Statistical analysis and predictive discrimination using DNA microarray data

Genomic features, patterns

$\updownarrow$

Physiological characteristics, clinical outcomes

Mike West

www.isds.duke.edu

Institute of Statistics & Decision Sciences

Duke University

Collaborators

Joseph Nevins	Genetics
Holly Dressman	Genetics
Jeff Marks	Surgery & Cancer Center
Carrie Blanchette	Surgery & Cancer Center
Rainer Spang	ISDS & Genetics
Harry Zuzan	ISDS & Genetics
Mike West	ISDS

Center for Bioinformatics & Computational Biology
Center for Genome Technology

(Breast cancer) discrimination

Two group problems: Binary outcomes

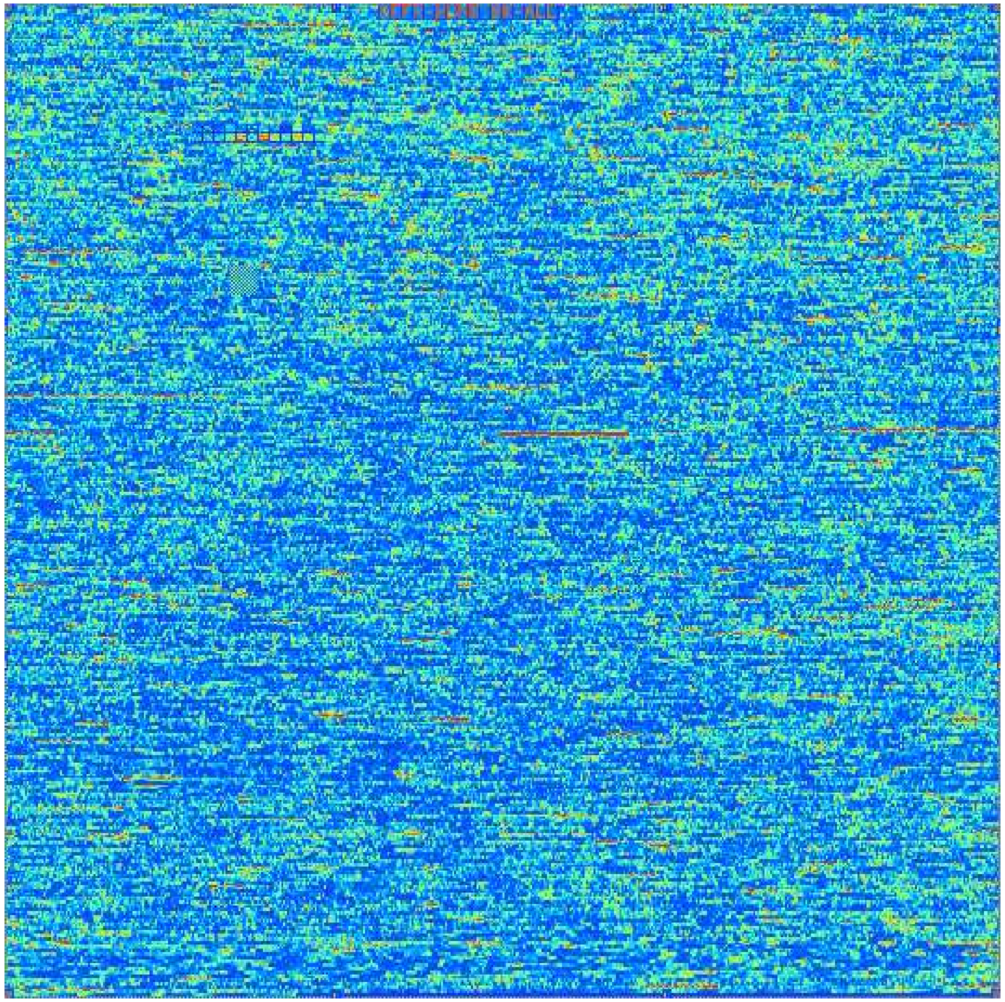
- e.g., ER+ versus ER−
- e.g., lymph node + versus lymph node −
- DNA microarray data: expression levels of $\approx$ 7000 genes (sequences) in RNA from tumour, tumour location, time point, ...
- 23 ER+, 20 ER−
- Discriminatory patterns of expression?
- Predictive validity? Predictive classification of tumours 50, 51, ...?
- Which genes are implicated? Surprises?
- Which tumours depart from general patterns? How?
- ... etc

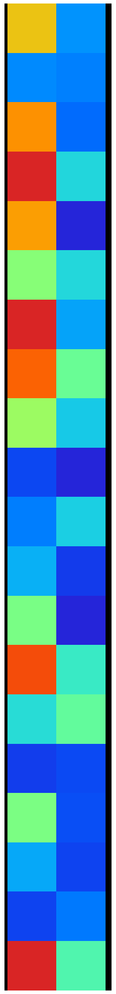
Expression array data

Microarray data: Affymetrix arrays

- ≈ 7000 genes (sequences)
- Data issues:
  - imaging, probe cell specific expression
  - data summaries in commercial software
  - ...
- Estimates of expression level by gene: **Absolute difference**
- Here: $\log_2(\max(1, \mathrm{AbsDiff}))$

One array, one probe set





Projecting large-scale expression data

- Binary regression: many predictor variables
- Possibly many interacting genes relate to status
- **Singular factor projection** of expression data
  - reduces dimension with no loss of information
  - summarises "important structure" in expression data
- Principal components decomposition
- Variances and correlations in expression fully "explained" by small number of factors
- Expression of (many) genes "driven" by (few) factors

Summary expression data

Notation:

- $x_{i,j}$ is expression level of gene i on microarray j
- p genes, n arrays: $n << p$

$$\mathbf{X} \quad = \quad [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \quad = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ x_{p,1} & x_{p,2} & \cdots & x_{p,n} \end{pmatrix}$$

Singular value (factor) decomposition

$$\mathbf{X} = \mathbf{ADF}$$

Factor loadings matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$

- patterns/relationships among genes

Latent factors are rows of $\mathbf{F}$

- patterns/relationships among arrays: $n << p$ factors

Supergenes=Factors: linear combinations of expression

Factors "drive" expression levels: gene i on array j :

$$x_{i,j} = a_{i,1}f_{1,j} + a_{i,2}f_{2,j} + \ldots + a_{i,n}f_{n,j}$$

Binary regression modelling

- Microarray j, expression profile $\mathbf{x}_j$
- Binary classification: 1 (ER+) or 0 (ER−)
- Probability array j is ER+ is $\pi(\mathbf{x}_j)$
- Standard probit model: $\pi(\mathbf{x}_j) = \Phi(\beta_0 + \mathbf{x}_j'\boldsymbol{\beta})$
- Linear regression on gene expression, mapped to probability scale
  - $\mathbf{x}_j'\boldsymbol{\beta} = \sum_{i=1}^{p} \beta_i x_{i,j}$
  - β_i is regression coefficient on gene i
- Statistical analysis: estimate coefficients, uncertainty

Supergenes in binary regression modelling

Regression on (many) genes reduces to regression on (few) supergenes

$$\mathbf{X}'\boldsymbol{\beta} = \mathbf{F}'\boldsymbol{\theta} \qquad\qquad \boldsymbol{\theta} = \mathbf{D}\mathbf{A}'\boldsymbol{\beta}$$

- n parameters, sample size n
- Ignore "stable" factors
- Use of stochastic regularisation: priors on $\boldsymbol{\theta}$
  - elements θ_j independent (orthogonality)
  - proper, "diffuse" priors: $\theta_j \sim T_k(0,1)$
  - neutral: implied priors for classification probability $\pi(\mathbf{x}_j)$
- Efficient analysis to estimate $\boldsymbol{\theta}$
- Markov chain Monte Carlo model fitting

Theoretical context and issues

- $\boldsymbol{\theta}$ *depends on design data* $\mathbf{X}$
- New arrays: new parameter, new priors
- Out-of-sample prediction: New tumours
- SVD analysis of **all** arrays

- **Underlying latent factor model genesis**
- SVD regression as a limiting case
- Consistent priors for $\boldsymbol{\theta}$ and underlying gene coefficients $\boldsymbol{\beta}$ as new data arises
- Generalised "g-prior"

Underlying latent factor models

Latent factor model for gene expression: tumour i

$$\mathbf{x}_i = \mathrm{B}\boldsymbol{\lambda}_i + \boldsymbol{\epsilon}_i$$

- $\boldsymbol{\lambda}_i \sim N(\mathbf{0}, \boldsymbol{I})$ and $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Psi})$
- patterns explained by (a few) latent factors: k =dim($\boldsymbol{\lambda}_i$)
- residual/idiosyncratic terms $\boldsymbol{\epsilon}_i$

Outcomes:

$$y_i \sim N(\boldsymbol{\lambda}_i'\boldsymbol{\theta}, 1)$$

- outcomes regress on latent factors in $\mathbf{x}_i$ – indirect regression on $\mathbf{x}_i$
- different outcomes relate to different latent factors

Underlying latent factor models: SVD regression case

- Latent factor model defines $p(y_i, \mathbf{x}_i, \boldsymbol{\lambda}_i)$
- Implied $p(y_i|\mathbf{x}_i)$**:** regression of y_i on $\mathbf{x}_i$
- Linear regression coefficient $\boldsymbol{\beta} = \mathbf{H}\boldsymbol{\theta}$
- $\mathbf{H}$ depends on $\mathbf{B}, \boldsymbol{\Psi}$

Some implications:

- Prior on $\boldsymbol{\theta}$ implies generalised $g-$prior on $\boldsymbol{\beta}$
- Limiting case: $\boldsymbol{\Psi} \rightarrow \mathbf{0}$ leads to SVD regression

Regression on genes via supergenes

- Efficient analysis of regression on $n << p$ supergenes
- Posterior (samples) for supergene vector $\boldsymbol{\theta}$
- Compute posterior (samples) $\boldsymbol{\beta} = \mathbf{AD}^{-1}\boldsymbol{\theta}$
- Bayesian/model justification of generalised inverse to $\boldsymbol{\theta} = \mathbf{DA}'\boldsymbol{\beta}$

Honest prediction and model assessment

Critical **predictive** assessment of discriminatory performance

- Predictions of new cases: validation sample
- "One-at-a-time" cross-validation of training data:
  - Take out microarray j
  - Fit model: Predict status of tumour j
  - Repeat for all arrays j

Gene screening

- Heterogeneity in data: "noise" from many "irrelevant" genes?
- Screen to smaller subsets - e.g., raw correlations with ER$+/-$ status
- Select "top k" and fit model on k genes
- Oestrogen receptor status example: $k = 100$
  - Multiple genes refine classification: minor effects
  - **Collective effects in addition to primary gene**

Gene screening in one-at-a-time cross-validation:

Different overlapping subsets of 100 for each hold-out case

Breast cancer data: ER status study

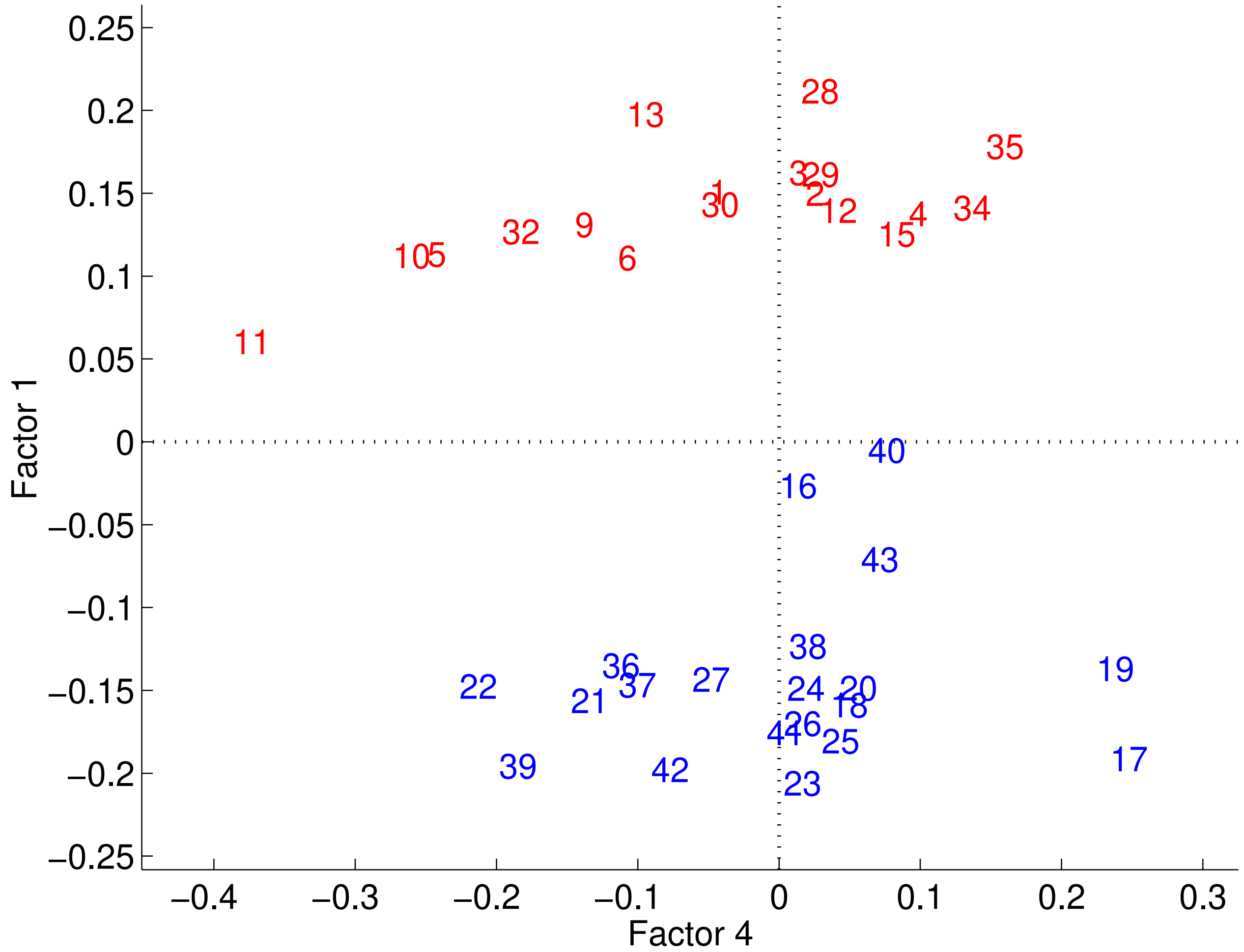
- Two batches: 43 (training sample) and 6 (validation sample)
- Two arrays (#7,8) removed: hybridisation problems, scratches, ...

ER status by immunohistochemical methods (summer 2000)

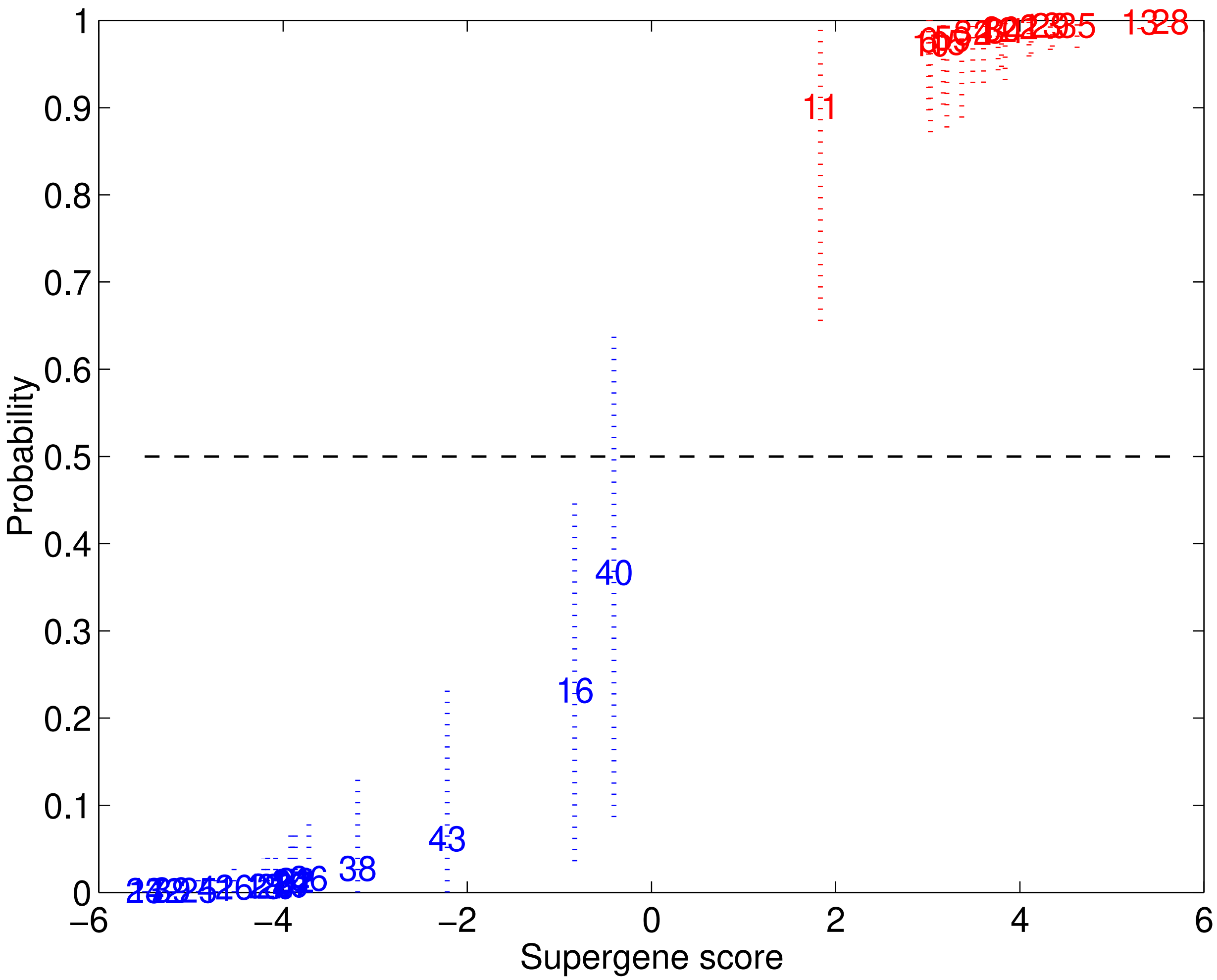
- Initial analysis questions ER+/− for some cases
- Checked by protein blot test (11/2000)
- Most confirmed: 3 cases (#14,31,33) differ
- Treat these 3 cases as of unknown status: add to validation set

38 training cases (18 ER+ and 20 ER−) **9 validation cases**

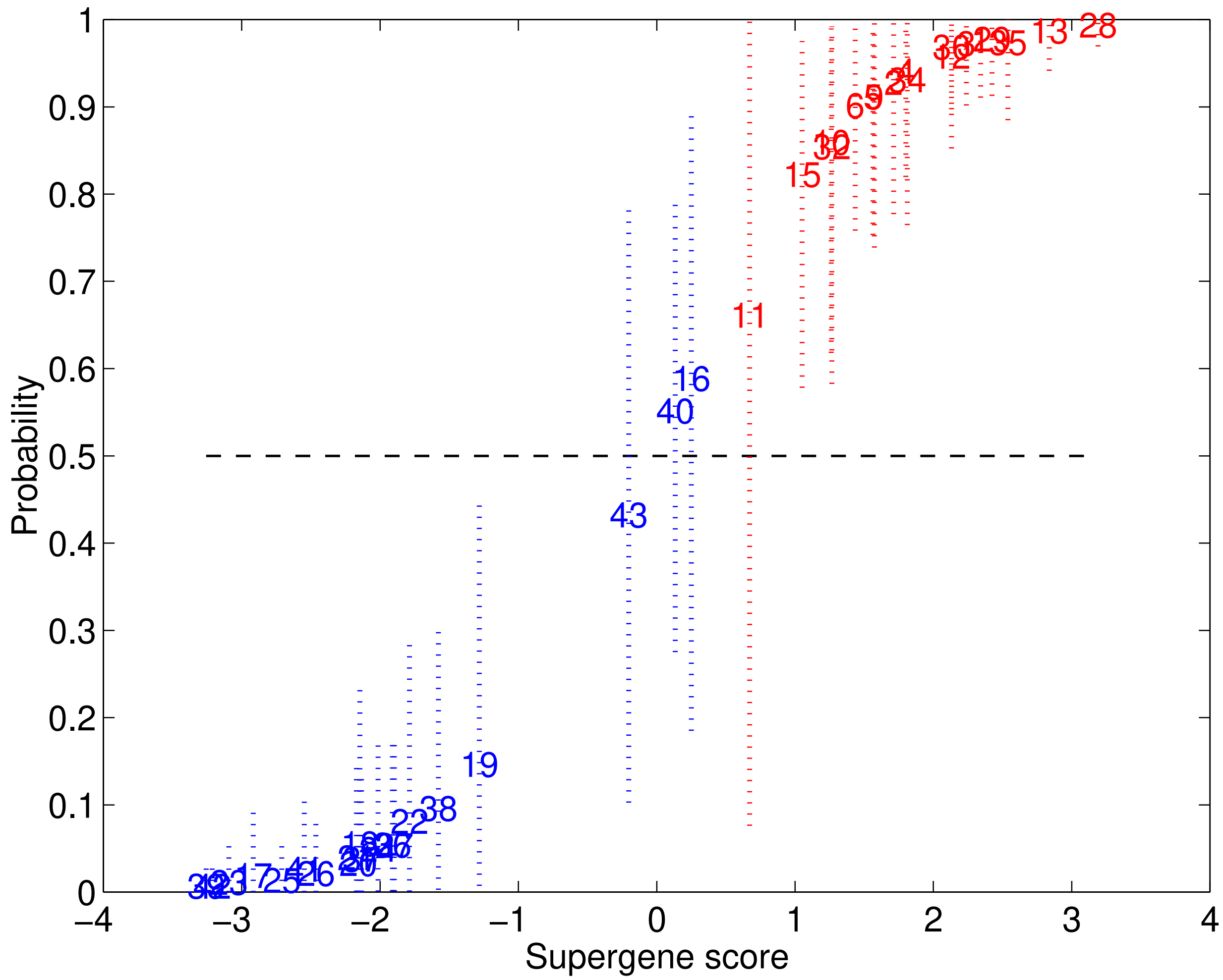
ER: Two factors underlying 100 "top genes"



ER: Fitted classification



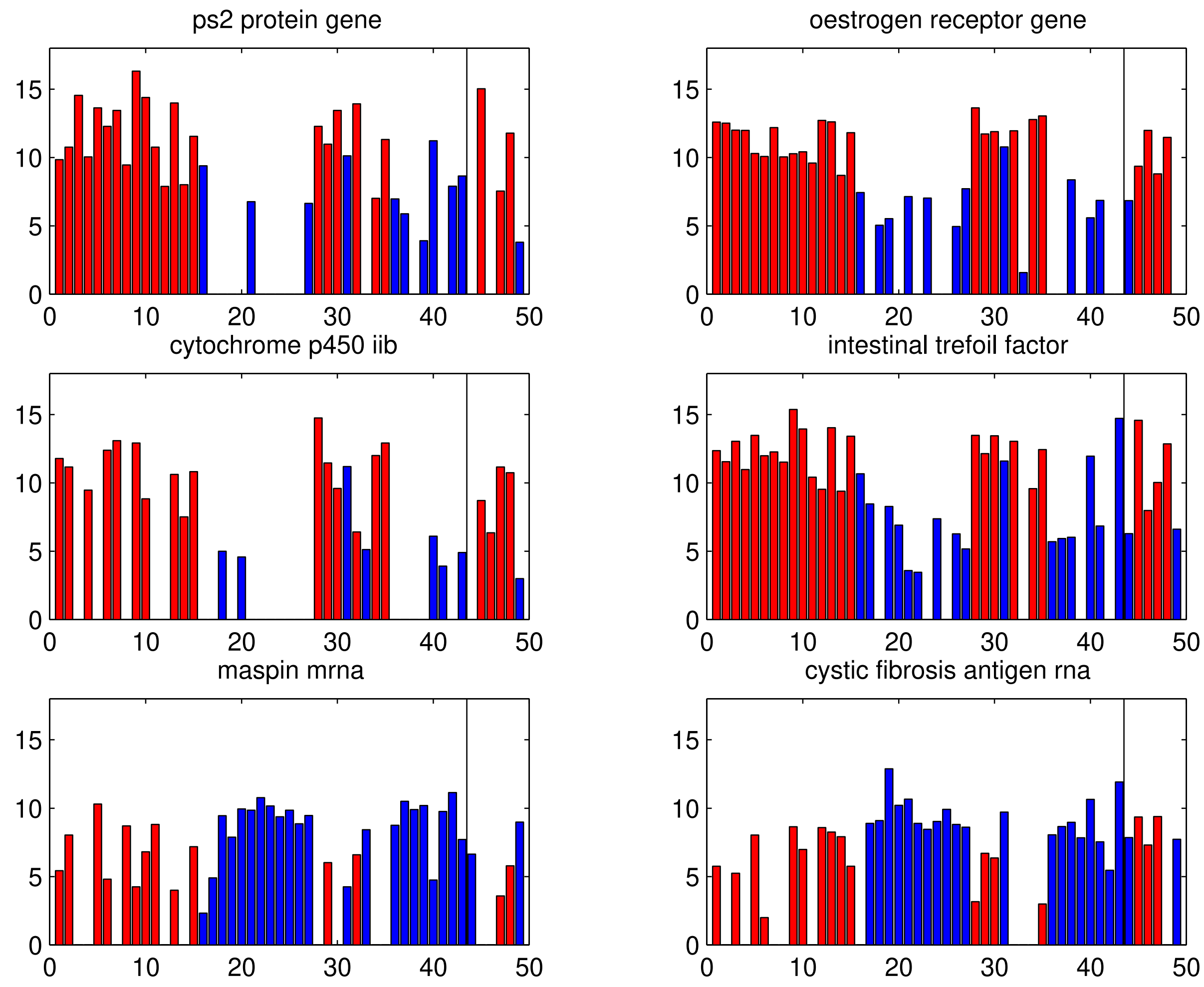
ER: Cross-validatory predictions



ER: Some "top" genes

• **ps2 protein gene (tff-1)**	ER regulated
• **mrna for oestrogen receptor**	receptor
• **cytochrome p450 iib mrna**	growth factor
• **intestinal trefoil factor mrna (tff-3)**	ER co-expressed
• **IGFBP-1**	ER regulated
• **hepsin (hepatoma serine protease)**	
• **Gata-3 tf**	High in ER+ cells
• **maspin**	ER related
• **cystic fibrosis antigen**	ER related; BC marker
• **p37nb mrna**	
• ...	
• **breast cancer, oestrogen regulated liv-1 protein mrna**	oestrogen induced

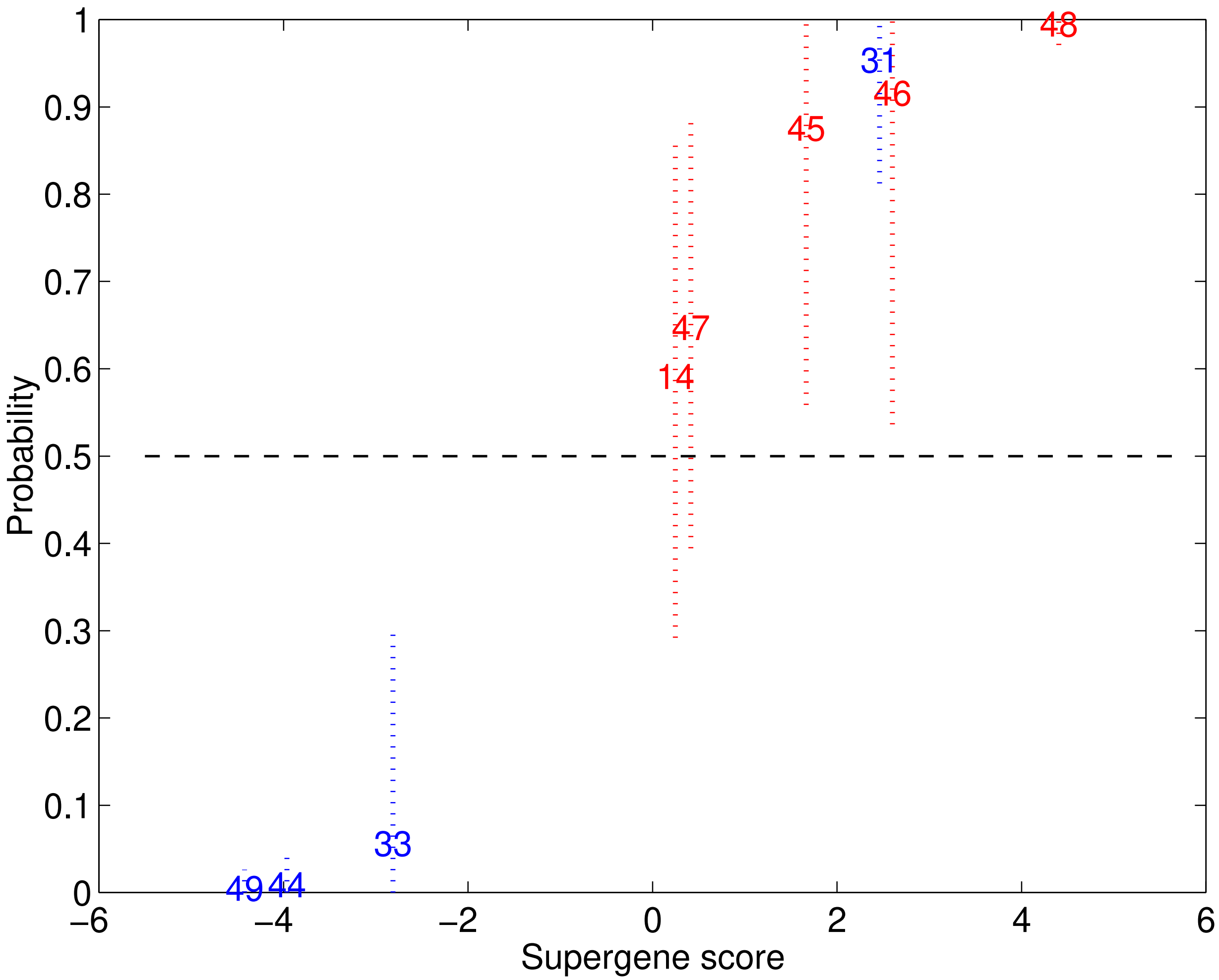
ER: Expression levels of some top genes



Tumours 16,40,43

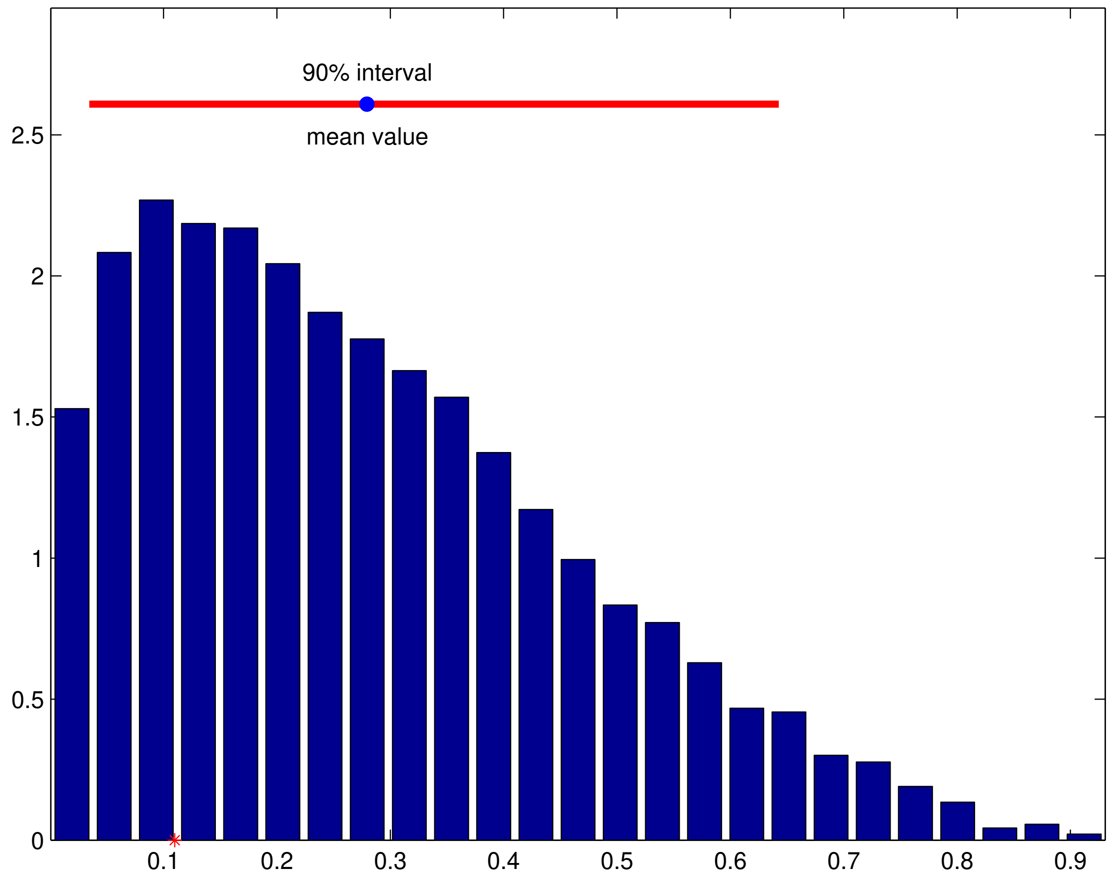
- Similar patterns: ER+ or ER−?
- High uncertainty about $Pr(ER+)$
- Oestrogen gene marginally down; other "up for ER+" genes up
- Mixed/conflicting story
- **High classification uncertainty results**
  - Other regulators of Ps2, Liv-1 ... ?
  - ER status determination ... ?
  - Changing from − to +?

ER: Predictions for validation sample



Classification and uncertainty

Classification probability for tumour 16



Choice of "point estimates" - Mean values "conservative"

Breast cancer nodal status

- Breast cancers classified by axilliary lymph nodal status
- Tumours metastasized to lymph nodes
- Most important risk factor in disease outcomes, therapy decisions

Data & Issues: Reported number of positive nodes

- $0 - 20+$, out of totals $2 - 37$
- crude categorization: reported Node+ versus Node−
- censored totals, “missed” positive nodes?
- tumours poised to metastasize to lymph nodes?

Breast cancer nodal status

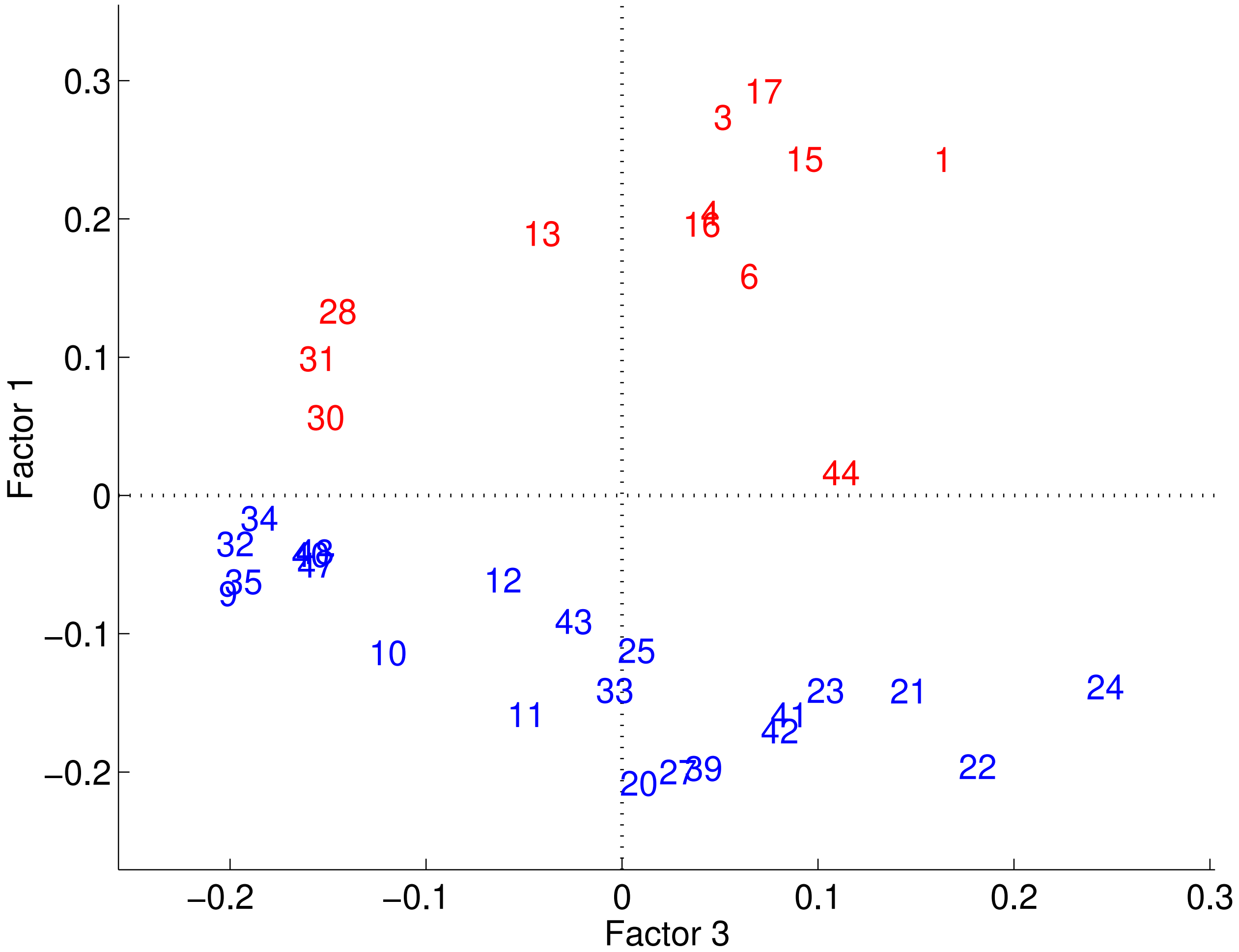
Clinicians define outcomes:

- 0: no positive nodes reported
- 1: at least 3 positive nodes reported

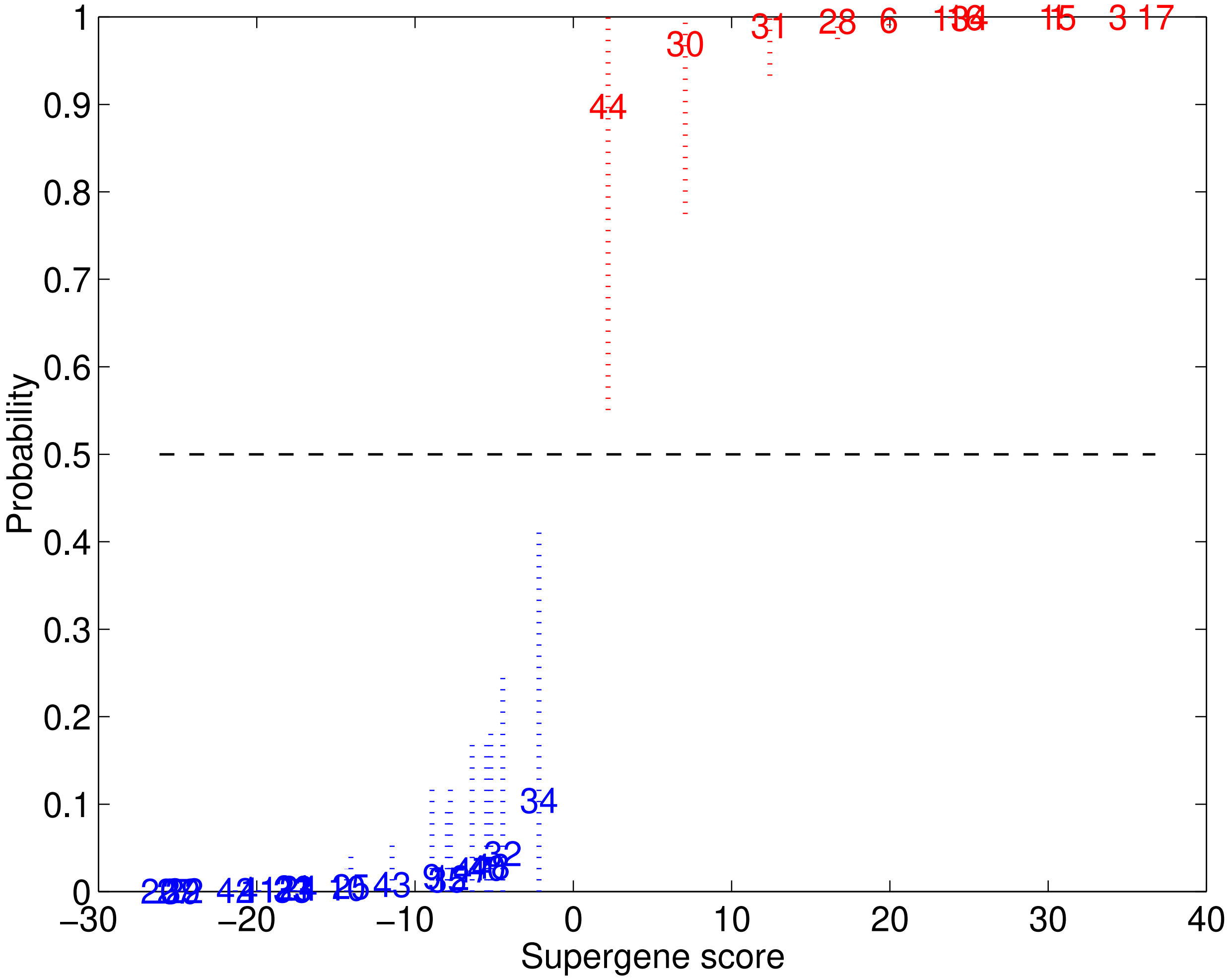
- to predict as validation cases: 1 or 2 positives reported

34 training cases (12 + and 22 −) **13 validation cases**

Nodes: Two factors underlying 100 "top genes"

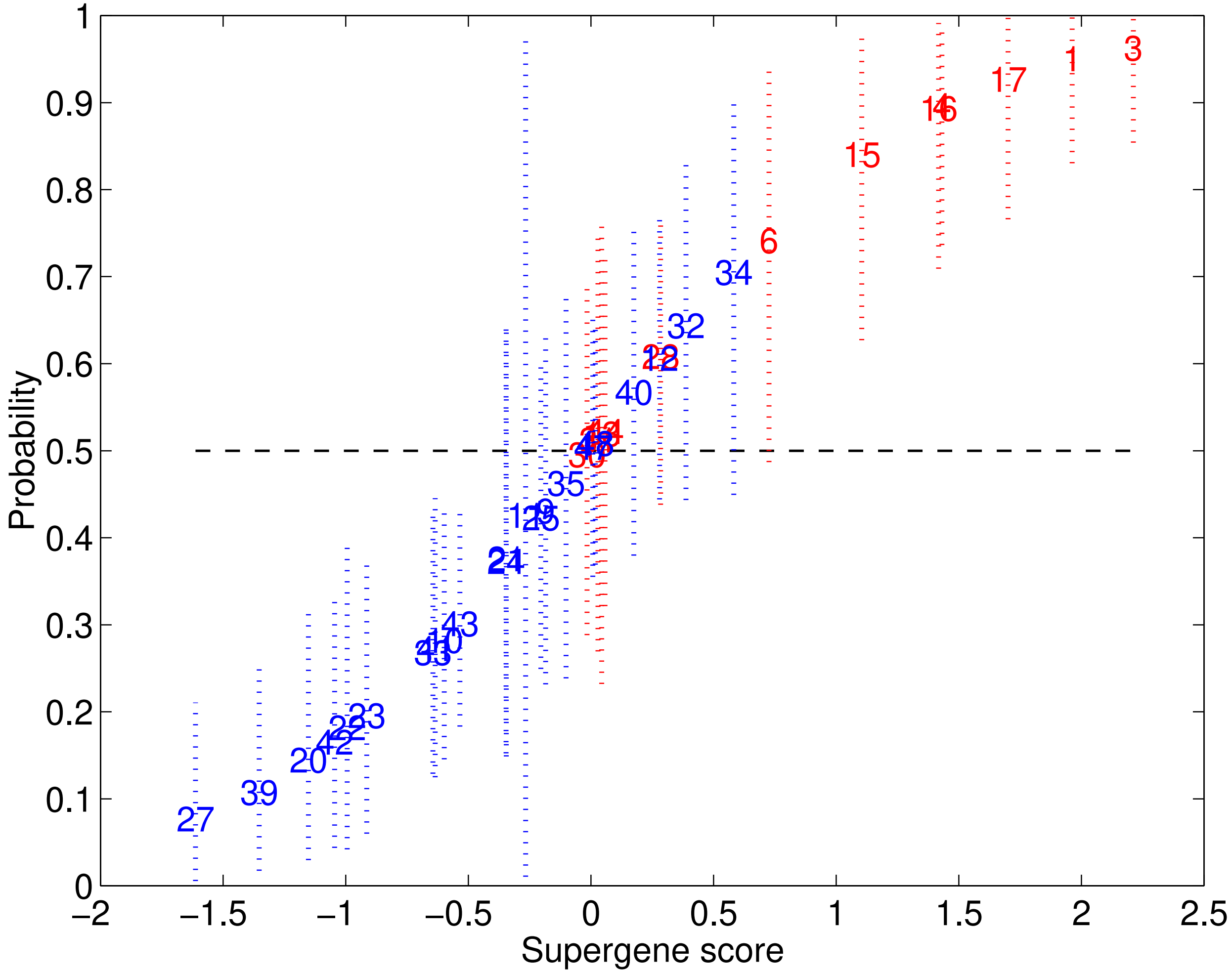


Nodes: Fitted classification

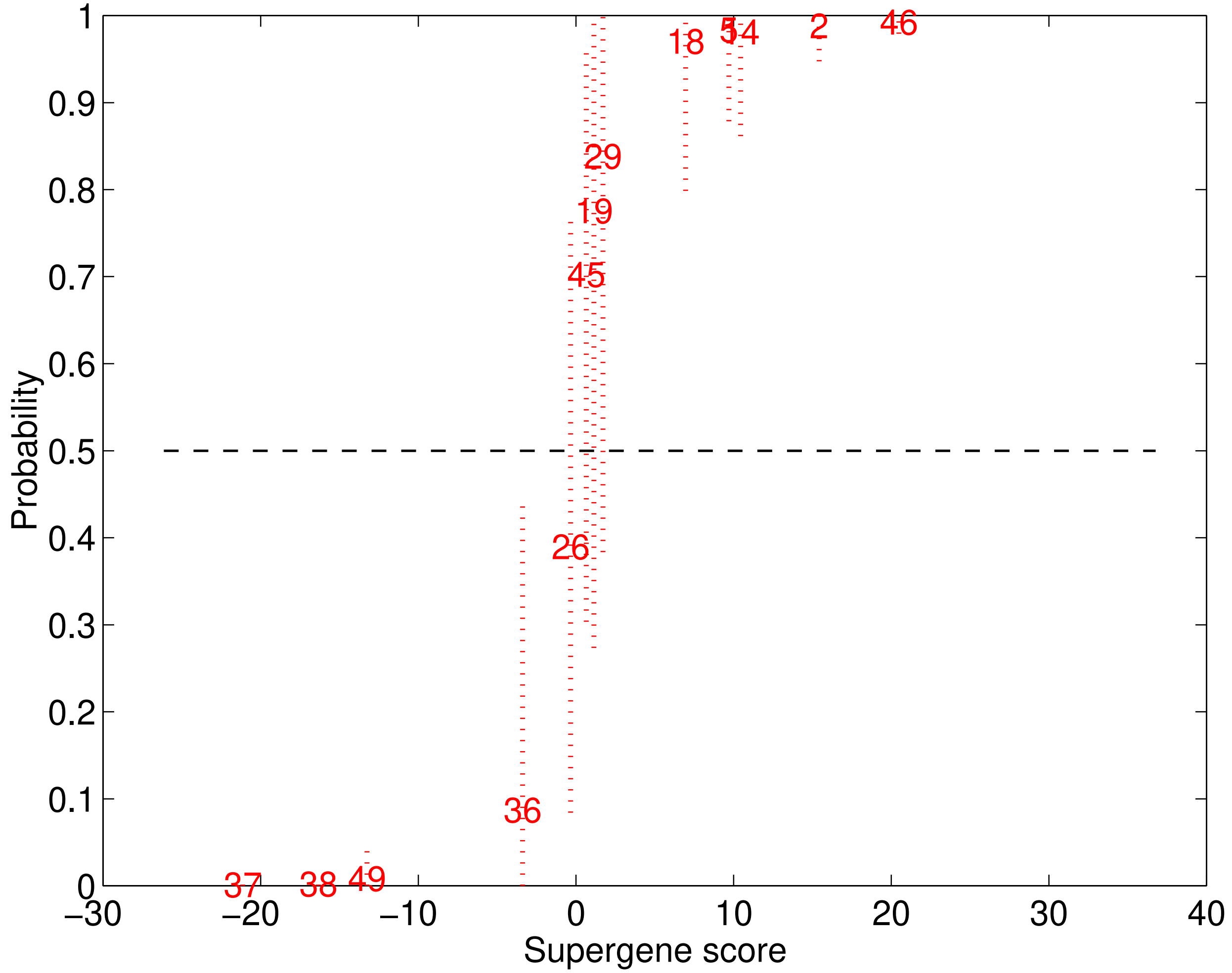


Case 44: 0/17 BUT positive intramammary lymph nodes

Nodes: Cross-validatory predictions



Nodes: Predictions for validation sample



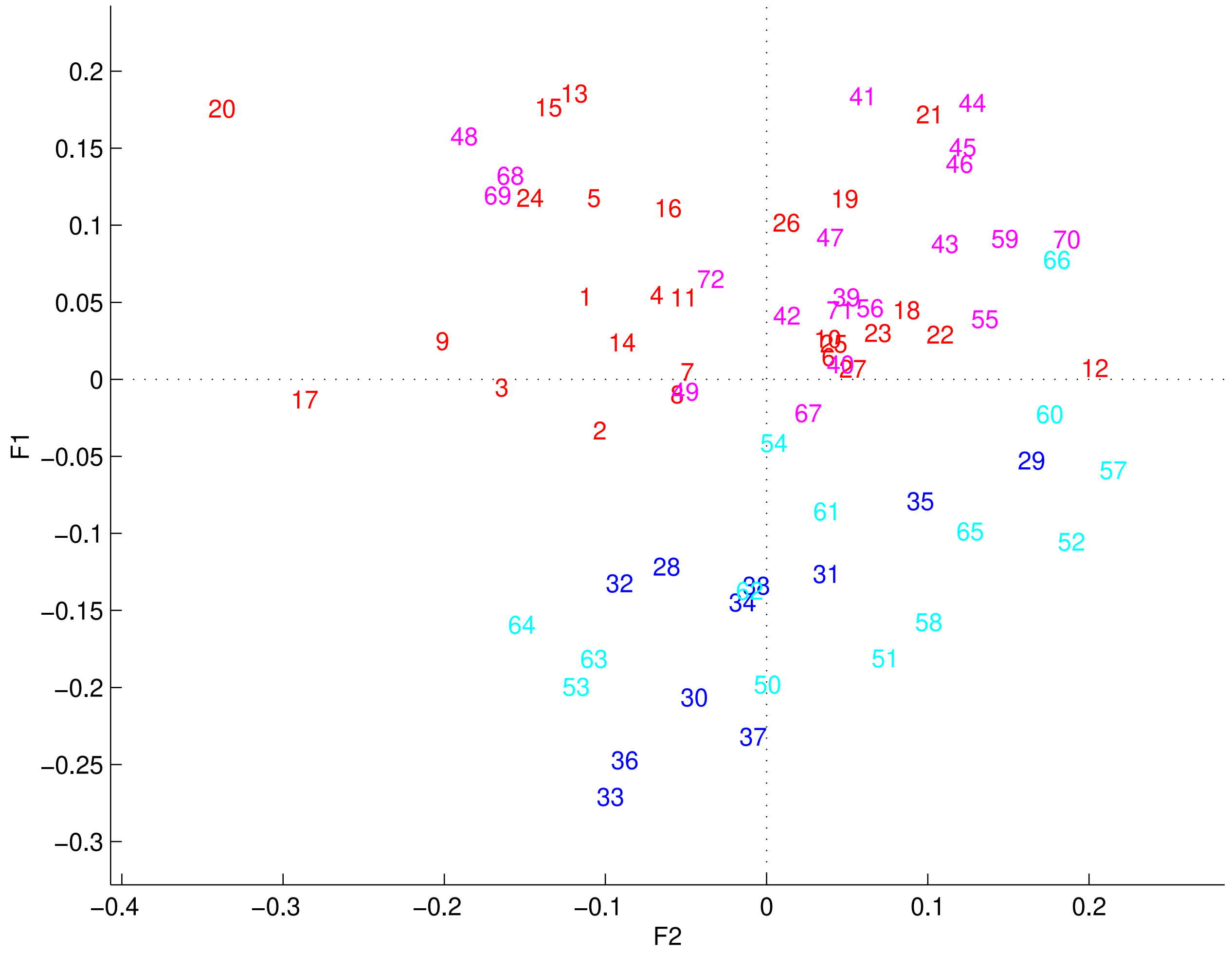
MIT ALL/AML leukemia study

Whitehead Institute, Lander group

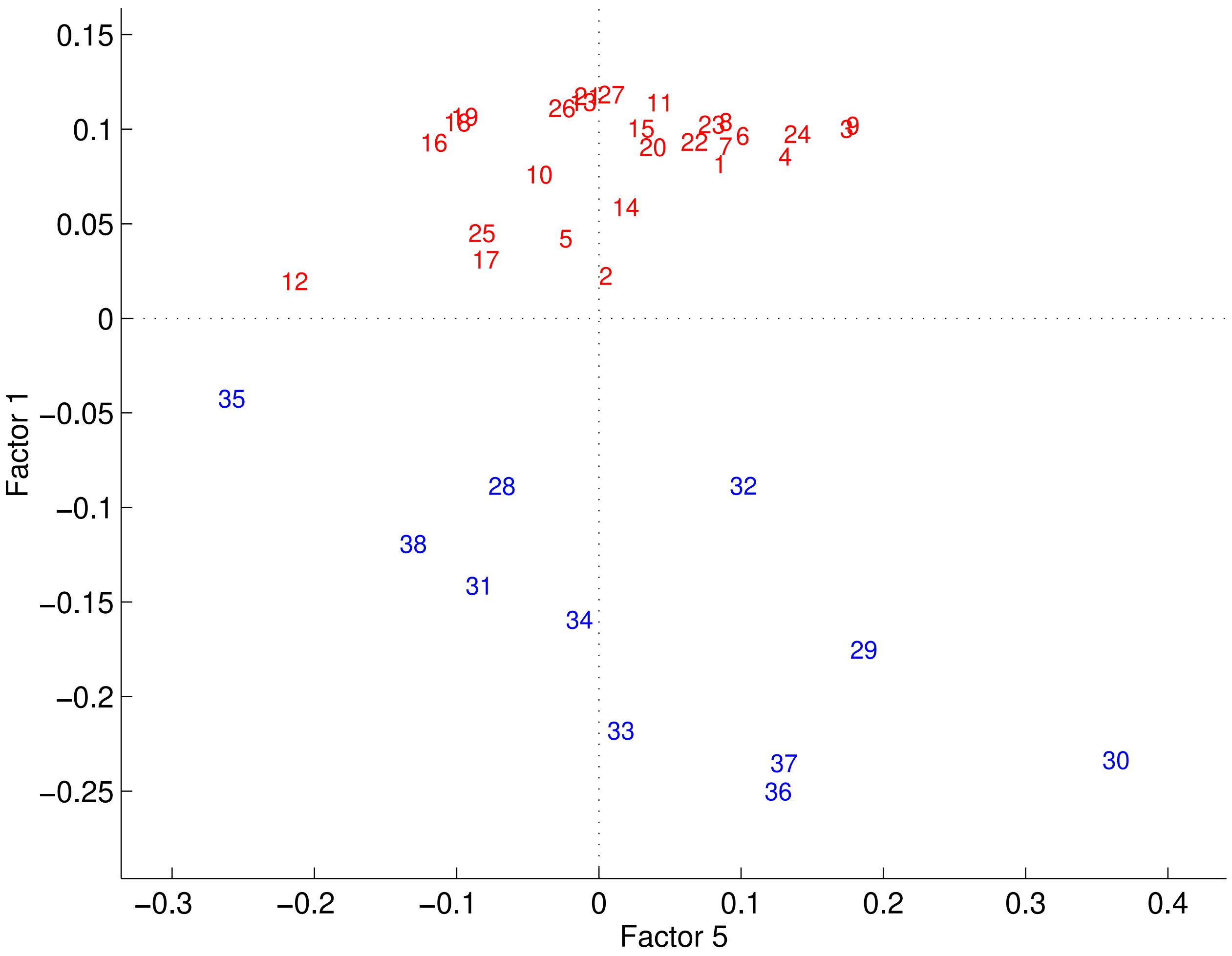
Golub *et al* Science, 1999

- 2 leukemias: ALL (1) and AML (0)
- "easily" identified on non-genetic bases
- 38 samples (27/11) on training arrays
- 34 samples (20/14) on validation arrays
- MIT (Whitehead) study:
  - data-based screen to 3,571 genes
  - some difficulty in predictive classification of 5 validation cases

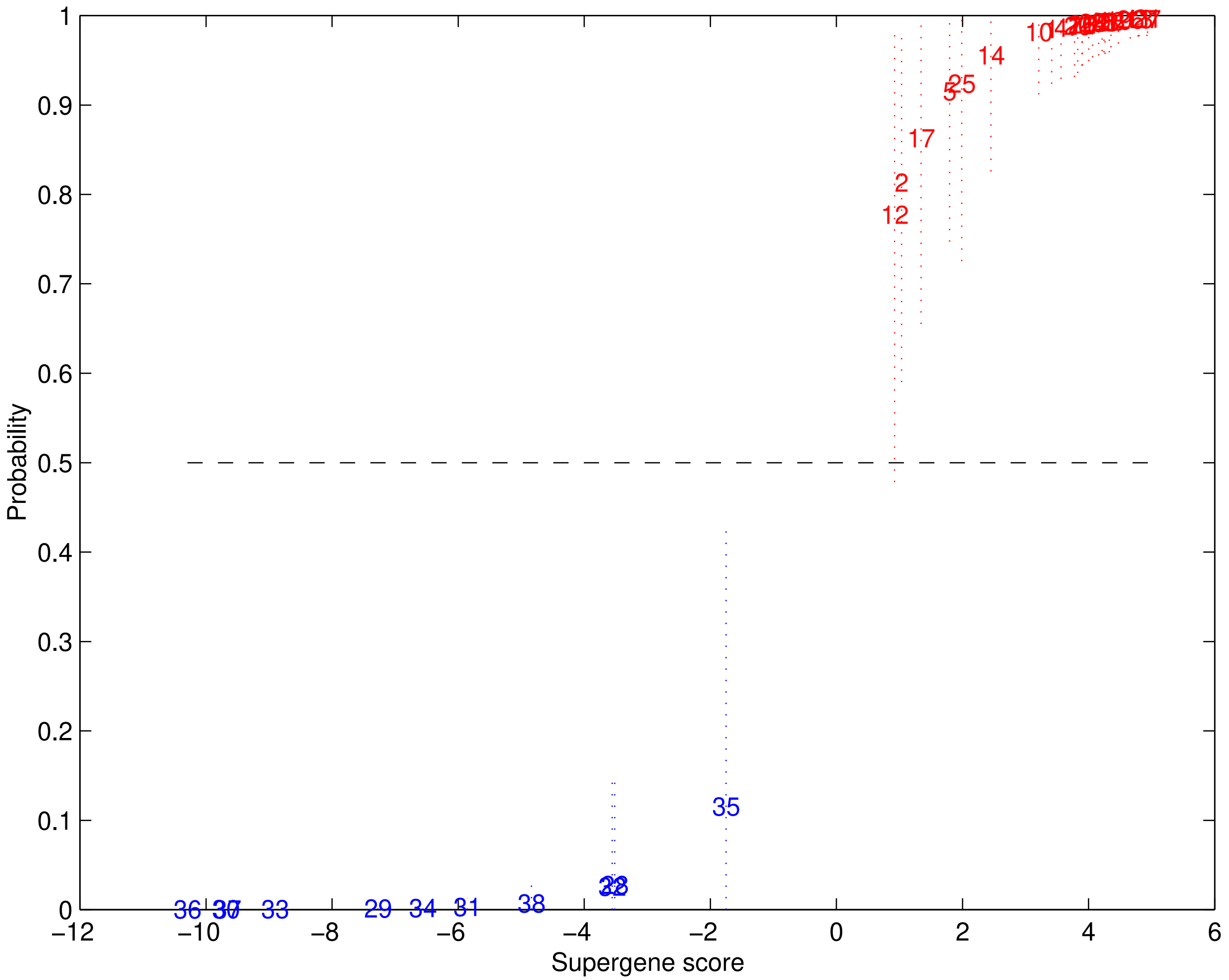
Leukemias: 2 factors in all genes



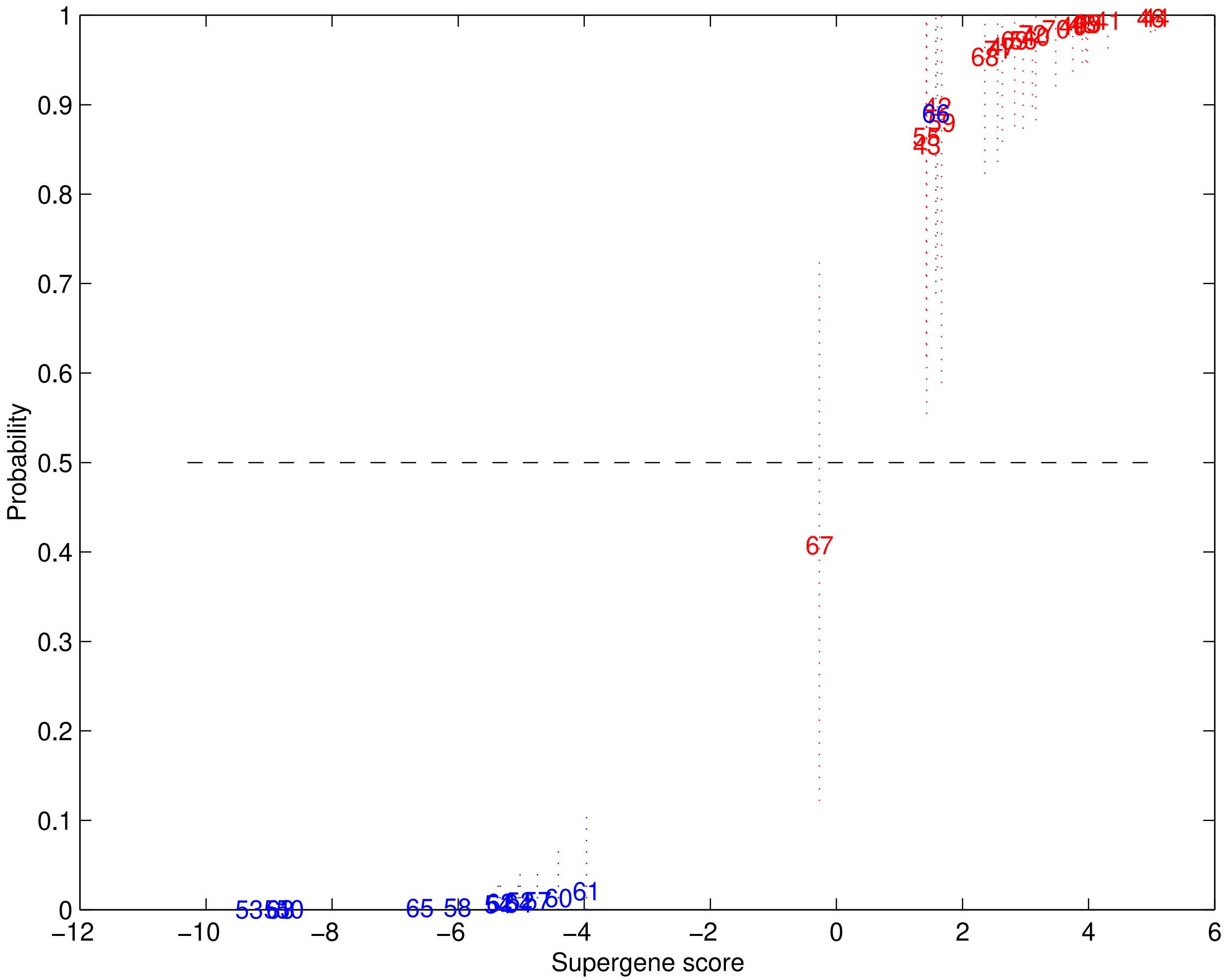
Leukemia: Two factors underlying 50 “top genes”



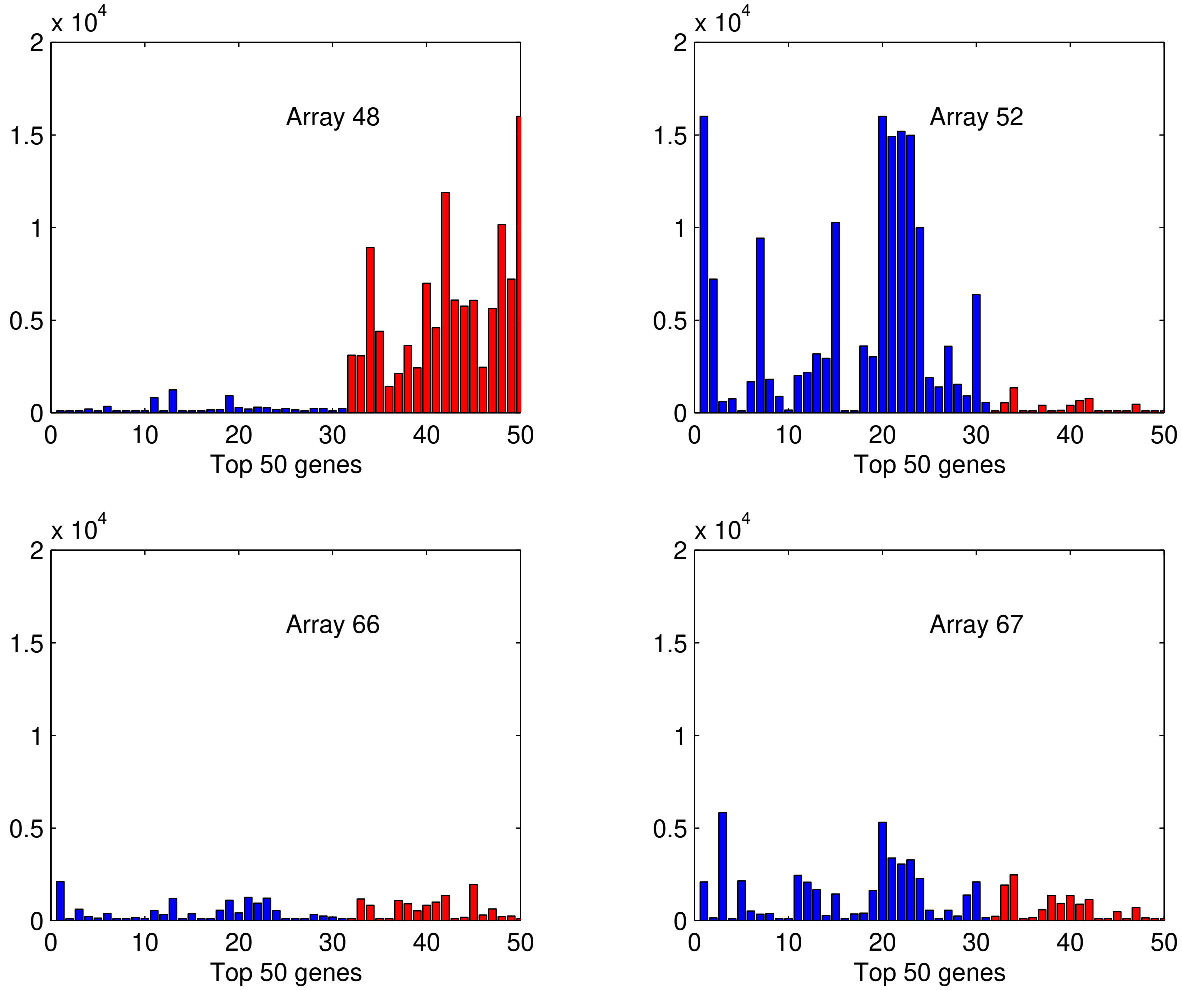
Leukemia: Fitted classifications on top 50 genes



Leukemia: Predictions for validation sample



Leukemia: Top 50 genes on four arrays



Data issues with Affymetrix arrays

- Hybridisation problems: RNA quality
- Fluorescent image scanning (registration, resolution)
- Global normalisation of expression, array to array
  - global scaling
  - non-linearities induced by varying hybridisation quality
- Local issues: scratches, patches, ...

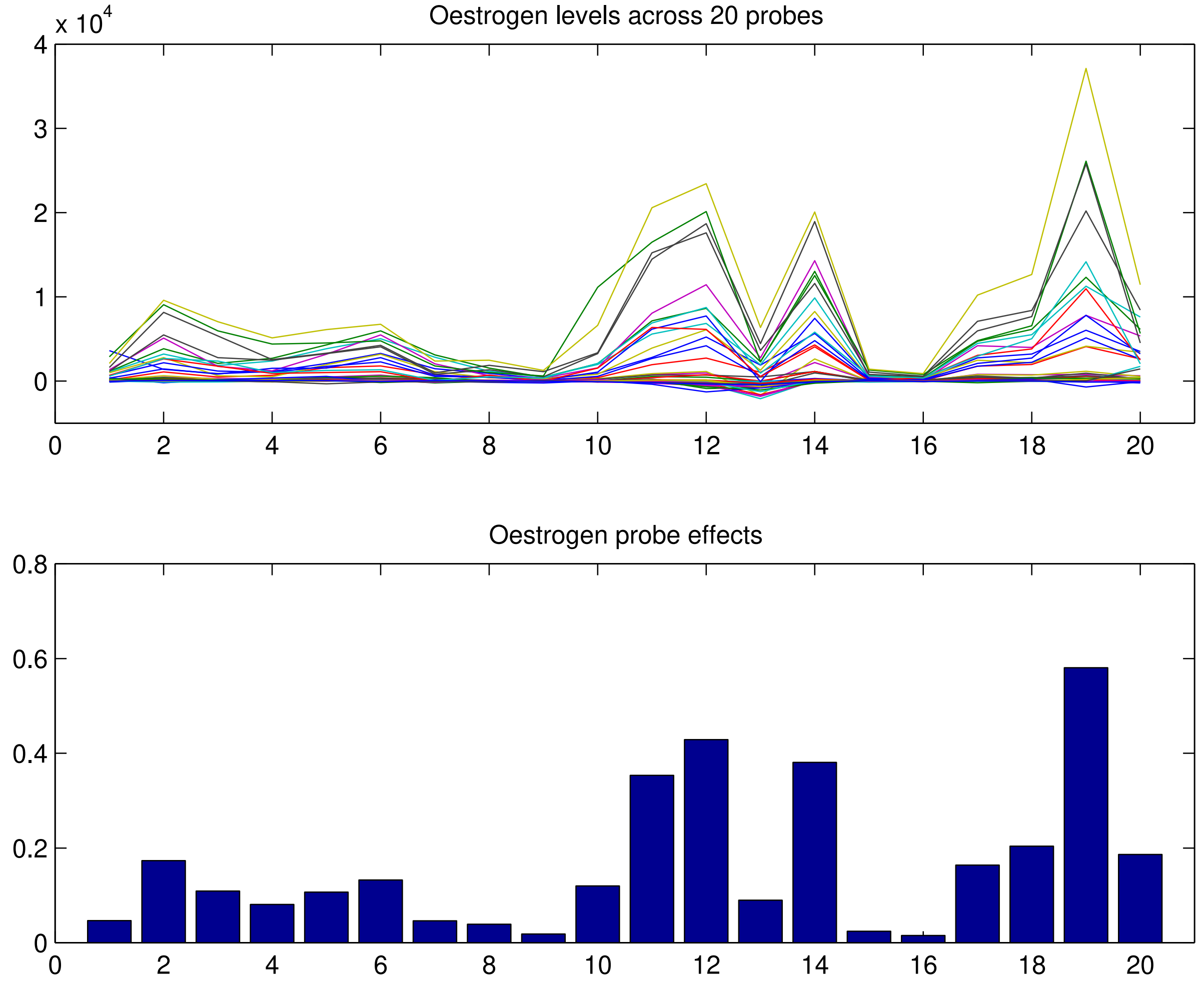
All distort expression summaries

- Pixel-level image model for background
- Bayesian image analysis: (non-negative) expression level *parameters*

More data issues

- 20 probe sequences per gene
  - "averaging" of pixel values within probe cells
  - "averaging" of probe cell averages
  - empirically based: global reliability?
- Marked variability across 20 probes for some genes
- 25mer specific hybridisation intensity
- **Alternatives:**
  - Model 25mer-specific hybridisation intensities (Li & Wong 2000)
  - Use all data: 20 measures per gene

Probe effects



Data quality and imaging issues

Image registration difficulties

- Scanning "grid" alignment problems
- Resulting probe cell summaries distorted
- Bayesian image registration methods to realign

Image background modelling

- Markov random fields at pixel level
- Aim to improve estimates of sequence-specific expression

Image registration issues

c.o.v. of probe cell expression levels – original

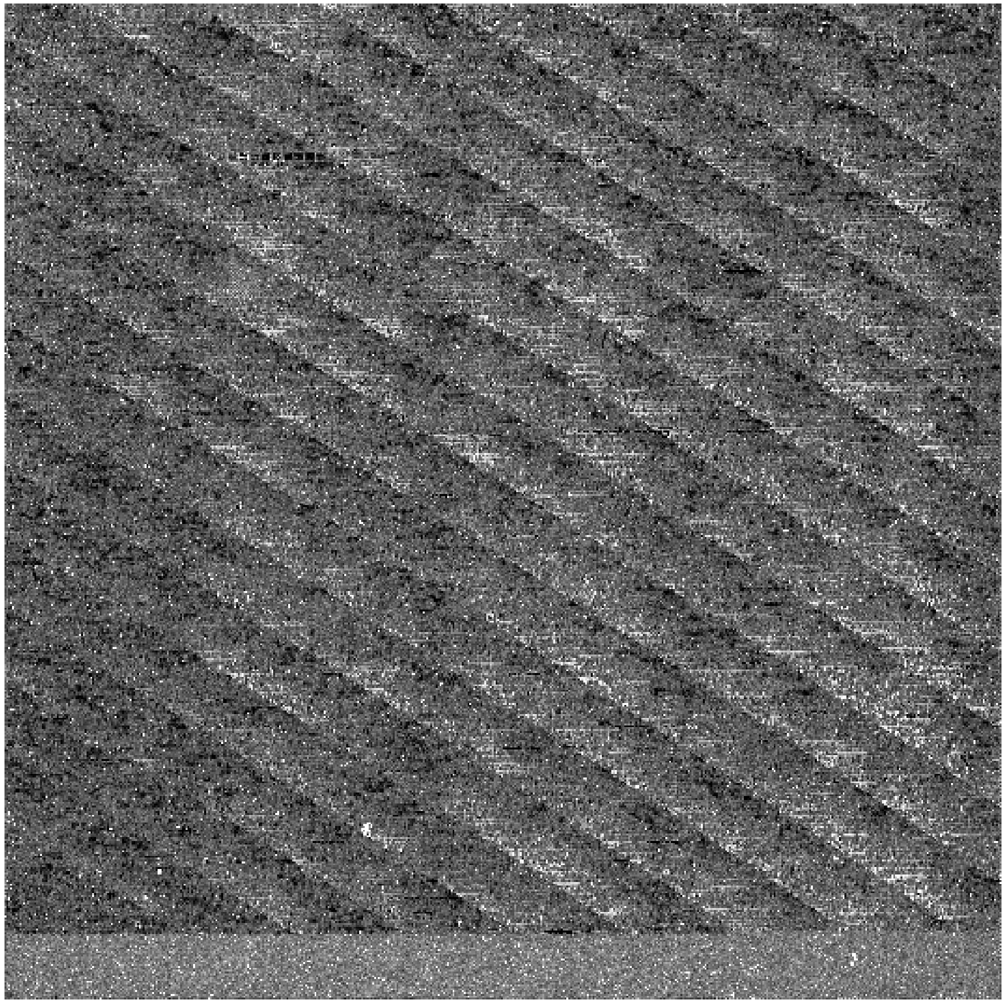
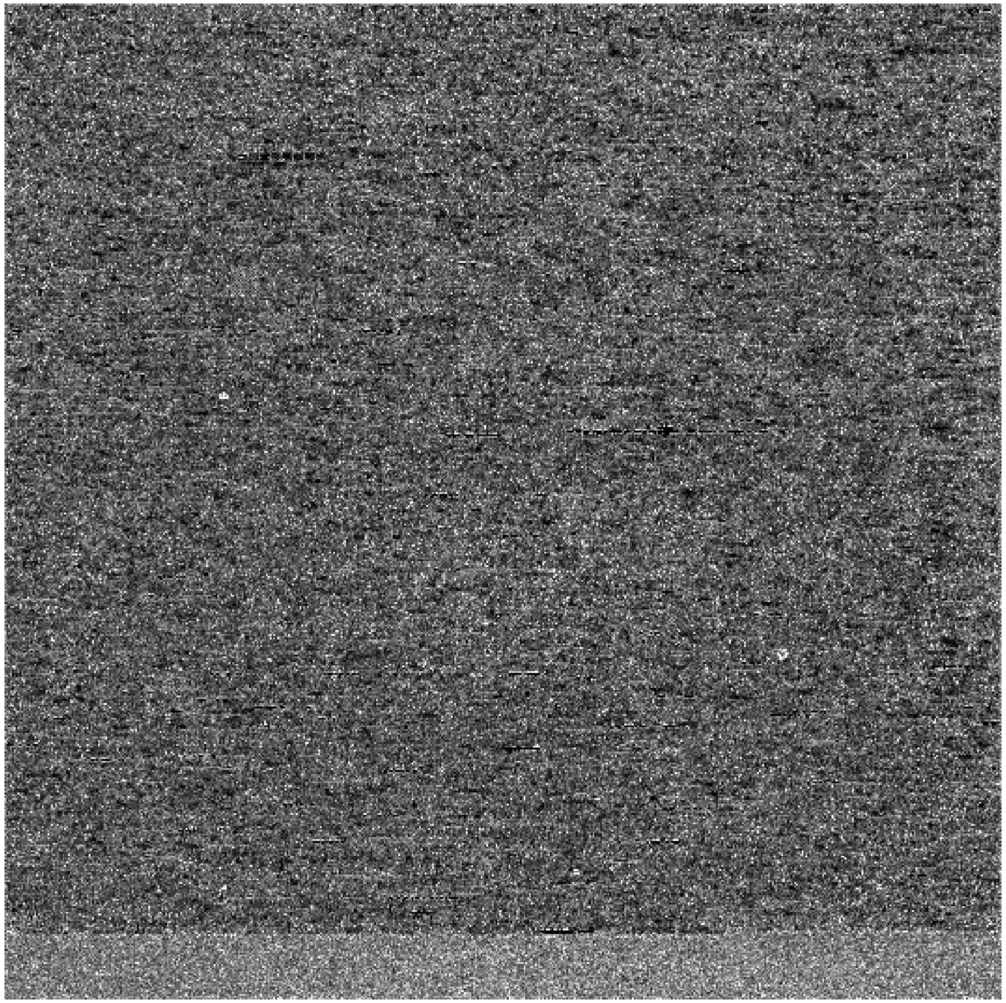


Image registration issues

c.o.v. of probe cell expression levels – aligned



Futures

Applications/extensions

- Other outcomes: e.g., genomic predictor of treatment outcome
  cancer states, remission/survival times, ...
- Exploration of relationships among genes
- Combining expression profiles with other clinical data

Statistical models

- Tumour heterogeneity issues
- Modelling progression of nodal status
- Refined factor models - to "de-noise" singular factor method
- Accounting for measurement errors in expression summaries
- Non-linear regressions