Data analysis and binary regression for predictive discrimination using DNA microarray data

Mike West
Institute of Statistics & Decision Sciences
Duke University

www.stat.duke.edu

IPAM Functional Genomics Workshop *November 2000*

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 ≈ 7000 genes (sequences) in RNA from tumour, tumour location, time point, ...
- 23 ER+, 20 ER−
- Discriminatory patterns of expression?
- Predictive classification of tumours 44, 45, ... ?
- 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: $\log_2(\max(1, \text{AbsDiff}))$

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: $n << p$
- $p = 7000\pm$ genes, $n = 43$ arrays

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

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 $\mathbf{F}$

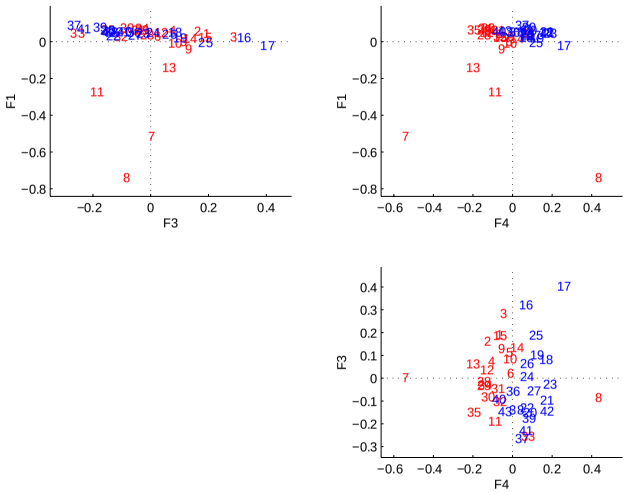
- patterns/relationships among arrays: $n << p$ factors

Supergenes=Factors: linear combinations of expression

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

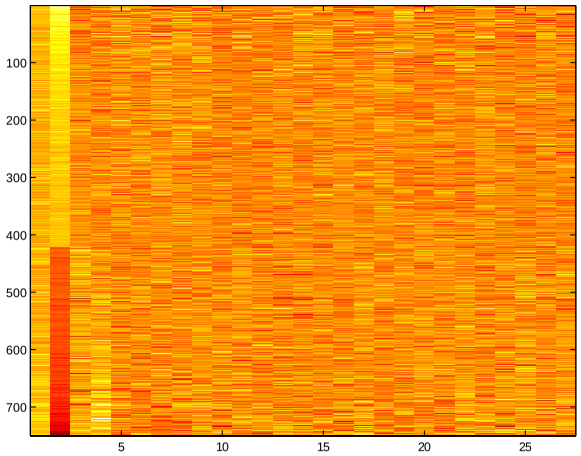
$$x_{i,j} = a_{i,1}f_{1,j} + a_{i,2}f_{2,j} + \ldots + a_{i,n}f_{n,j}$$

Arrays on 3 supergene factors: Coloured for ER+/ER-



Factor weight vectors 750 genes

Weight vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots,$



Binary regression modelling

- Microarray j, expression profile $\mathbf{x}_j$
- Binary classification: 1 (ER+) or 0 (ER−)
- Probability array j is ER+ is $p(\mathbf{x}_j)$
- Standard probit model: $p(\mathbf{x}_j) = \Phi(\mathbf{x}_j'\boldsymbol{\beta})$
- 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

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

- n parameters, sample size n
- Ignore "stable" factors
- Use of stochastic regularisation: priors on $\boldsymbol{\theta}$
  - elements θ_j independent (orthogonality)
  - $\theta_j \sim N(0, \tau_j^2)$ with prior on τ_j
  - neutral: implied priors for classification probability $p(\mathbf{x}_j)$
- 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

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

- $\boldsymbol{\lambda}_i \sim N(\mathbf{0}, \boldsymbol{I})$ 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: $\boldsymbol{\Psi} \to \mathbf{0}$ 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) $\boldsymbol{\beta} = \mathbf{A}\mathbf{D}^{-1}\boldsymbol{\theta}$
- Bayesian/model justification of generalised inverse to $\boldsymbol{\theta} = \mathbf{D}\mathbf{A}'\boldsymbol{\beta}$

Cross-validation (honest) prediction

Critical **predictive** assessment of discriminatory performance

- One-at-a-time analysis
- Take out microarray j
  - Fit model
  - Predict status of tumour j
- Repeat for all arrays j

Kernel regression structure

Marginalising over $\boldsymbol{\beta}$ implies

$$\mathbf{y} \sim N(\mathbf{0}, \mathbf{K})$$

with *kernel covariance matrix*

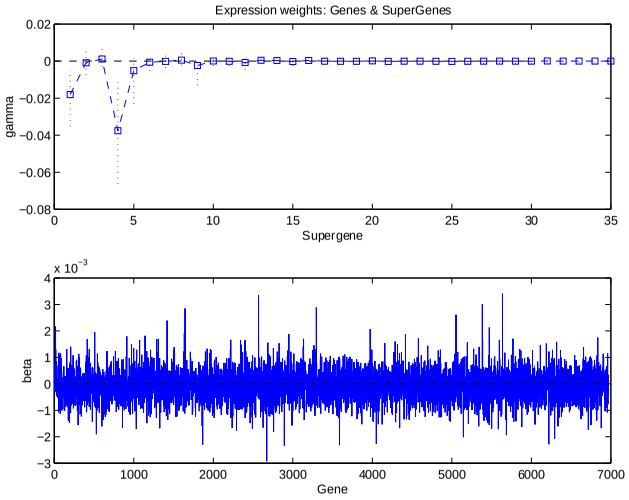
$$\mathbf{K} = \mathbf{F}'\mathbf{T}\mathbf{F} + \mathbf{I}$$

with

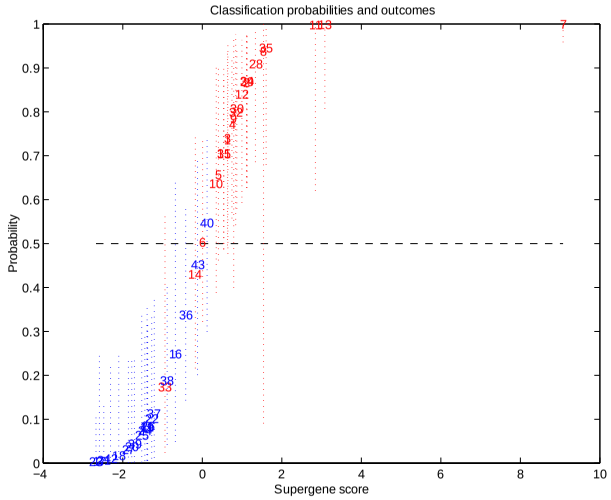
$$\mathbf{T} = \mathrm{diag}(\tau_1^2, \ldots, \tau_n^2)$$

- correlations between arrays
- effective dependence structure with respect to classification
- key role of $\mathbf{T}$
- effective *non-linear classifier*

ER status: Estimated regression coefficients

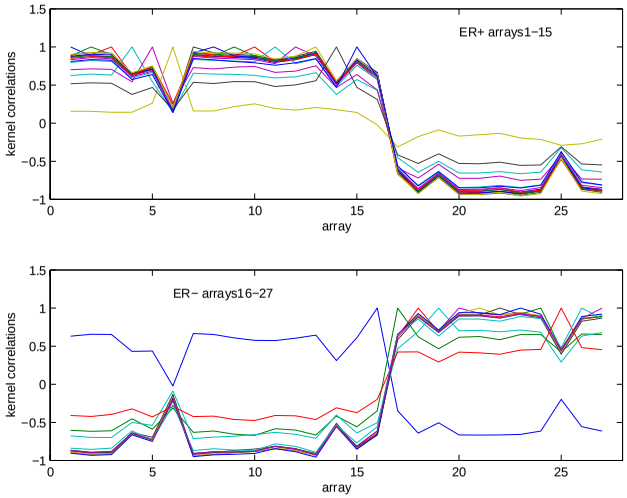


Fitted classification

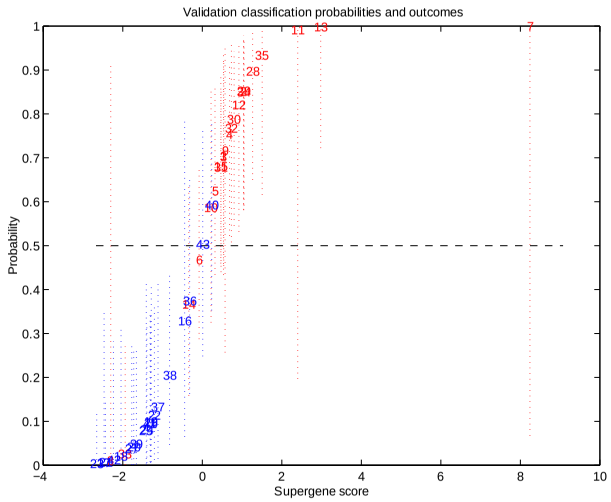


Estimated kernel correlation structure

First run of 27 tumours/arrays only



Cross-validatory predictions



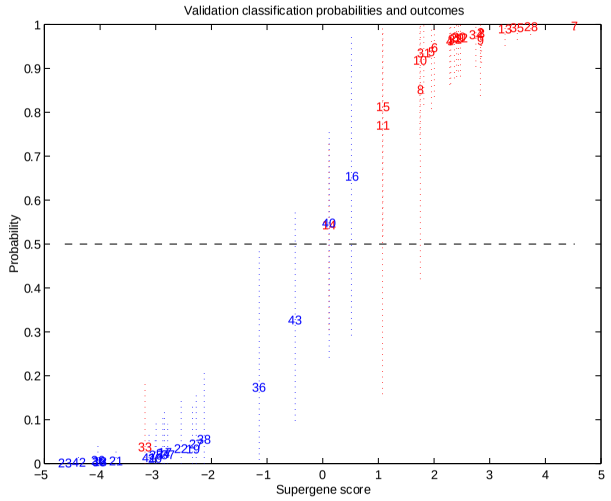
Gene screening

- 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**
  - Interesting cases 33 (ER−), 16, 40 (ER+)

- **Tumour 33**: Classified ER+ (non-Duke diagnosis)
- Reclassify as ER+ and refit model: "Perfect" classification

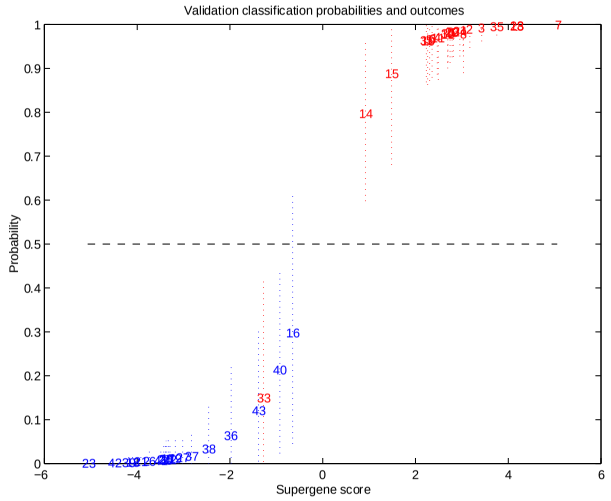
Cross-validation predictions: Top 100

One-at-a-time analysis

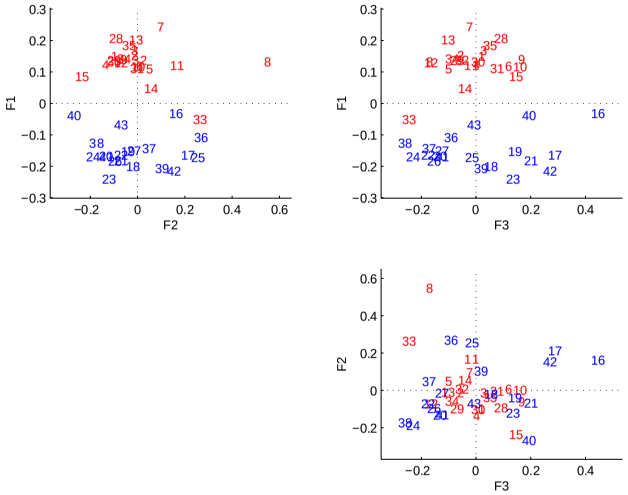


Cross-validation predictions

Overall top 100



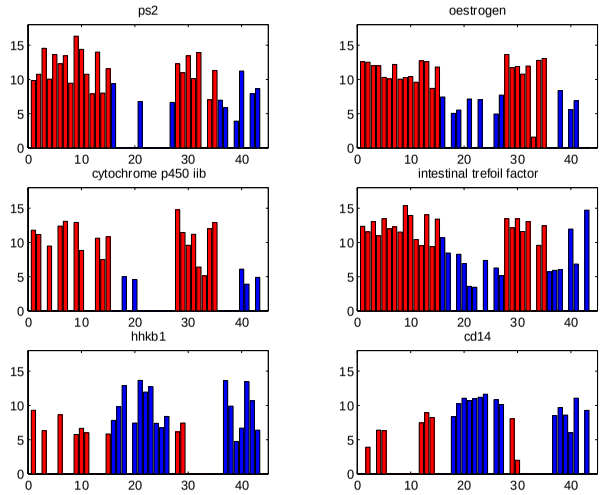
Arrays on pairs of 3 factors: Top 100



Some "top" genes: "up" favours ER+

- ps2 protein gene
- mrna for oestrogen receptor
- cytochrome p450 iib (hiib1) mrna
- intestinal trefoil factor mrna
- hepatoma mrna for serine protease hepsin
- insulin like growth factor binding protein [placenta]
- p37nb mrna
- c-myb gene
- ccaat displacement protein
- clone 23948 mrna sequence
- nat1 gene for arylamine n-acetyltransferase
- ...
- breast cancer, oestrogen regulated liv-1 protein mrna

Expression levels of some top genes

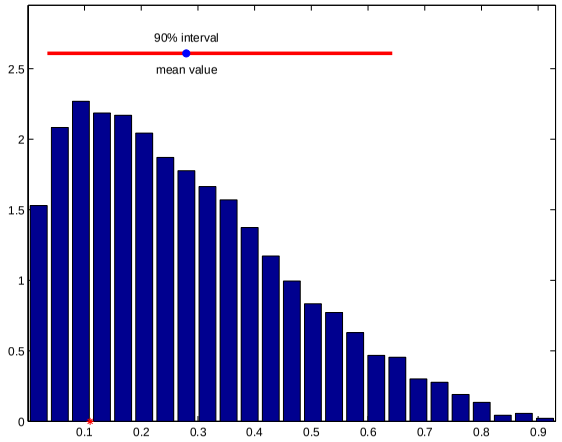


Tumours 16, 40

- Similar patterns: ER+ or ER−?
- High uncertainty about $Pr(ER+)$
- Oestrogen gene marginally "down" - Ps2 and Liv-1 higher
  Both regulated by oestrogen receptor
- Other "up for ER+" genes high on arrays 16, 40
- Mixed story in data on arrays 16, 40
- High classification uncertainty results
  - Other regulators of Ps2, Liv-1 ... ?
  - ER status determination ... ?
  - Evolving from − to +?

Classification and uncertainty

Classification probability for tumour 16



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

Breast cancer nodal status

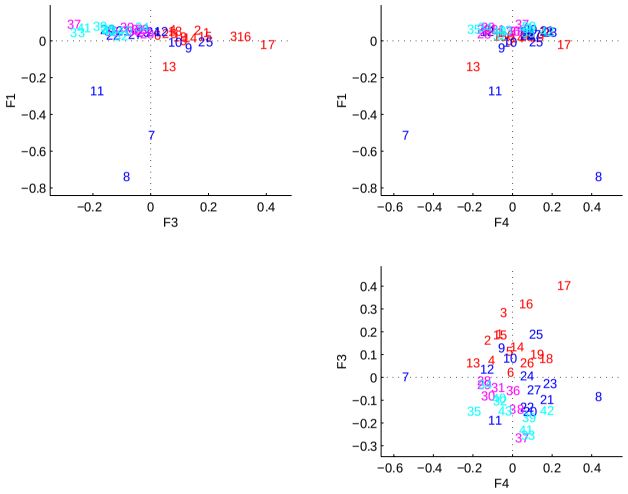
- Same breast cancer arrays, classified by axilliary lymph nodal status:

  primary, lymph node-negative breast cancer
  versus
  primary, lymph node-positive

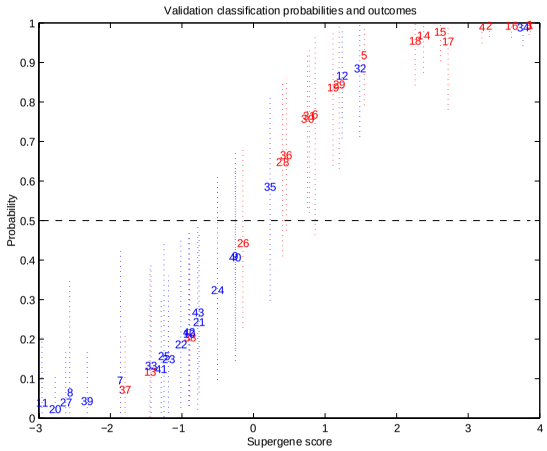
- Expression: expected to be highly heterogeneous
- Data confirms this: Analysis of all 7000+ genes
  - no clear discrimination expected
  - none found
- Clearer picture based on "Top 100" - Similar story to ER
- Data issues: Consistency of samples

Arrays on 3 supergene factors

27 initial arrays, 16 later arrays



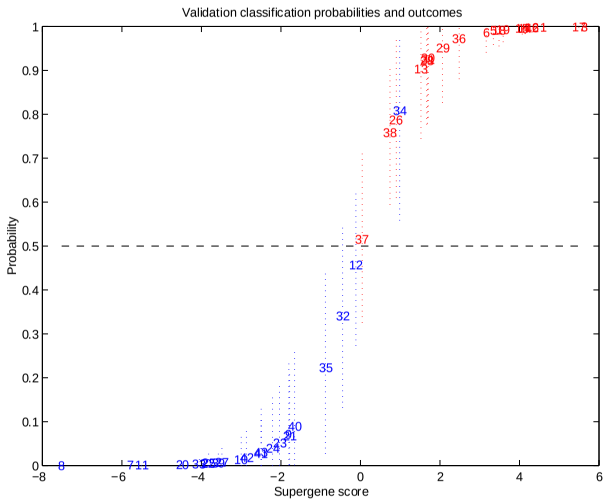
Cross-validation predictions: 1-at-a-time/Top 100



Case 37: 1+/37: most "extreme"
Case 34: 0+/13

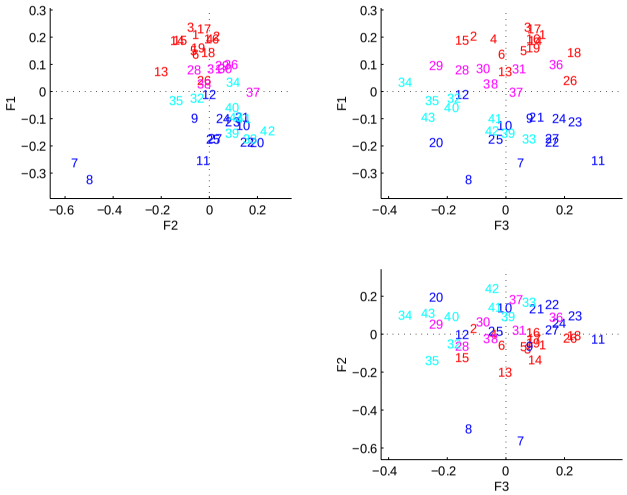
Cross-validation predictions: Top 100

Overall top 100



Arrays on 3 supergene factors: Top 100

27 initial arrays, 16 later arrays



Some "top" genes: "up" favours Node–

- alk-4 mrna
- bloom syndrome protein (blm) mrna
- wilm tumour-related protein
- mrna for actin-related protein
- retinoid x receptor beta (rxr-beta)
- fkbp-rapamycin associated protein (frap)
- ribosomal protein s4
- histone h1.1
- receptor tyrosine kinase ligand lerk-7 precursor (eplg7) mrna
- mrna for kiaa0063 gene
- mrna for glycerol kinase
- ...

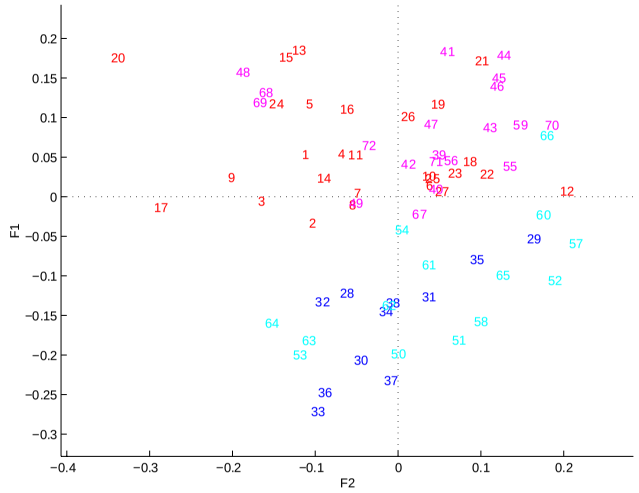
MIT ALL/AML leukemia study

Whitehead Institute, Lander group

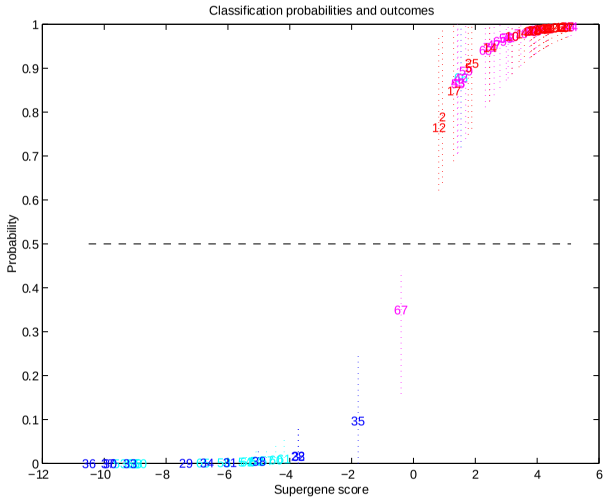
Golub *et al* Science, 1999

- 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:
  - some difficulty in predictive classification of 5 validation cases

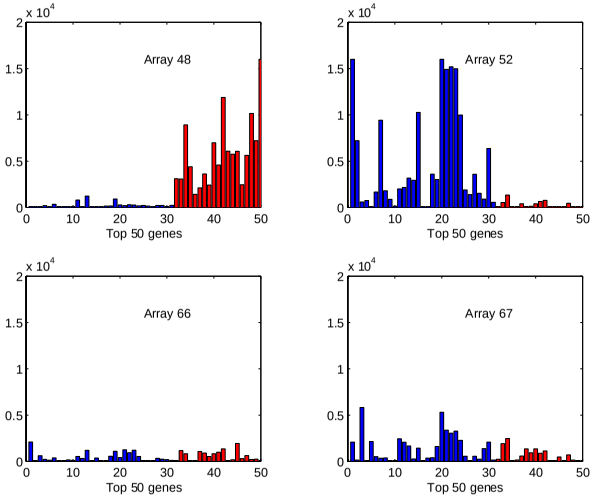
Leukemias: 2 factors



Validation predictions on top 50



Top 50 genes on four leukemia 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

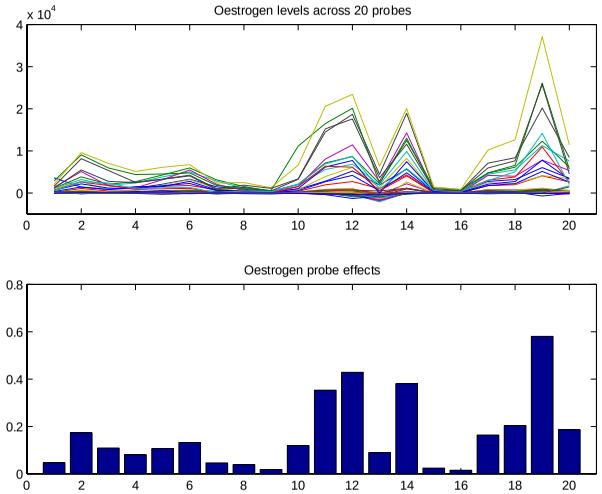
- Pixel-level image model for background
- Bayesian image analysis: (non-negative) expression level *parameters*

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



Futures

Applications/extensions

- Other outcomes: e.g., genomic predictor of treatment outcome
- Multiple outcomes: e.g., cancer stages/states
- Measured outcomes: e.g., *time* to remission
- Exploration of relationships among genes
- Combining expression profiles with other clinical data

Statistical models

- Refine "empirical" singular factor method
- **Latent supergene factors** - to "de-noise" singular factor method
- Accounting for measurement errors in expression summaries
- Non-linear regressions