STA 278/BGT 208
GENE EXPRESSION ANALYSIS

Statistical Models, Methods & Computation

Mike West

Institute of Statistics & Decision Sciences

www.isds.duke.edu

Computational & Applied Genomics Program

www.cagp.duke.edu

Statistics in Genomics and Biological Profiling

- y - clinical or physiological state, outcome
  - few dimensions
- x - molecular, clinical, biological data
  - DNA microarrays - Gene expression: few-40,000 genes
  - Other molecular data, clinical data, ...
- Breast, ovarian cancer (risk groups, recurrence time outcomes)
- Cardiovascular disease (disease states)
- Biological response to environmental exposures
- Cell cycle patterns: common regulators of gene subsets
- Interrelationships: gene pathways and transcriptional control
- Gene discovery and prioritisation

One Example Motivating Context

Breast cancer discrimination & prediction: Binary/2-group problems

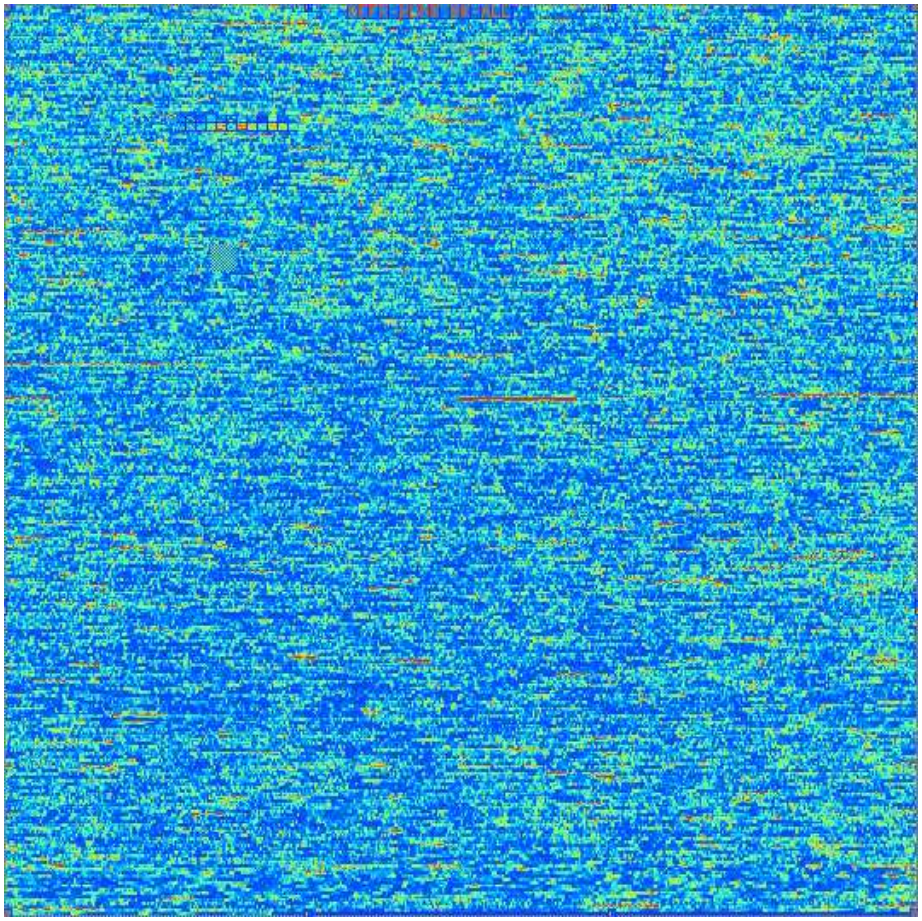
- e.g., ER+ versus ER−
- e.g., lymph node + versus lymph node −
- DNA microarray data: expression levels of $5-40,000$ genes (sequences) in RNA from tumour, tumour location, time point, ...
- n_1 cases ER+ and n_0 cases ER− (by design or observation)
- Discriminatory patterns of expression?
- Predictive classification of *next* patient/tumour?
- Which genes are implicated? Biology?
- Which tumours depart from general patterns? How?
- ... etc

Expression array data

Microarray data: Affymetrix arrays

- $7 - 40,000$ genes (sequences)
- Each represented by 15-20 ($\times 2$) short oligonucleotide sequences
- Data issues:
  - management, manipulation, multiple formats
  - imaging: extraction, summary data production
  - data quality: experimental and processing errors
  - estimates of expression level by gene: multiple methods
- **Statistical modelling, analysis and inference**

One Tumour Sample



Older Affymetrix array: 7,000 genes *Current: 40,000 genes*

Key Statistical Themes

Modelling high-dimensional distributions: $p(x)$, $p(y|x)$

- Understanding high-dimensional structure: x
  - management and exploration challenges
  - exploratory analysis – subsets, clusters, visualization
  - parametric models
  - complex dependencies: collinearities
- Relating to prediction of outcomes: y
  - Regression and prediction models
  - Scale of parameters, number of observations
  - Variable selection: which xs? Multicollinearities, redundancies, ...
- Major role for statistical computation & visualisation

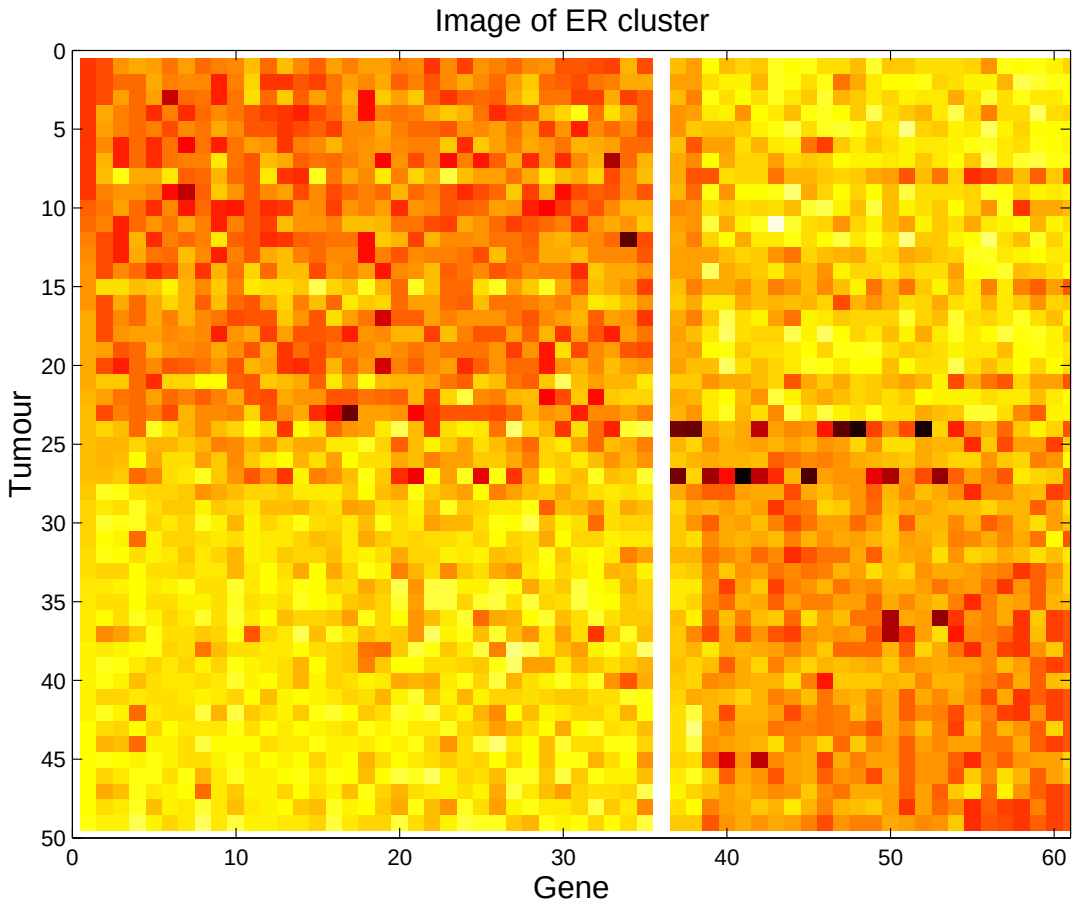
Organising & Exploring Data Subsets: Clustering

- For gene & gene-gene relationship discovery
- To identify aggregate patterns, or "factors," within groups

Cluster analysis:

- grouping genes with related patterns over samples
- grouping samples with related patterns across gene subsets
- both ..
- Look at results ... interpretation?
- Selection of gene subsets for clustering?
- Clustering methods: $k-$means, hierarchical, ...
- Software

Simple Cluster Display



Genes selected via binary regression model analysis

Statistical Factor Ideas

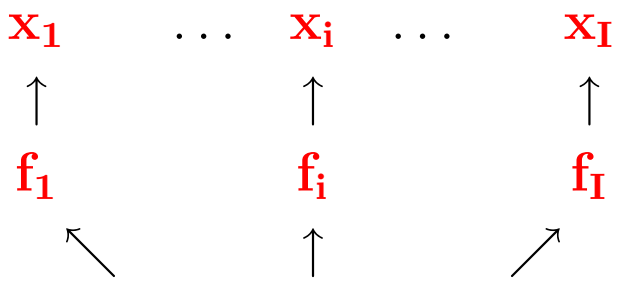
Direct (empirical) data decompositions

- PCA, SVD
- (Several) other decomposition methods

Factor Models:

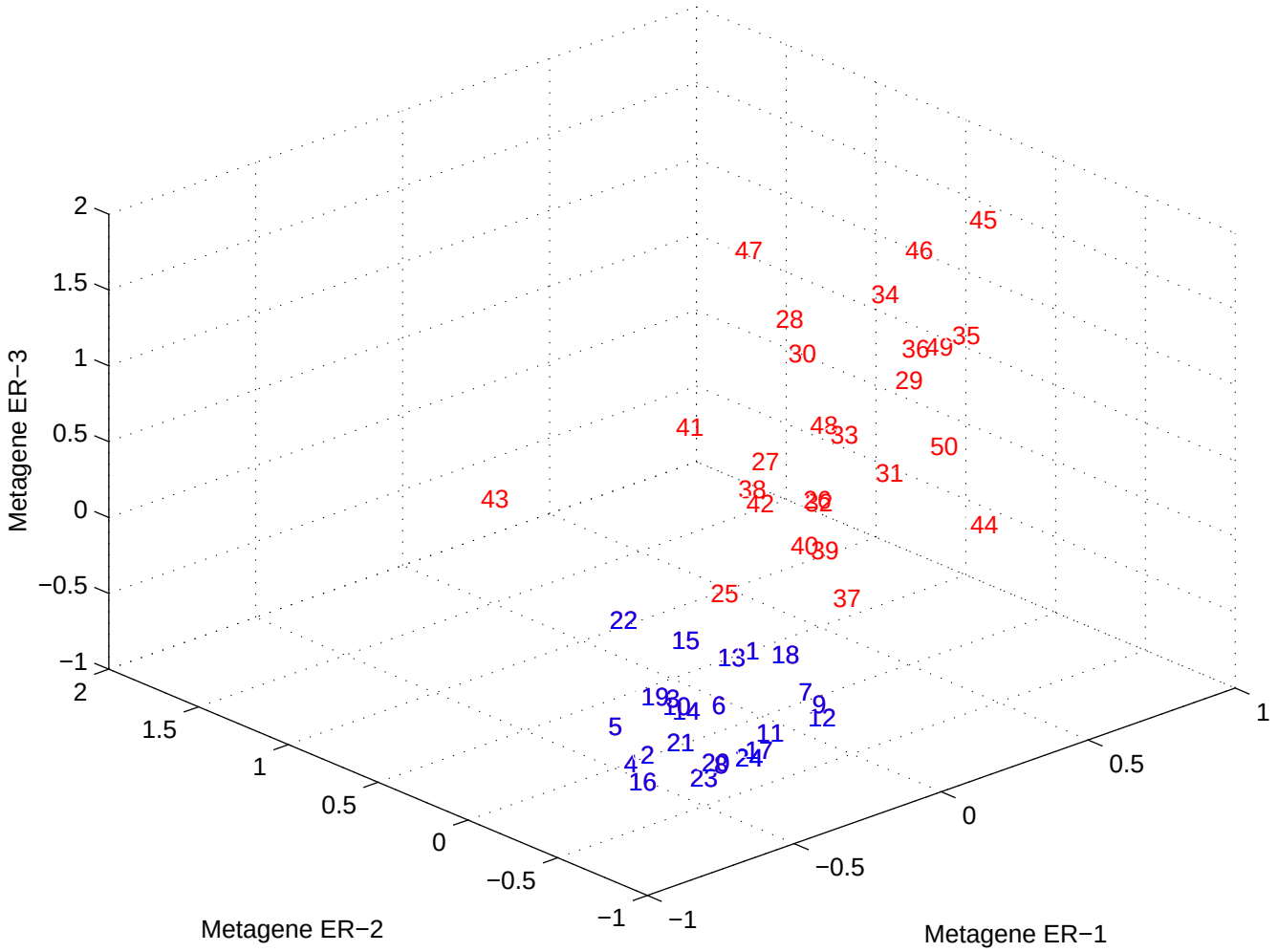
$$\mathbf{p(x)} = \int \mathbf{p(x|f)p(f)df} \qquad \text{e.g., } \mathbf{x = Af} + \epsilon$$

- $\dim(x)$=0,000s, $\dim(f)$= 0s-00s



Common patterns across genes induced by "low dimensional" factors

Binary Predictions: Simple ER Example



Factor Models

- Dimension reduction – factors are latent, but "real" structure
- Prediction and regression – factors may be predictors of y
- Links to cluster analysis: clustering by "common pattern" = "factor"
- Statistical & computational issues

Genomic applications:

- Underlying biological/gene pathways in genomic studies
- Finding models using factors to predict clinical states
- *"Metagene"* factors: characterise state

Models and Methods for Regression and Prediction

eg., Linear regression: $y = x'\beta + \epsilon$

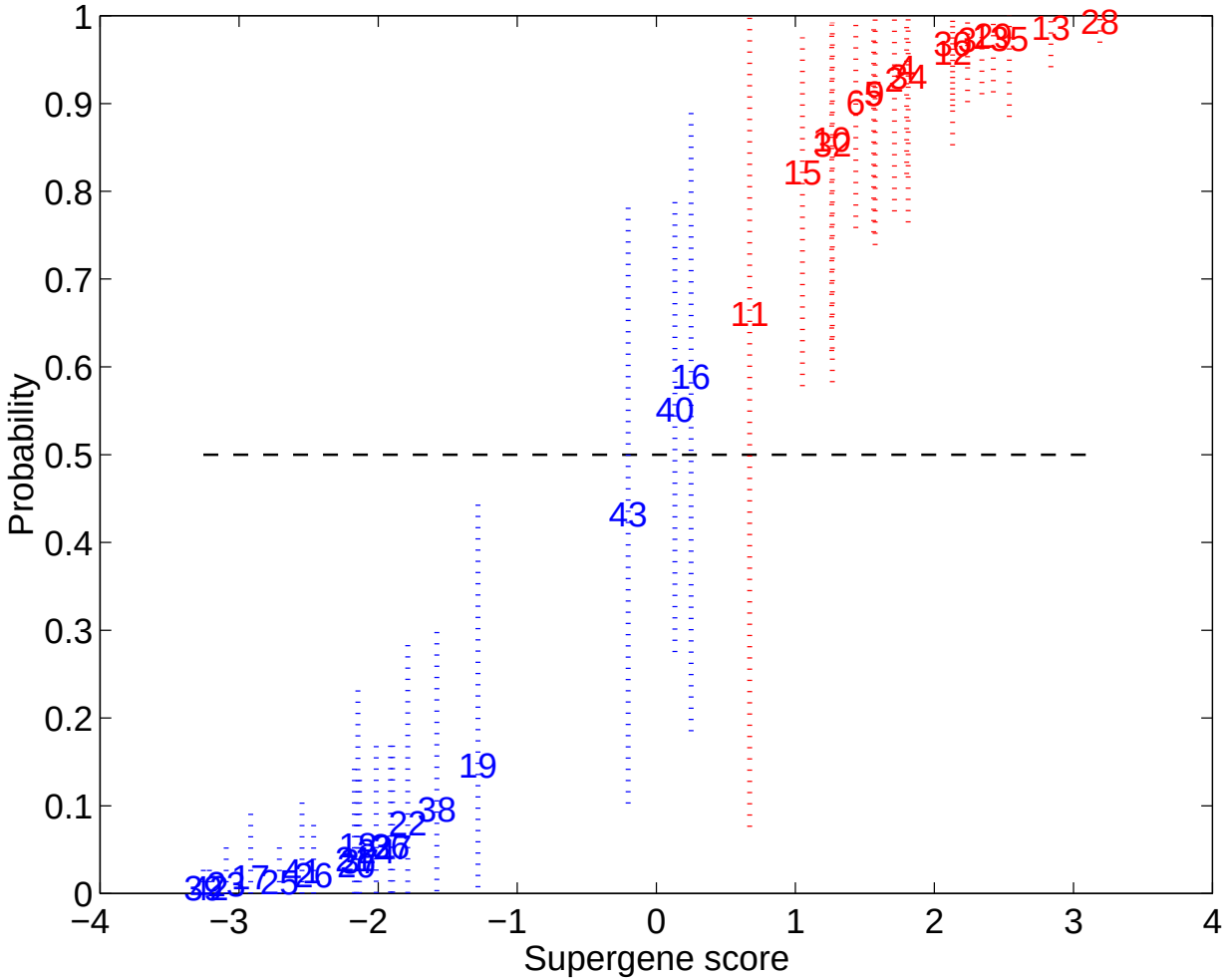
eg., Binary regression: Probability of ER+ is $\pi(x)$

binary regression models: $\pi(x) = \Phi(x'\beta)$

General Statistical Issues:

- Estimate regression coefficients, uncertainty: β
- Select genes (variables) - subset of x
- Factors as predictors? Other non-genomic variables?
- Multicollinearities!
- $p >> n$
- Deriving predictions

Binary Predictions: Simple ER Example



Likelihood & Bayesian Inference

Likelihood Inference: Maximise and summarise

$$L(\beta) = \prod_{i=1}^{n} p(y_i|\beta, x_i)$$

Requires identifiable model: $p < n$

Bayesian Inference: Explore, sample and summarise posterior density

$$p(\beta) \propto p(\beta) \prod_{i=1}^{n} p(y_i|\beta, x_i)$$

for chosen priors $p(\beta)$

Regression Models: Inference & Computation

Model Fitting, Search & Computation

- Bayesian analysis: Regularisation for prediction
- Search and optimisation methods to find MLEs, posterior modes
- Simulation-based computations unavoidable in non-linear models
- Large-scale model search: combinatorial search

Statistical Computation

Computation: Iterative methods for a given model

- Newton-Raphson search, gradient algorithms
- EM (Expectation-Maximisation) algorithm
- Simulation: e.g., Markov Chain Monte Carlo (MCMC)

Model/variable selection:

- Which genes go in x?
- Sequential search – forward/backward variable selection
- Stochastic search
- Combinatorics: k genes (or factors) from p

e.g., $p = 100, k = 4$: about 4m models. Tiny example.

Non-linear Regression

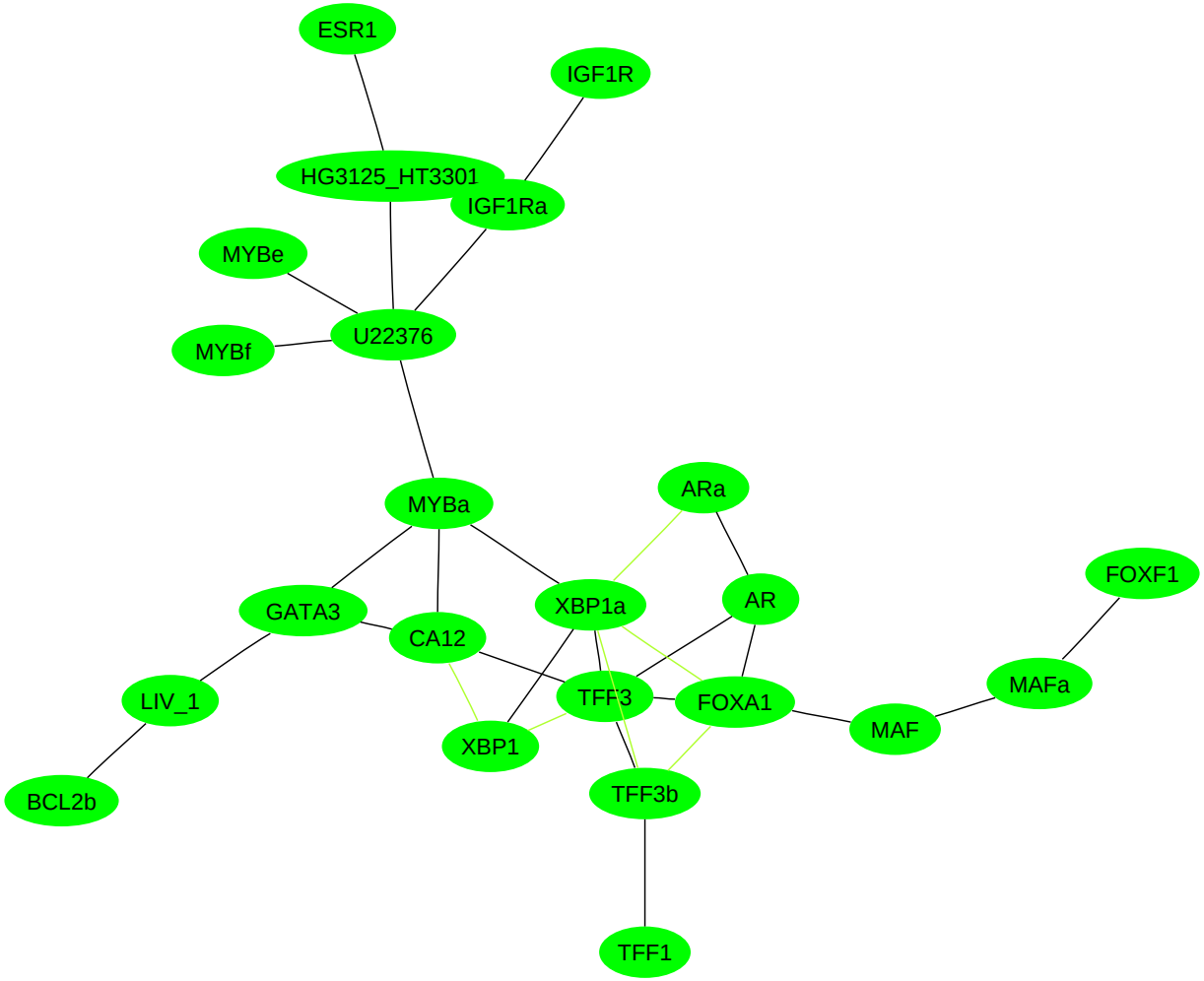
- Non-linear regression models for binary, categorical outcomes
- Survival data modelling: recurrence of cancer
  - Non-Gaussian
  - Censored samples
  - Non-parametric data exploration, parametric models
- Non-linear association (correlation is linear)
- Recursive partitioning methods: Statistical Tree Models

Statistical Exploration of Associations: Graphical Models

Bayesian Networks: Graphical Models

- High-dimensional distribution of x
- Graph representation:
  - Nodes are x_i
  - Edges represent "dependencies"
  - Understanding dependencies, defining "cliques"
  - Relationships to factor models
- Role of graph theory; role of computation
- Links to biological networks in genomic applications
- Links and interfaces to text-based information systems

An ER SubGraph of a 7,000 Node graph



Class Website

www.isds.duke.edu/courses/Spring04/sta278

- Outline, overview and topics
- Statistics and biology texts, other support materials, notes
- Computing, software and data sets

Course Goals, Projects & Research

Class Computing Orientation

COMPUTING: ISDS Unix system: isds.duke.edu

- Student accounts: NetID and setup
- Matlab, R, C++, ...
- ssh remote login

COMPUTING: Other computers (Linux/FreeBSD, Windows, Mac)

- Matlab (OIT site license; Mathworks; Computer Store); toolboxes
- Other software (cluster, graphics, graph drawing, ...)
- R download sites; Bioconductor (later)
- Matlab and R tools – google