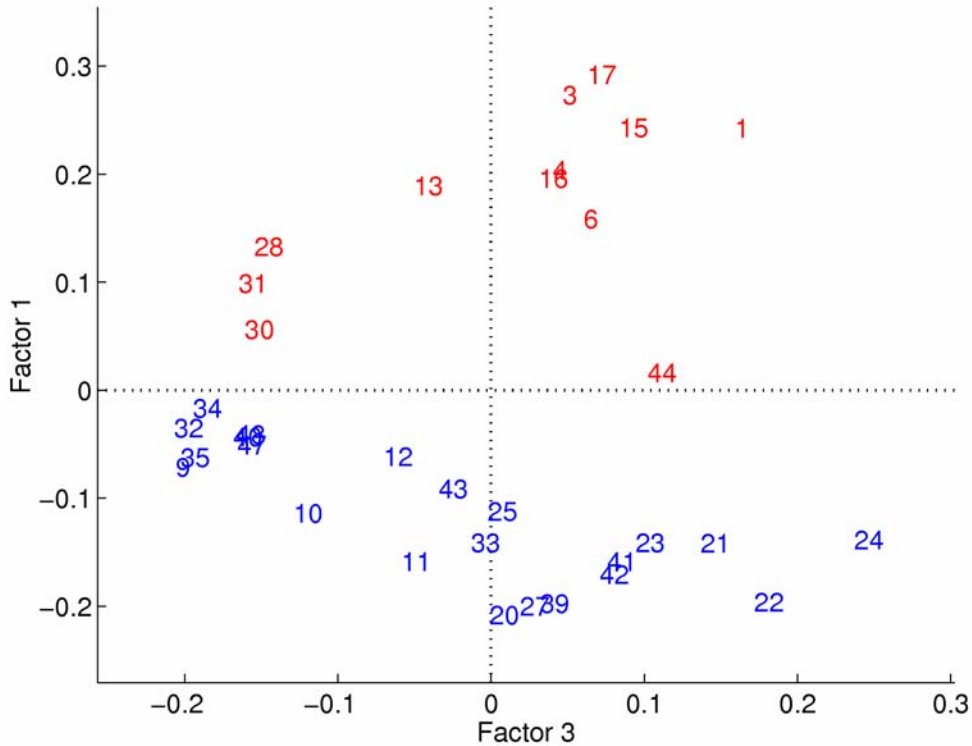
Binary example:
Node negative vs >4 positives

Interesting case:

#44: 0/17 nodes

… BUT

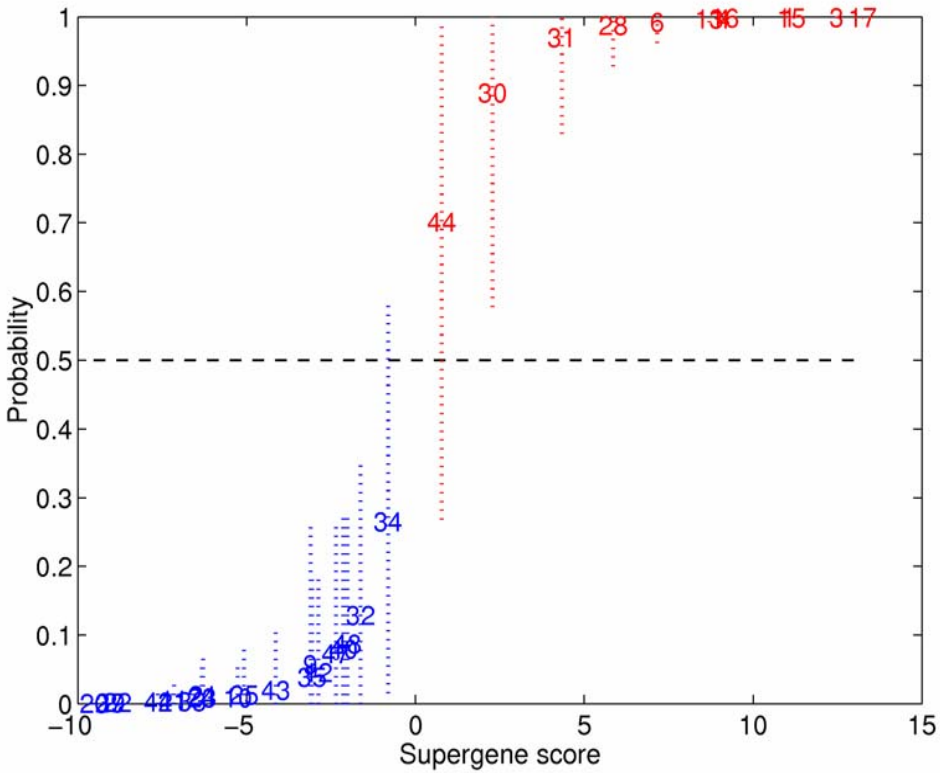
positive intramammary lymph nodes

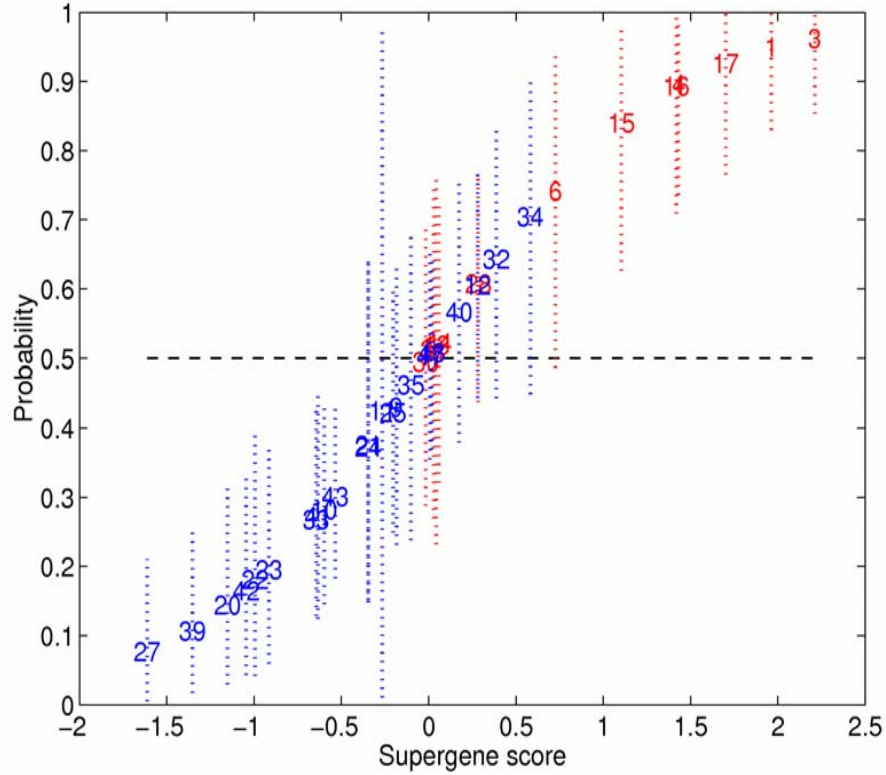


Duke PNAS 2001 Breast Cancer Data

Predicting Metastatis

CV predictions on 100 top genes





CV predictions properly reselecting 100 genes each analysis

Heterogeneity: less clear cut

Example: MIT ALL/AML Leukemia Data

(Golub et al Science 1999)

2 Leukemias: ALL(1) AML(0)

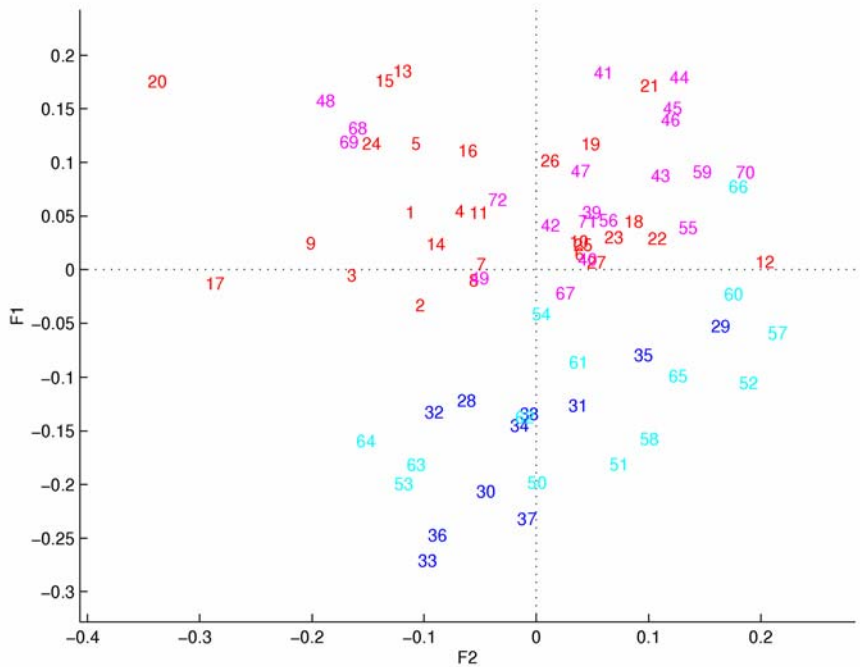
Clear non-genomic differences

38 training samples (27/11)
34 validation (20/14)

MIT: data-based screen p=3571

Some difficulty in predictive classification of 5 cases

2 factors: all genes



Training samples: ALL AML
Validation samples: ALL AML

MIT ALL/AML Leukemia Data

Binary factor regression:

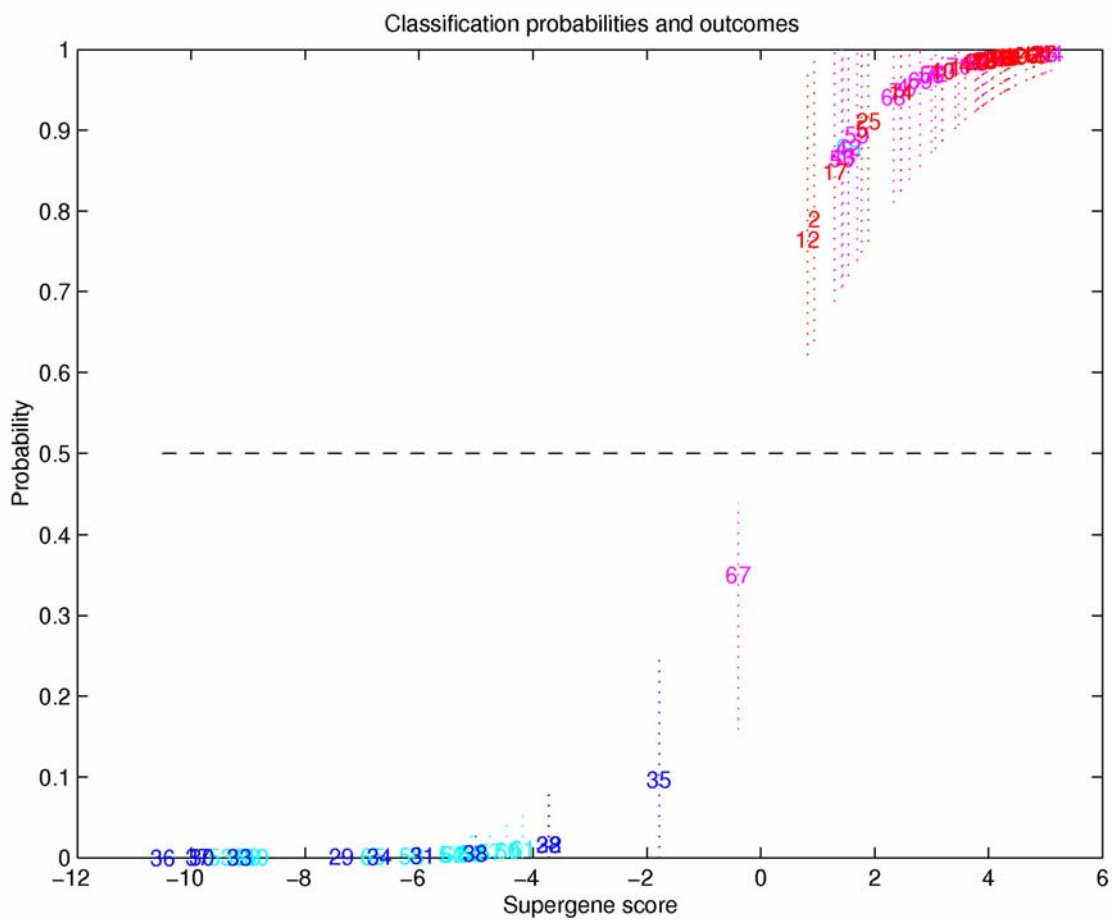
50 genes

Fitted training data, predicted validation

Unrealistically 'clean'

Cases 66, 67?

'Misclassified' cases?



MIT ALL/AML Leukemia Data

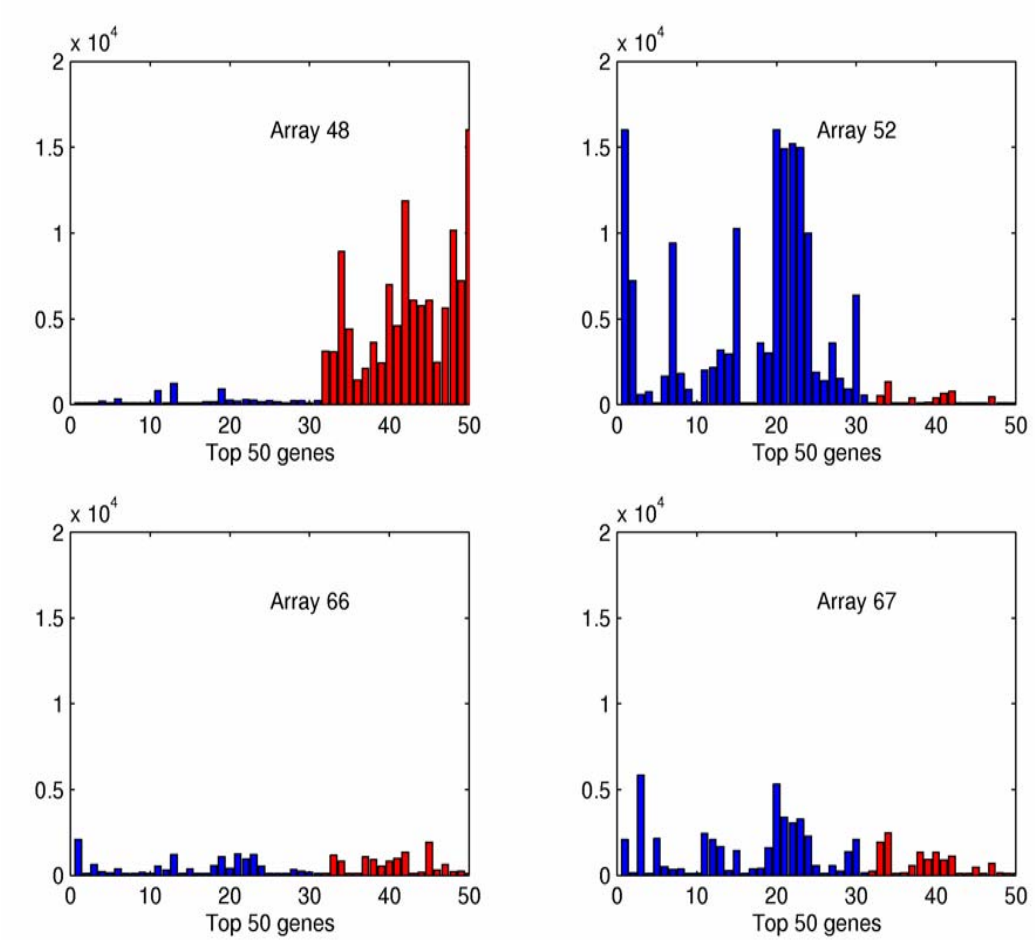
Cases 66, 67?

Probably data quality

Poor normalisation?

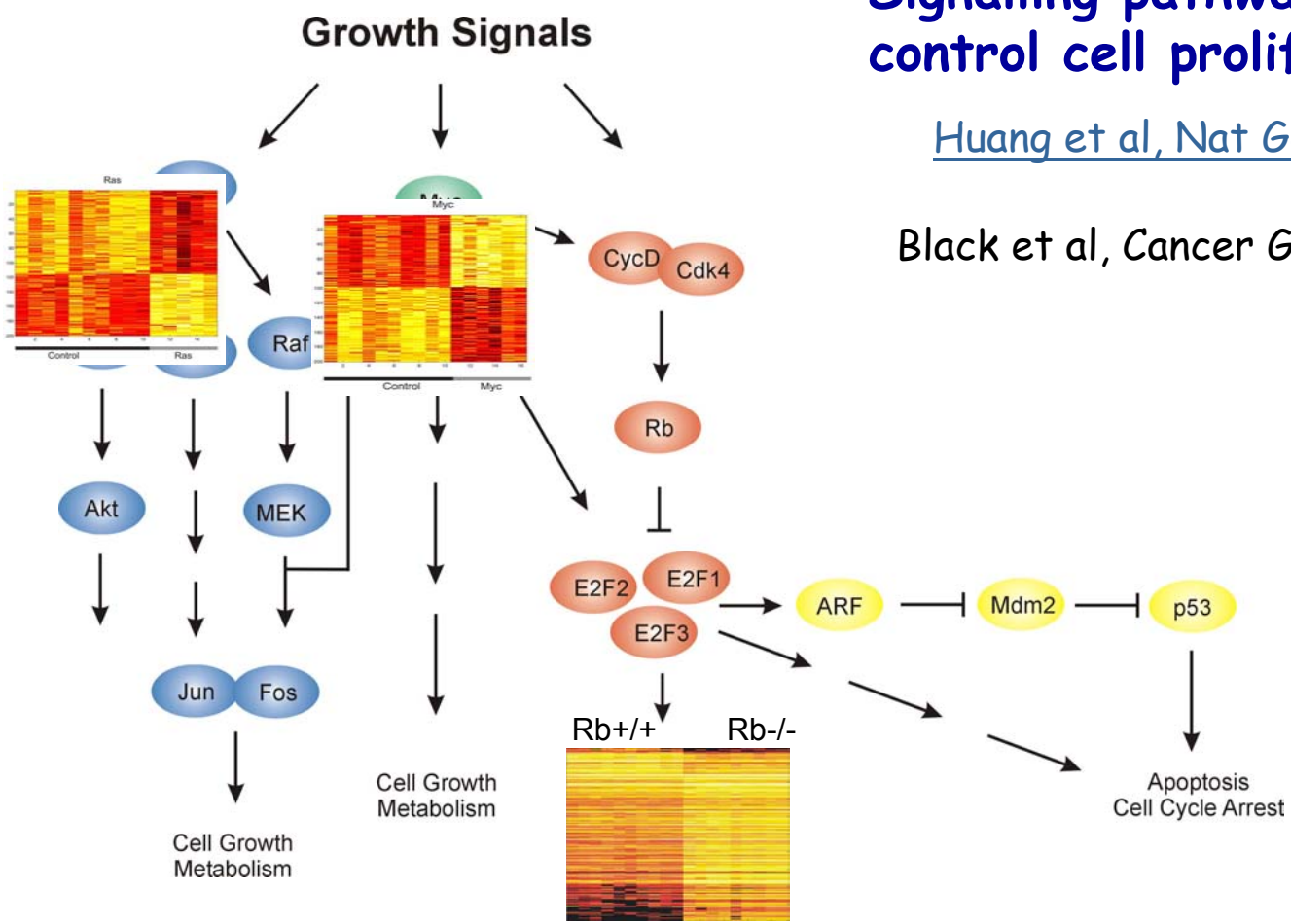
Low/degraded RNA?

Expression levels: top 50 genes on 4 arrays



Predictive Profiling

- Signatures of Oncogenic Pathway Deregulation -

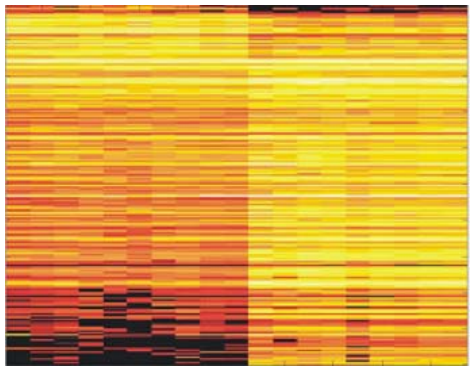


Signalling pathways that control cell proliferation

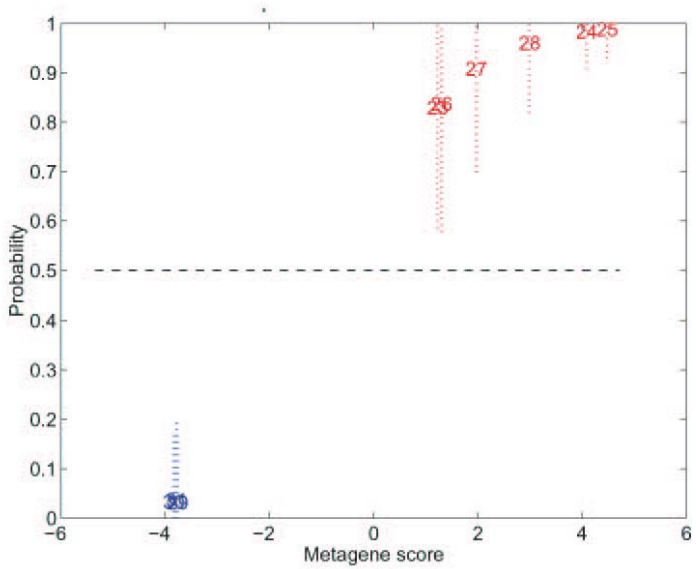
Huang et al, Nat Gen 2003

Black et al, Cancer Gen 2003

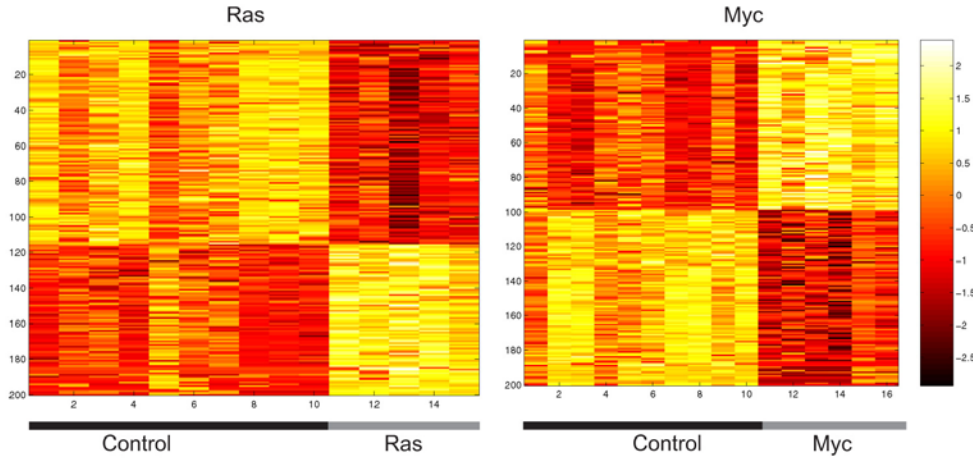
Signatures from Binary Regression in Cell Line Experiments



Rb knockout



Myc, Ras over-expression



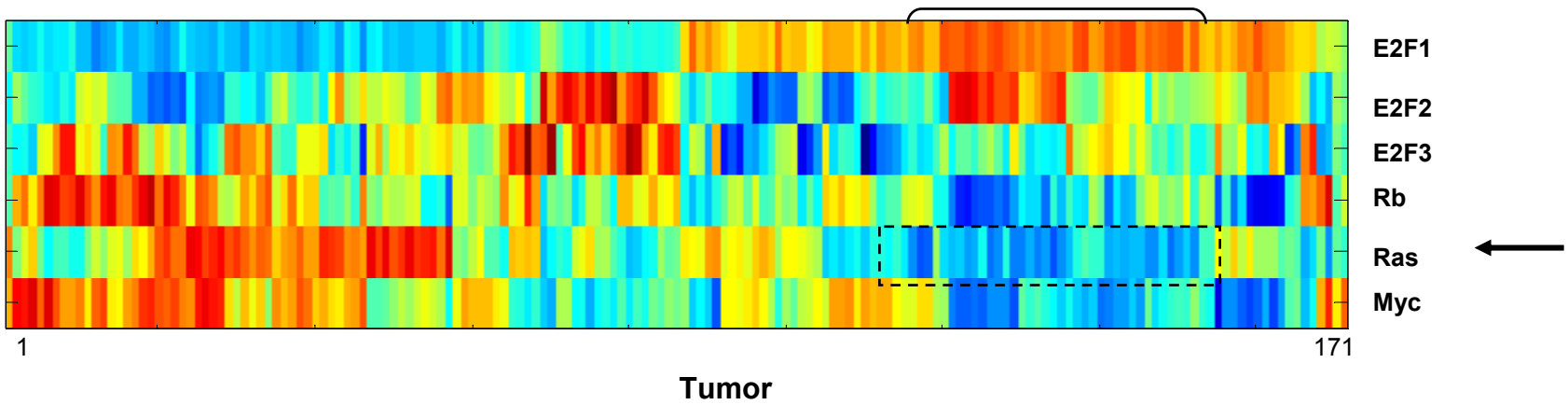
Out-of-sample prediction

Signatures Predict Differences in Oncogenic Activity in Mouse Tumours

Huang et al, Nat Gen 2003

Signatures of Oncogenic Deregulation

current studies to define multiple signatures
for charactising human tumour states



ABSS04