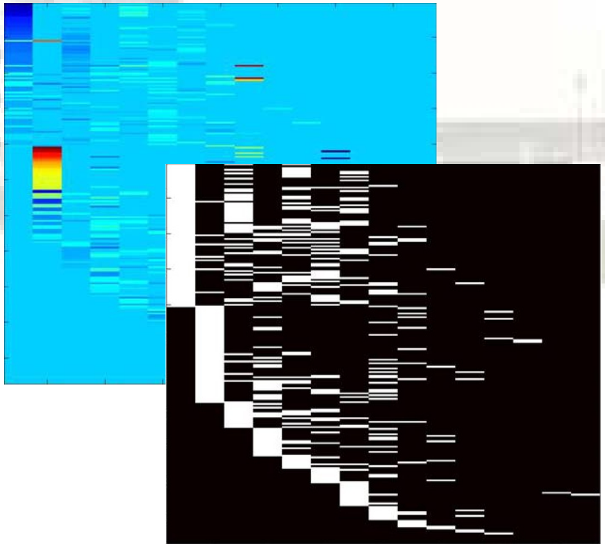
Aspects of Statistical Modelling & Data Analysis in Gene Expression Genomics



Mike West
Duke University

Flightplan

#1 **Genomics, Microarrays, Data:**
Big picture

#2 **Bayesics - Regression and Shrinkage:**
Gene expression as predictors

#3 **Patterns and Factors:**
Prediction via pattern profiling

#4 **Sparse Modelling:**
Regression subset-structure uncertainty

#5 **Sparse Models and Profiling:**
Gene expression as response: Designed experiments

#6 **Sparse Models and Profiling:**
Gene expression as response: Latent factor models

Flightplan

#1 Genomics, Microarrays, Data:
Big picture

#2 Bayesics - Regression and Shrinkage:
Gene expression as predictors

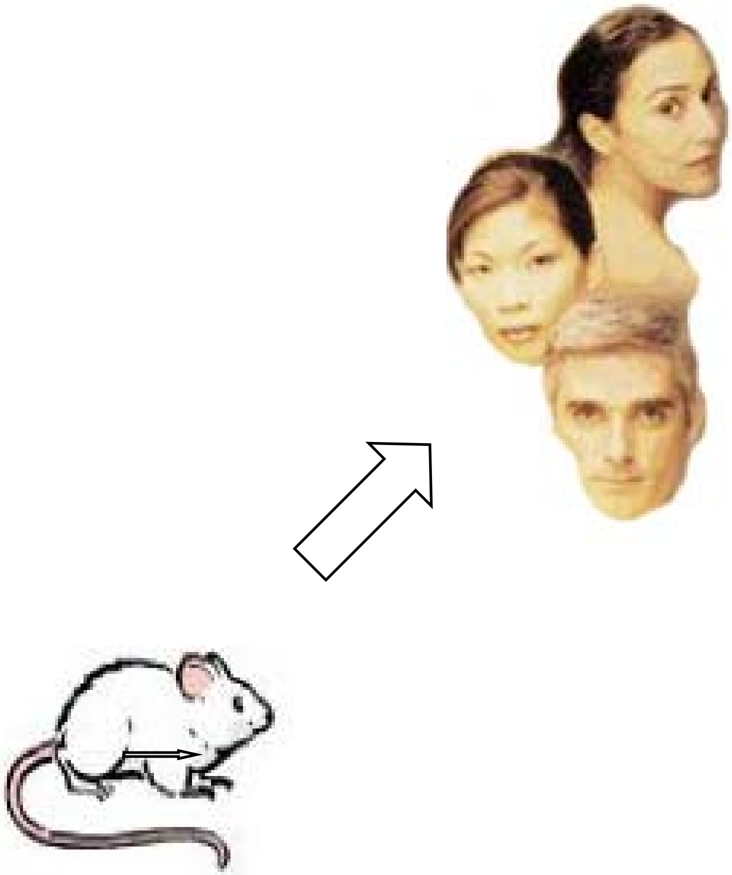
#3 Patterns and Factors:
Prediction via pattern profiling

#4 Sparse Modelling:
Regression subset-structure uncertainty

#5 **Sparse Models and Profiling:**
Gene expression as response: Designed experiments

#6 Sparse Models and Profiling:
Gene expression as response: Latent factor models

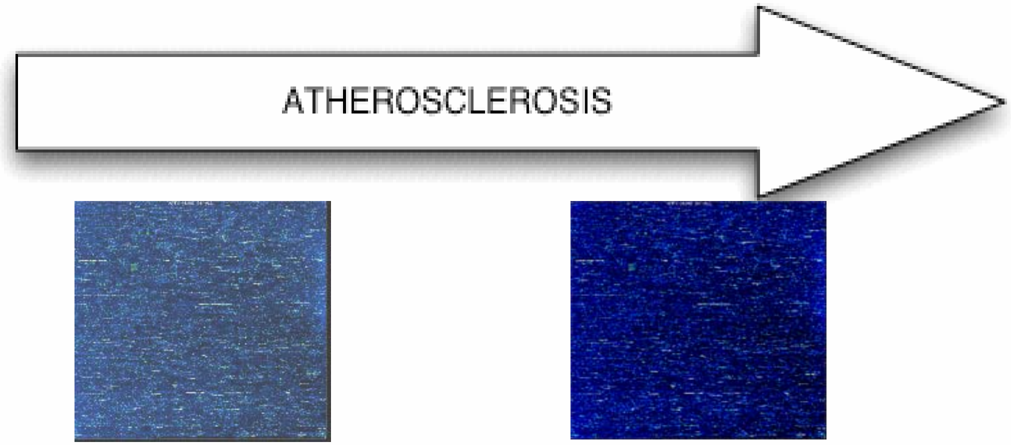
Animal Model to Human: Predictive Pattern Translation



Gene expression profiles:

Signatures of states

Cardiovascular Genomics: Atherosclerosis



Apo E -/-, 6 wk Chow Diet

PRECLINICAL DISEASE

Apo E -/-, 12 wk Western Diet

EARLY/INTERMEDIATE DISEASE

Age, Diet, Gender, WT/ApoE

2^4x5+ balanced factorial design

Human aorta gene expression:

Gene/pathway identification

Disease state prediction

Mice models:

Disease state characterisation

Mouse ➡ Human mapping

Large-scale multifactoral design

Gene expression (aorta) response

Action is interactions

Multivariate ANOVA

12,500 genes in parallel

$X \;=\; \beta$	WT, 6wk, chow, fem (baseline)
$+\ \mu$	male
$+\ \delta$	fat diet
$+\ \alpha$	age=12wk/old
$+\ \gamma$	ApoE genotype
$+\ \mu\delta$	fat diet & male
$+\ \mu\alpha$	12wk/old & male
$+\ \mu\gamma$	ApoE & male
$+\ \delta\alpha,\ \delta\gamma,\ \alpha\gamma$	
$+\ \mu\delta\alpha,\ \ \mu\delta\gamma,\ \ \mu\alpha\gamma,\ \ \delta\alpha\gamma,\ \ \mu\delta\alpha\gamma$	
+ noise	

Highly Multivariate Multifactorial Experiments

Multiplicities: Full multivariate analysis – simultaneous inference: identify "real" effects

Bayesian multivariate Anova: Multiple shrinkage estimation

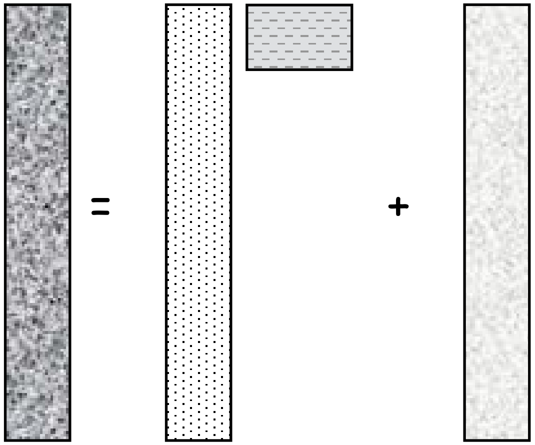
A precursor experiment: Beware the flood (RNAi, etc)

Analysis goals: Gene identification, pattern profiling of states/interactions of design variables

Translation to human samples: Predict expression signatures

Multivariate Anova

$$X = BH + N$$



Genes by samples $X_{(p \times n)}$

Gene-specific regression parameters $B_{(p \times k)} = \{\beta_{g,j}\}$

Fixed design $H_{(k \times n)}$

Design factor j: Main effect, interaction, … $\beta_{g,j}$

Multiple tests/comparisons – simultaneous inference

Substantial noise component

SPARSE Multivariate Anova

Sparsity priors: $\#\{\beta_{g,j} \neq 0\} =$ small

$$\beta_{g,j} \sim (1-\pi_{g,j})\delta_0(\beta_{g,j}) + \pi_{g,j} N(\beta_{g,j}|0,\tau_j)$$

Extends traditional sparsity priors:

Gene g, Design factor j: $\pi_{g,j}$ ~ sparsity

Invites informative, structural modelling

Hierarchical models/priors within design factors

Large p - Shrinkage and Sparsity

Model-based, automatic shrinkage - Simultaneous "multiple tests"

Decision theory/false discovery?
Estimation versus Decision?
How many comparisons/hypotheses?

$$\pi^*_{g,j} = Pr(\beta_{g,j} \neq 0|X)$$

$$E(\beta_{g,j}|\beta_{g,j} \neq 0,\ X)$$

Computation and

posterior summarisation

MCMC methods

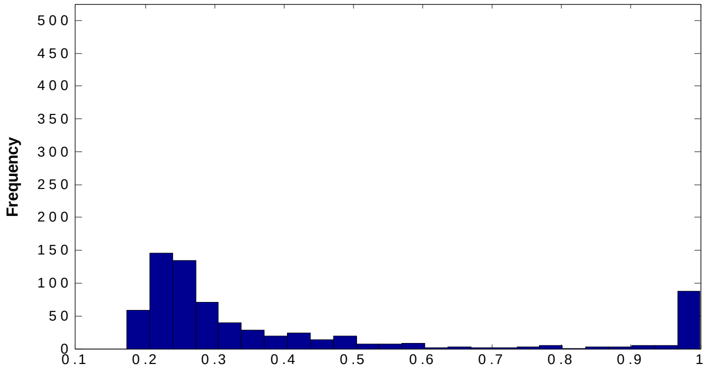
Hierarchical Sparse Anova Model

New hierarchical model:

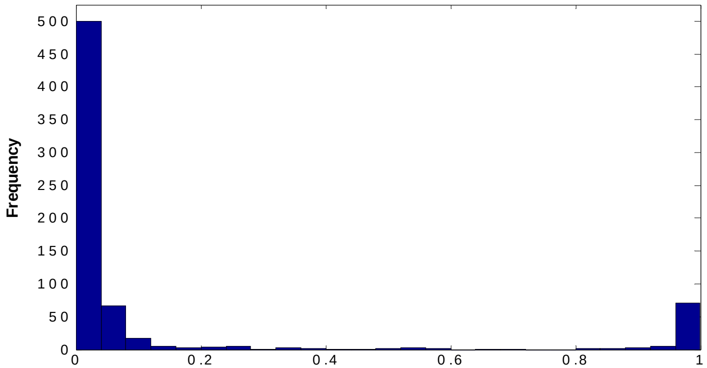
$$\beta_{g,j} \sim (1-\pi_{g,j})\delta_0(\beta_{g,j}) + \pi_{g,j} N(\beta_{g,j}|0,\tau_j)$$

$$\pi_{g,j} \sim (1-\rho_j)\delta_0(\pi_{g,j}) + \rho_j Be(\pi_{g,j}|s_j r_j, s_j(1-r_j))$$

$$p(\rho_j, \tau_j, r_j)$$



$$\pi^*_{g,j} = Pr(\beta_{g,j} \neq 0|X)$$



Design factor j: ρ_j ~ sparsity

It matters: improved signal isolation

MCMC methods vital

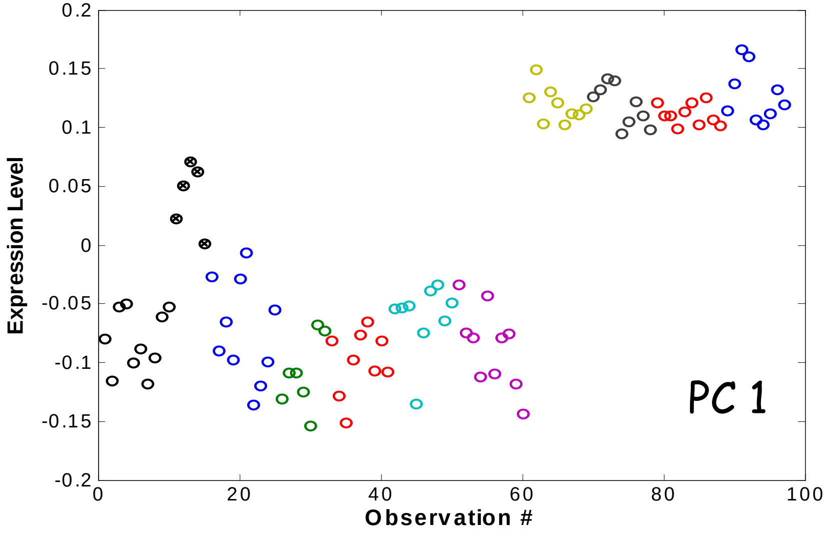
(Lucas et al 05)

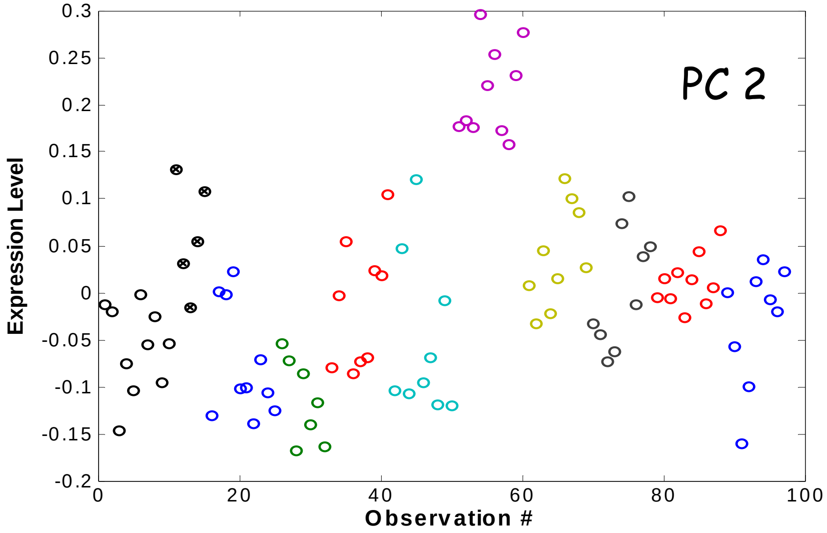
Multivariate Sparse Regression

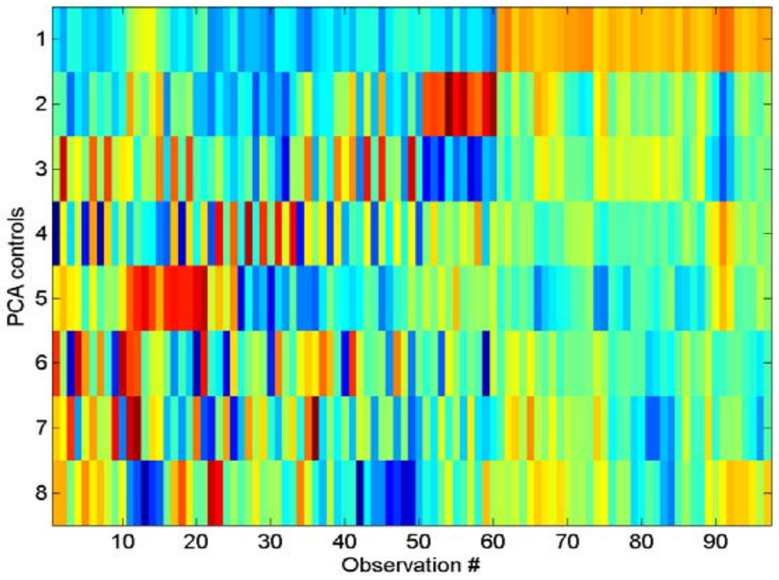
Regressions example:
"housekeeping gene control factors"

Gene-Sample-Study artifacts

Oncogene interventions experiments



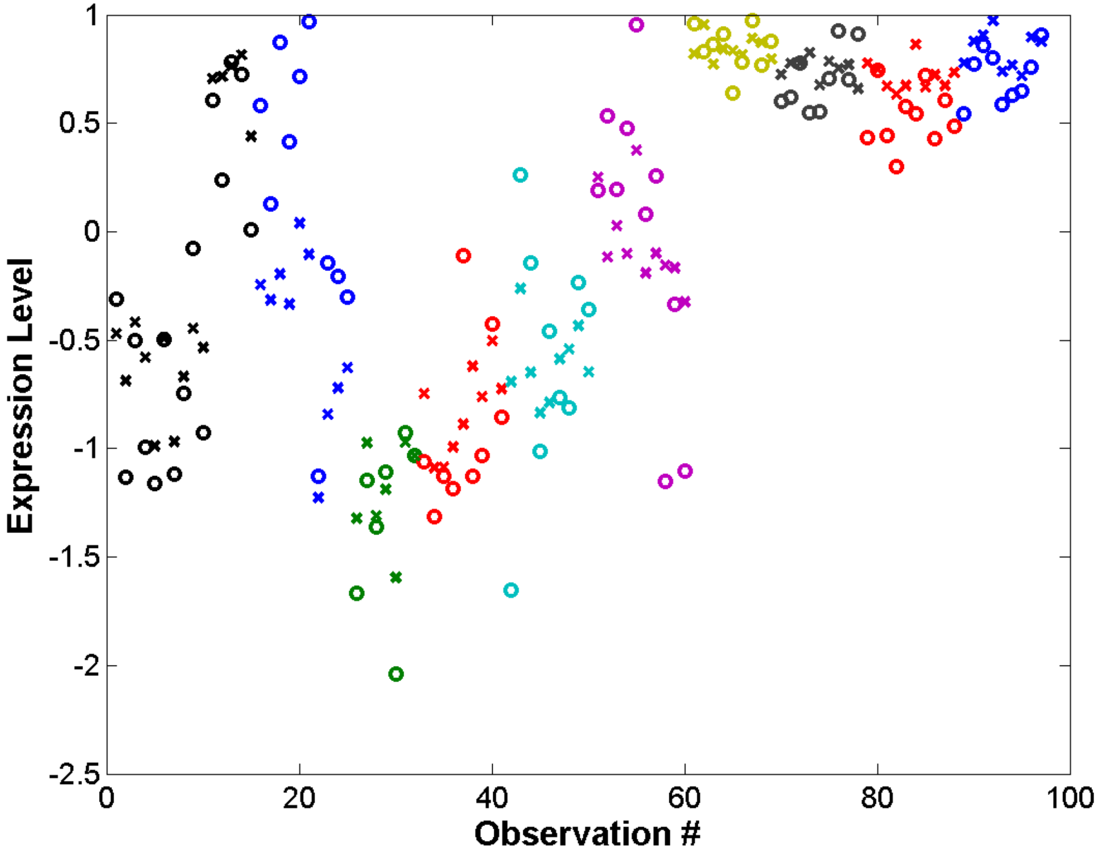




(Lucas et al 05)

Multivariate Sparse Regression: Bias Corrections

Major but Sparse effects:
Selective impact across genes

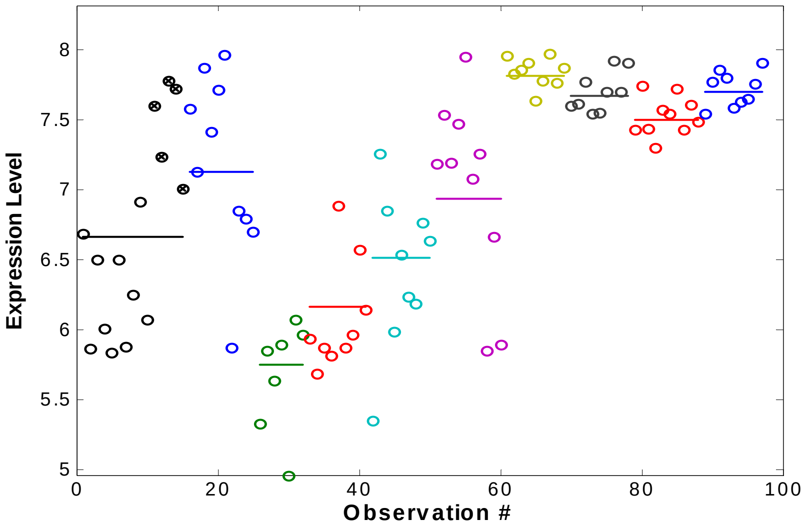


o Gene expression data

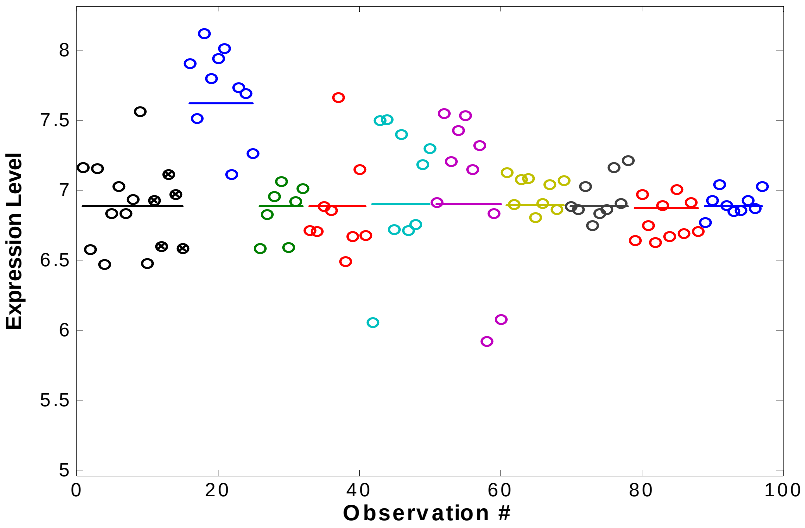
x Fitted component on control factor 1

(Lucas et al 05)

Bias Correction & False Discovery Control



No control factors



With control factors

Oncogene experiments - Myc up-regulated gene:

NFE2L1

Nuclear factor (erythroid-derived 2)-like 1
Involved in the regulation of apoptosis

Directly induced by Myc

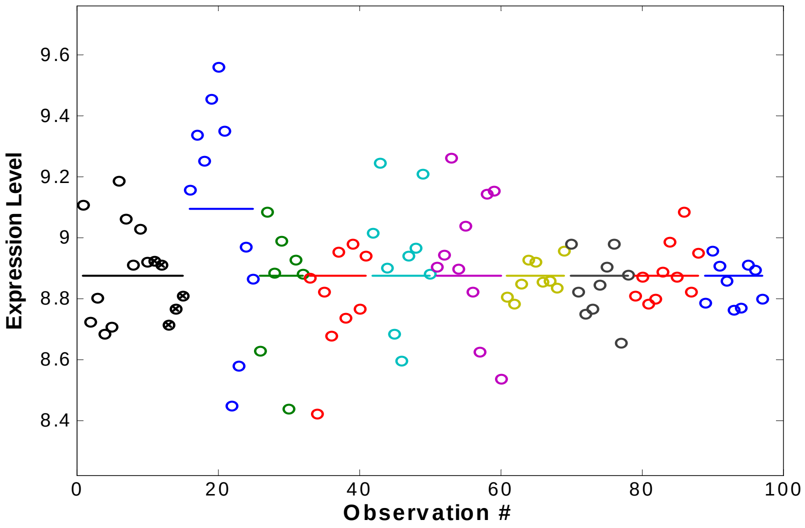
consensus binding sequence
(T/C)GCGCA(C/T)GCGC(A/G)

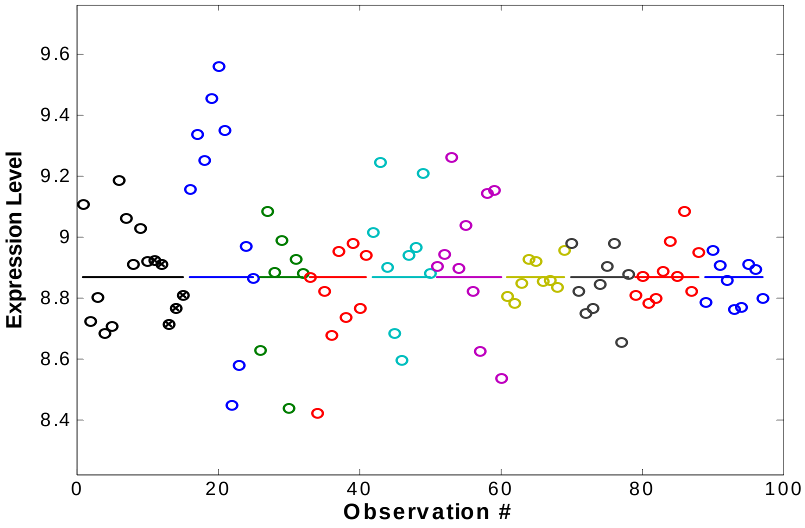
Myc binding site

GENES & DEVELOPMENT (2003-01-15)

(Lucas et al 05)

Multivariate Sparse Anova Model for Variances





Sparse ANOVA for noise variances

$$X = BH + N$$

$$N = \{\nu_{g,i}\}, \quad \nu_{g,i} \sim N(0, \exp(\kappa_{g,i}))$$

$$K_{(p\times n)} = \{\kappa_{g,i}\}$$

$$K = \Theta H + N_K$$

Sparsity prior for $\Theta_{(p\times n)} = \{\theta_{g,j}\}$

Oncogene interventions experiments:
False discovery

MCMC methods vital

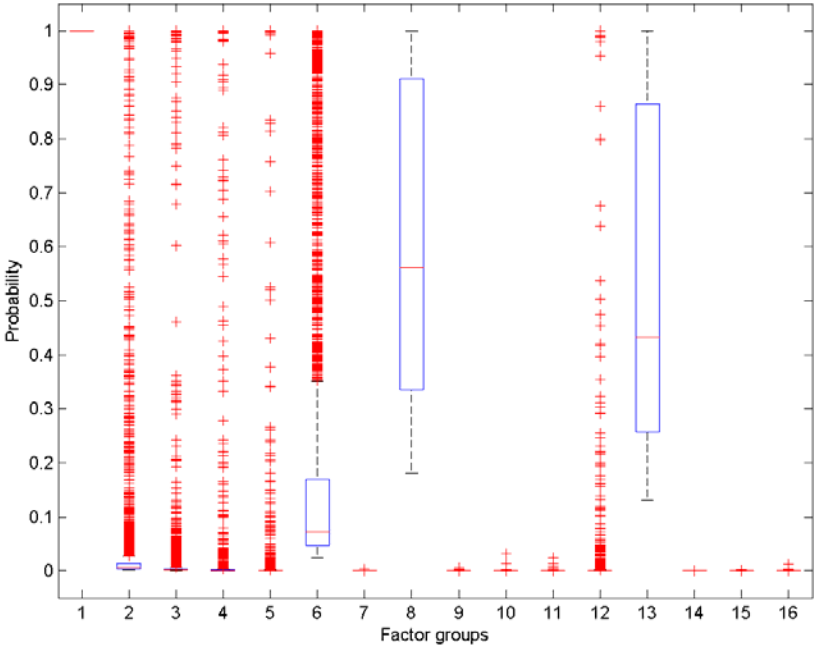
(Lucas et al 05)

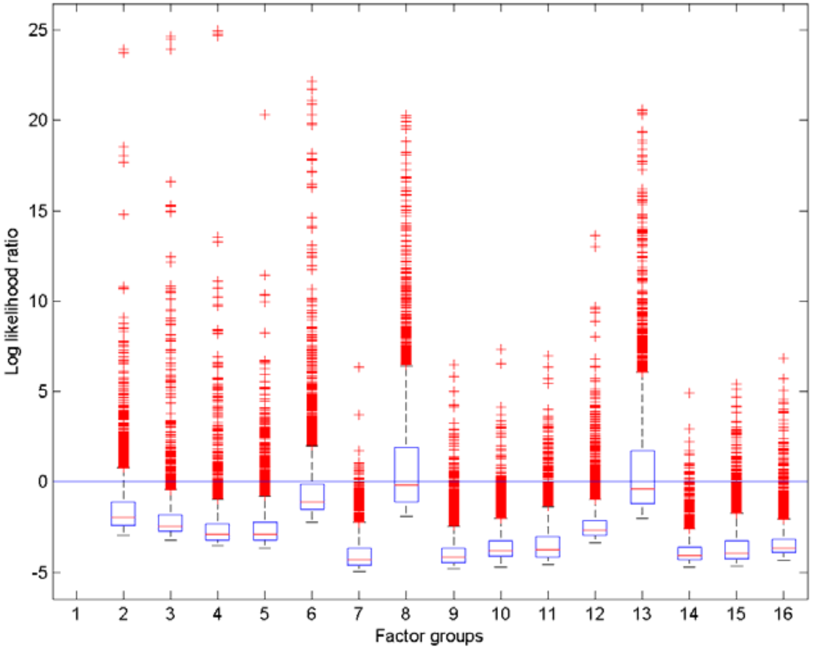
Gene-Environment Associations in Mice Experiments

Probabilities and log-likelihood ratios

- SHRINKAGE

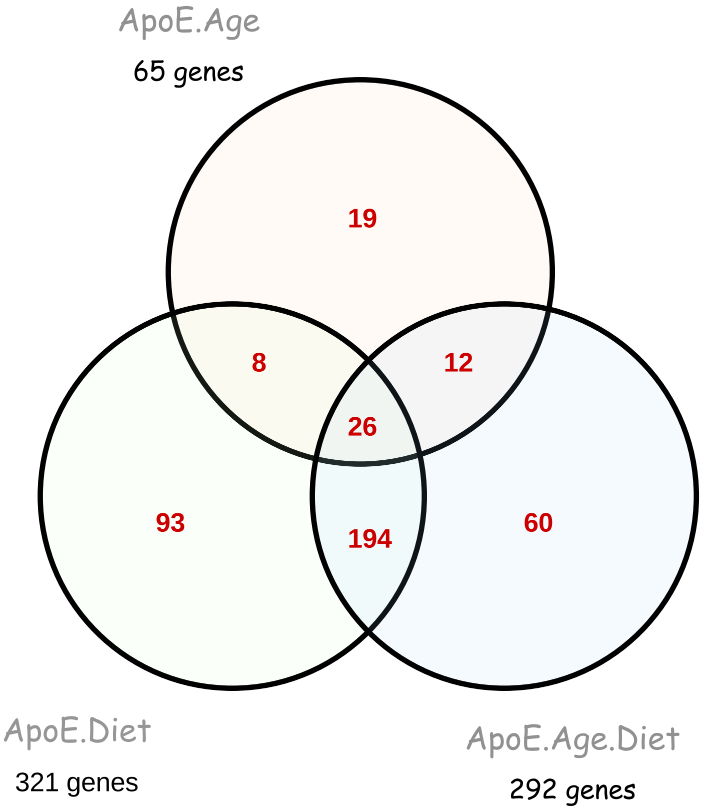
$$\pi^*_{g,j} = Pr(\beta_{g,j} \neq 0|X)$$

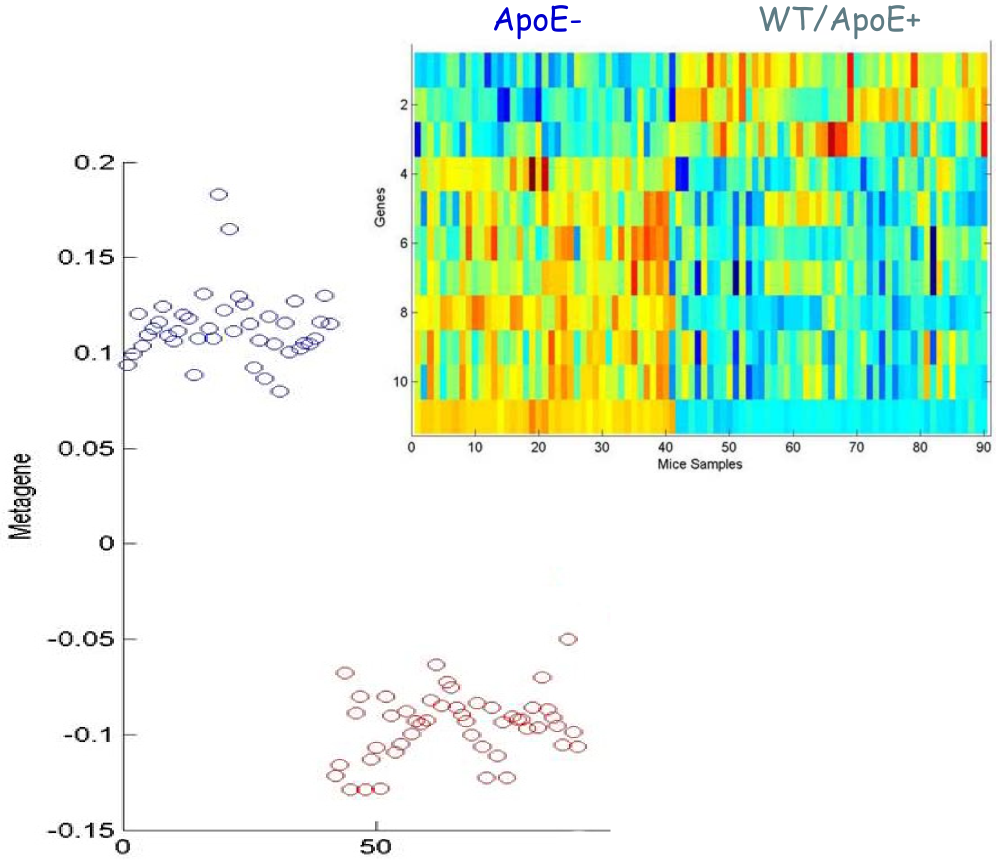




Pattern Profiles: Signatures of States

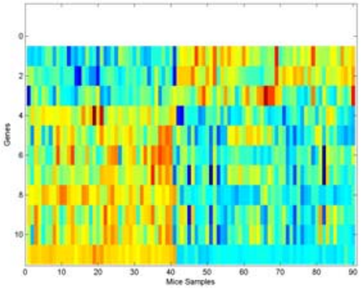
'significant' effects define characterising metagenes

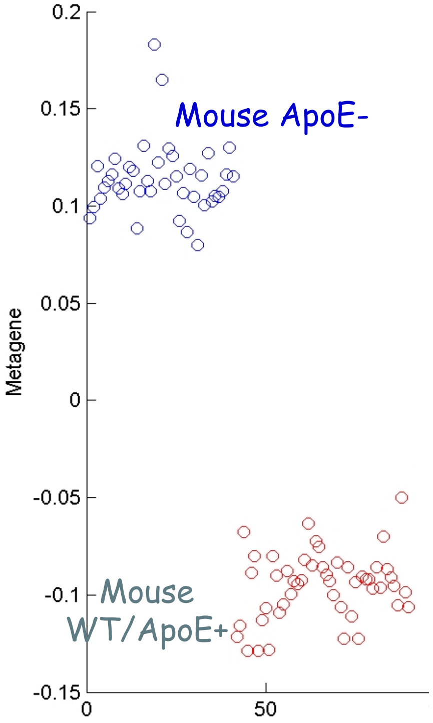




Pattern Profiling: Mouse-Human Translation

Mapping genes from mouse to human normalise/evaluate/extrapolate signatures



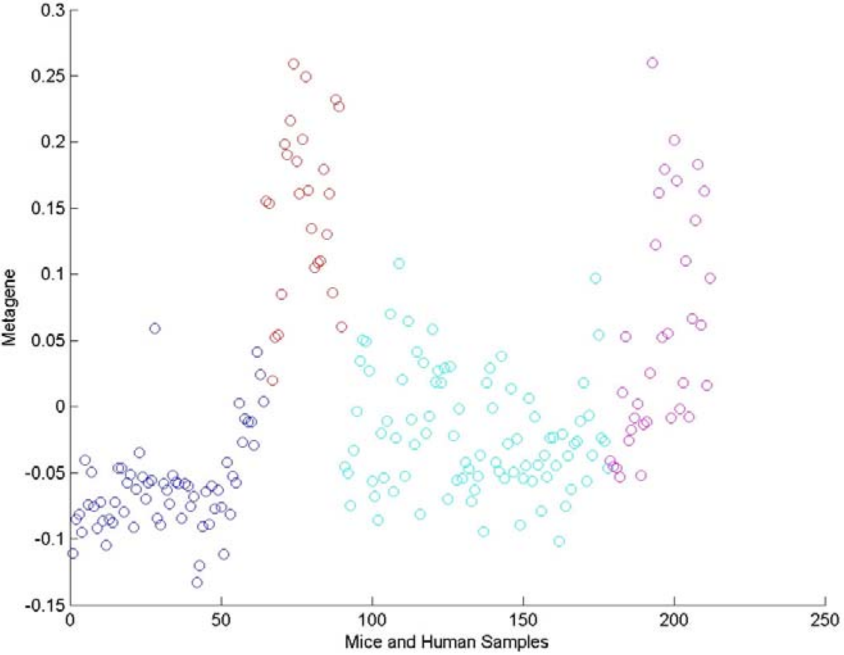


Human D+/-

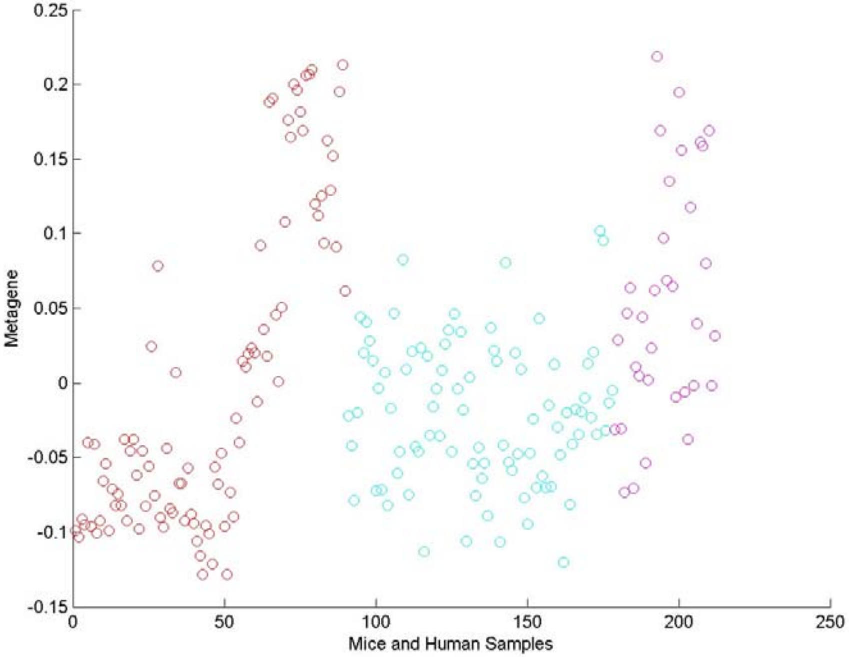
Aorta lesions, advanced atherosclerosis

Pattern Profiling: Mouse-Human Translation

ApoE.Age



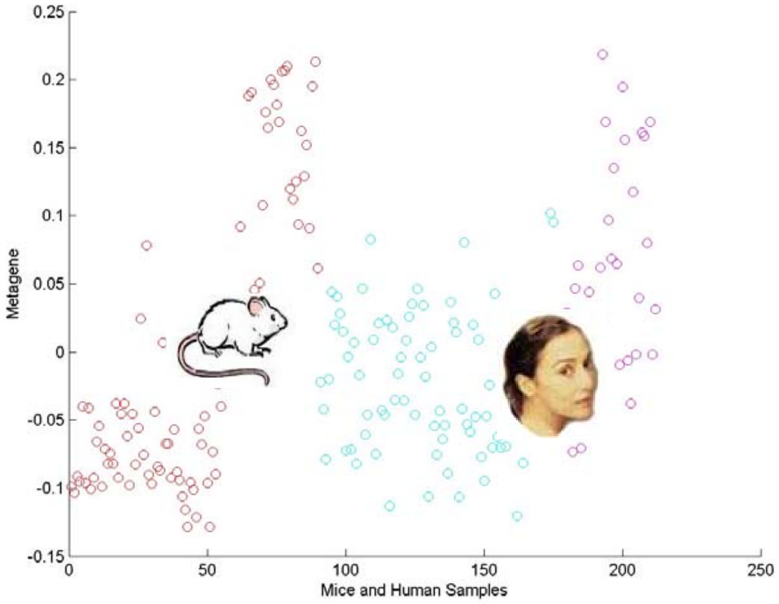
Intersection of
ApoE,
ApoE+{Age,Diet}



Translation – Needs and Next Steps

Improved pathway characterisation:
implicated genes

Disease Risk Signature



More/better phenotypes in humans

Blood/serum assays for gene expression

Metabolites in serum and blood:
metabolic/genomic fusion

(Karra et al 04; Seo et al 05)

Flightplan

#1 Genomics, Microarrays, Data:
Big picture

#2 Bayesics - Regression and Shrinkage:
Gene expression as predictors

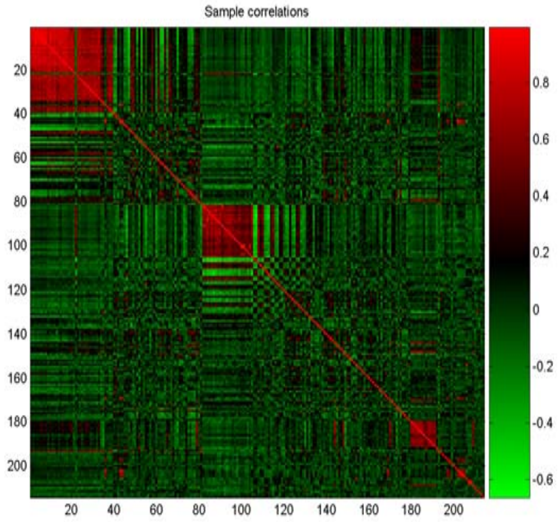
#3 Patterns and Factors:
Prediction via pattern profiling

#4 Sparse Modelling:
Regression subset-structure uncertainty

#5 Sparse Models and Profiling:
Gene expression as response: Designed experiments

#6 Sparse Models and Profiling:
Gene expression as response: Latent factor models

Latent Factor Models & Patterns of Association

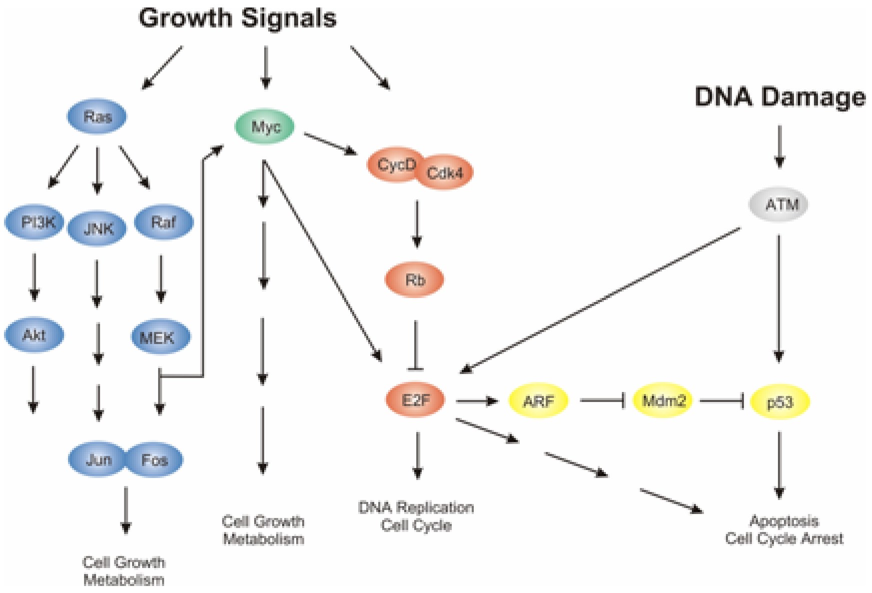


Decompositions of p(x)

Latent structure underlying associations

Cancer Studies:

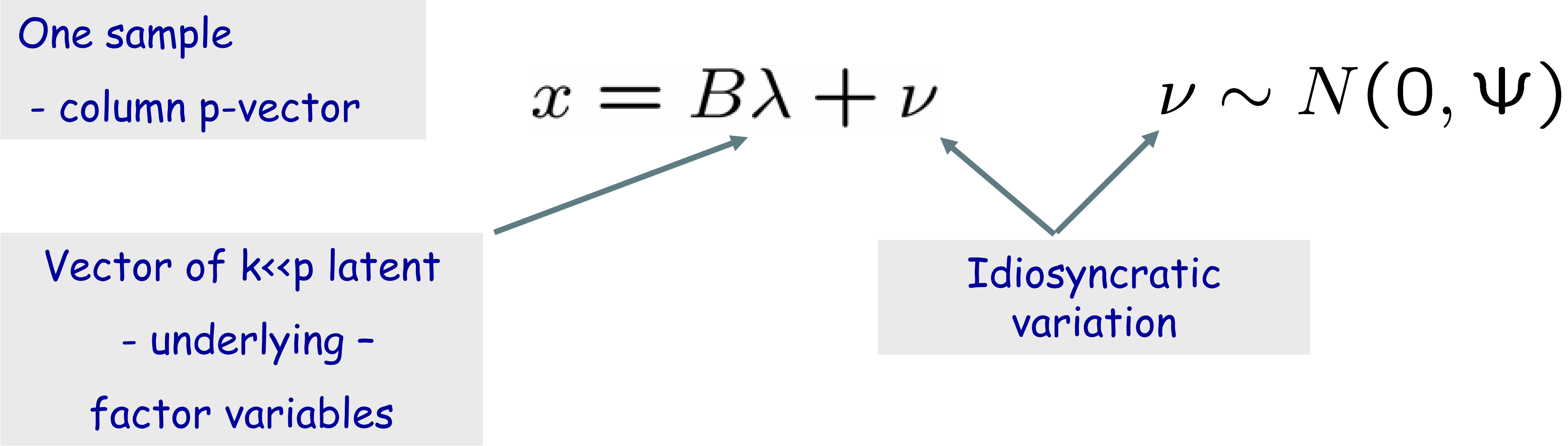
Multiple deregulated pathway components



Latent Factors:

intersecting sub-pathways

Latent Factor Model Structure



One sample

- column p-vector

$$x = B\lambda + \nu \qquad \nu \sim N(0, \Psi)$$

Vector of k<<p latent

- underlying –

factor variables

Idiosyncratic variation

Latent factors: $\lambda \sim N(0, T)$

Model of covariance matrix: $V(x) = BTB' + \Psi$

(West 03, Valencia 7; +Aguilar 01 JBES; +Lopes 04 Stat Sinica)

Sparse Models:

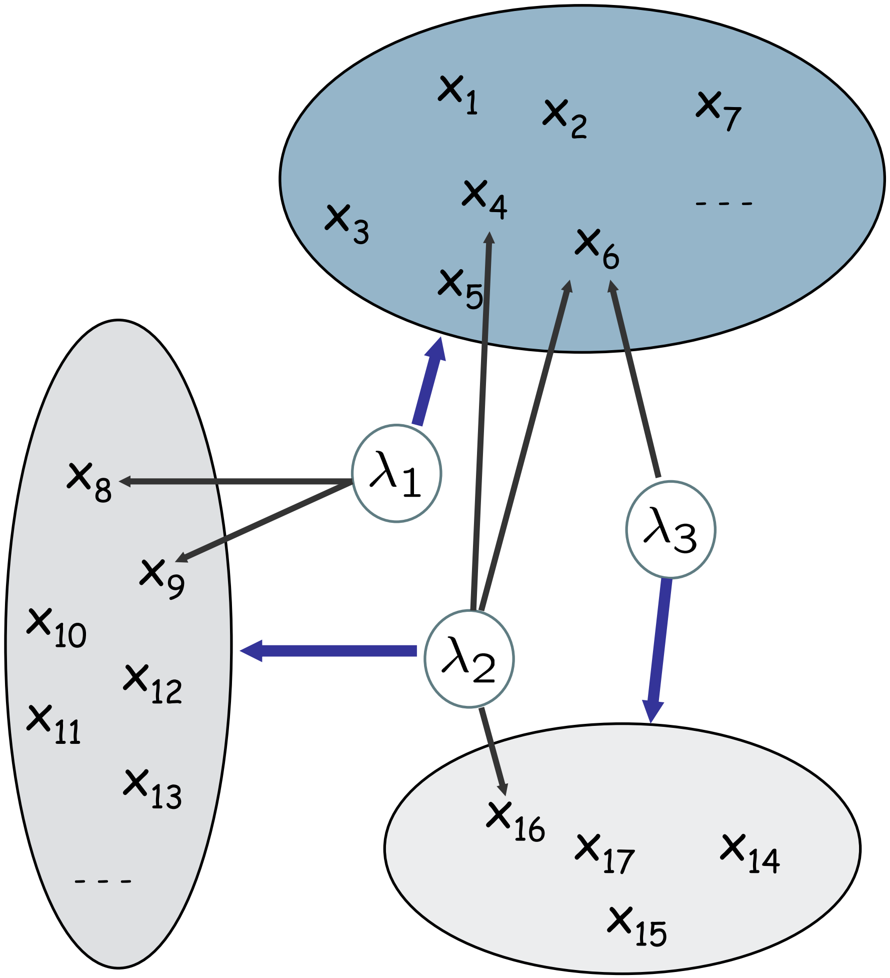
One factor – few or many variables

One variable – 0,1, or few factors

$$B = \{\beta_{g,j}\}$$

Row (variable) g, factor j:

$\#\{\beta_{g,j} \neq 0\} =$ 0,1, .., small



(West 2003, Valencia 7)

Sparsity and Structure

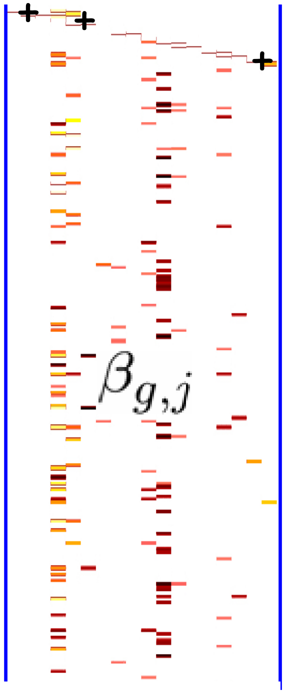
Uncertain sparsity patterns:

$$\beta_{g,j} \sim (1-\pi_{g,j})\delta_0(\beta_{g,j}) + \pi_{g,j} N(\beta_{g,j}|0,\tau_j)$$

$$\pi_{g,j} \sim (1-\rho_j)\delta_0(\pi_{g,j}) + \rho_j Be(\pi_{g,j}|s_j r_j, s_j(1-r_j))$$

$$p(\rho_j, \tau_j, r_j)$$

B =



(West 2003, Valencia 7)

Structure:

Parametrisation of **B** - Identification

"founders" of factors

Computation: Global MCMC

$$X = B\Lambda + N$$

$$B = [b_1, \ldots, b_k], \quad \Lambda = [\lambda_1, \ldots, \lambda_n]$$

$$\Pi = \{\pi_{g,j}\}$$

$$p(b_j|\text{else}) = \prod_{g=1}^{p} p(\beta_{g,j}|\text{else})$$

Other (hyper) parameters "easy"

$$p(\Pi|\text{else}) = \prod_{g=1,j=1}^{p,k} p(\pi_{g,j}|\text{else})$$

Not parallelizable:

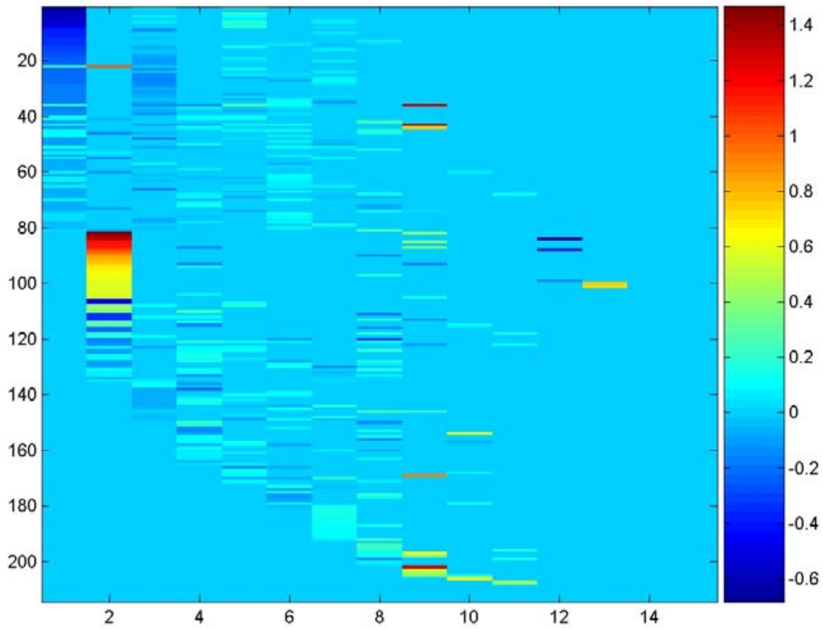
Serial Gibbs/MCMC

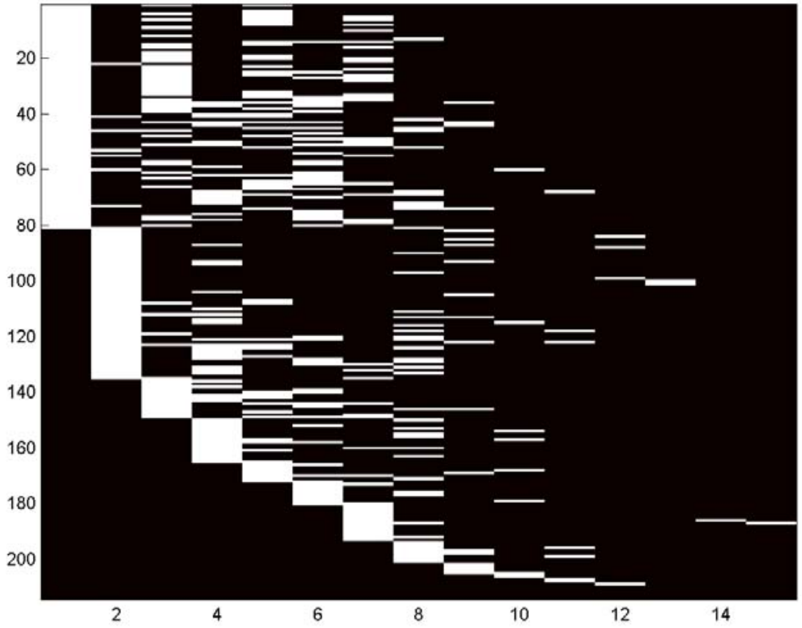
$$p(\Lambda|\text{else}) = \prod_{i=1}^{n} p(\lambda_i|\text{else})$$

Parallel within Gibbs iterates

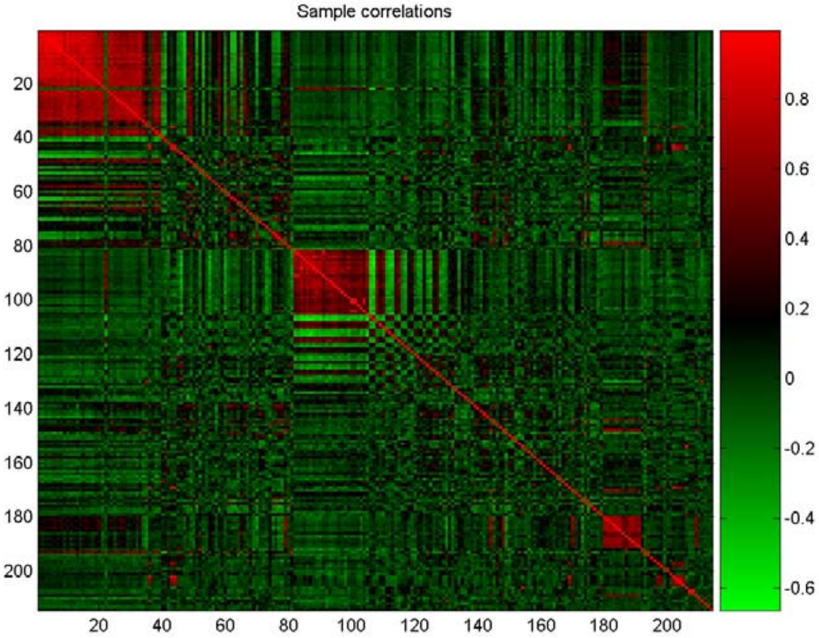
Inference on Sparsity and Structure

Aspects of p(B,Π|X)

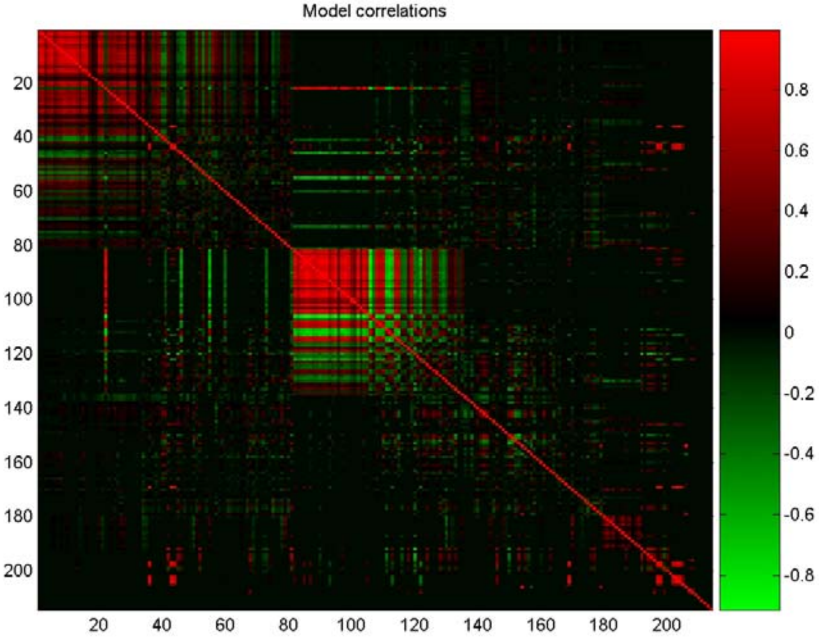




Factor Models and Association Patterns

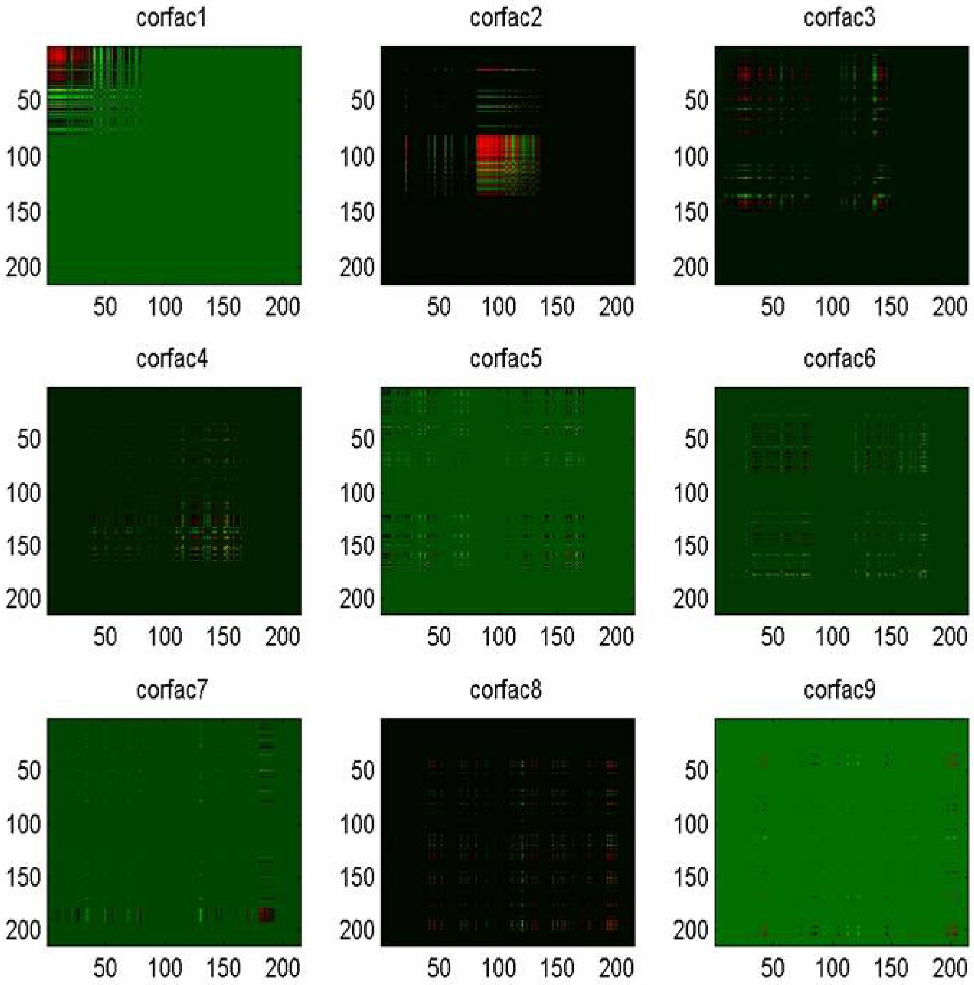


Sample correlations



Fitted correlations
in B'TB+Ψ

Decomposing Associations by Factor



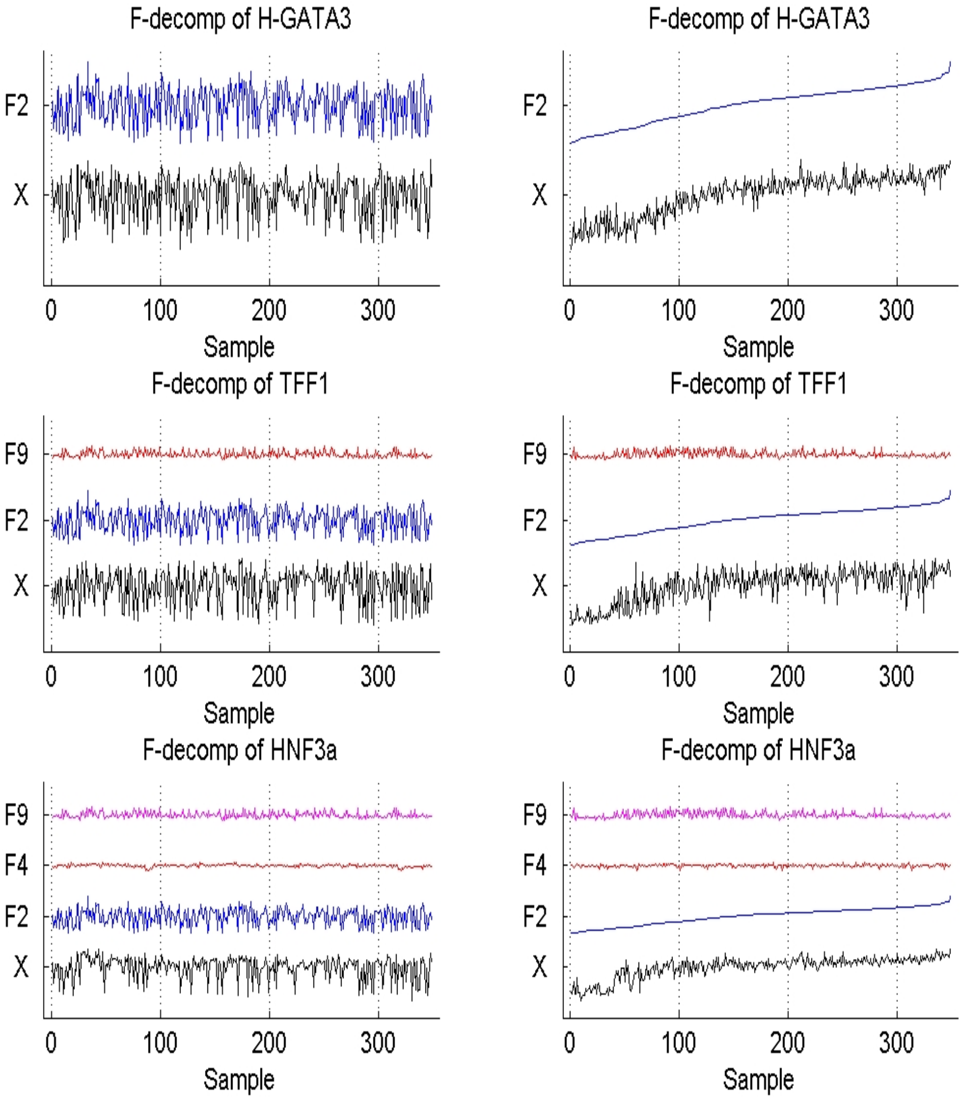
Covariance decompositions:

$$BTB' = t_1 b_1 b_1' + t_2 b_2 b_2' + \cdots$$

Decomposition by Factor Signatures

Data decompositions:

$$x_g = \beta_{g,1}\lambda_1 + \beta_{g,2}\lambda_2 + \cdots$$



Breast cancer gene expression

Factor 2: ER, 9: HER2

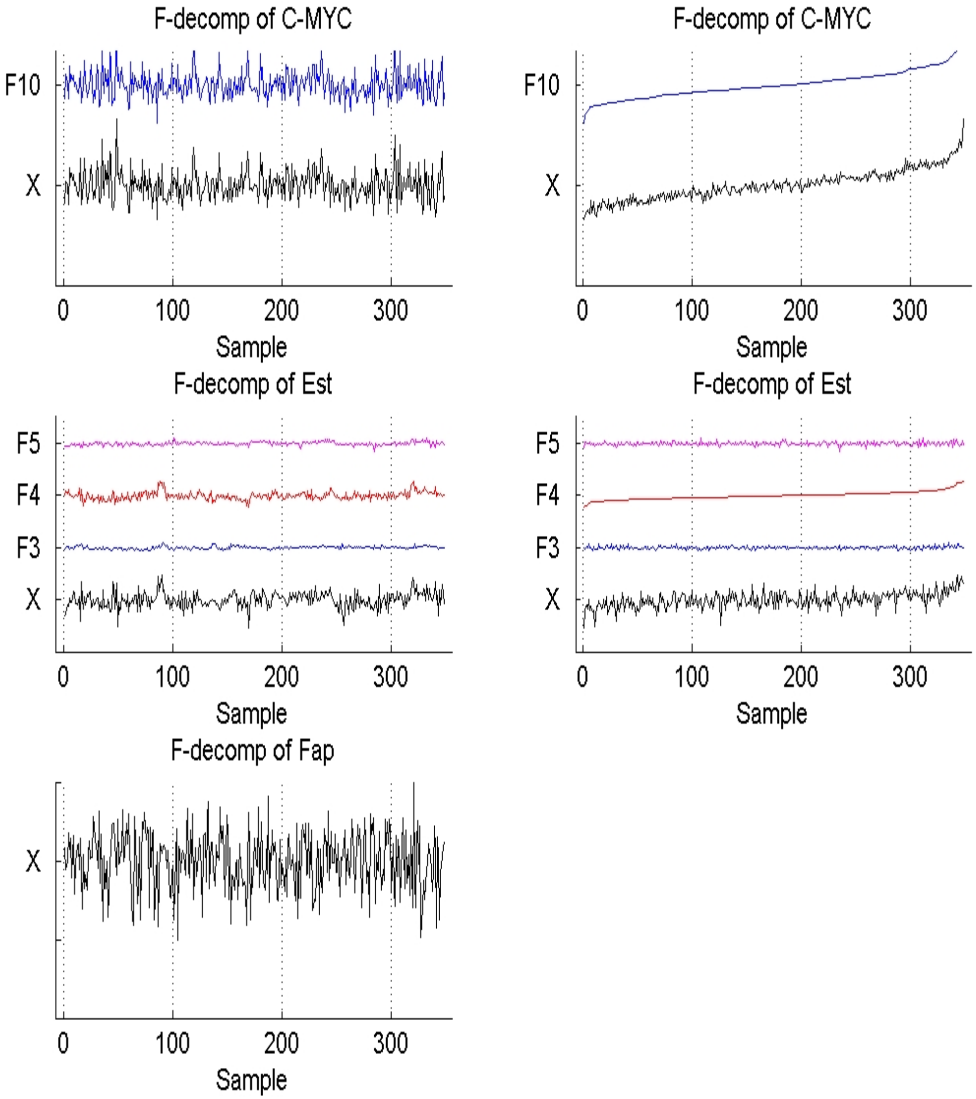
Decomposition by Factor Signatures

Discovery of gene-factor associations

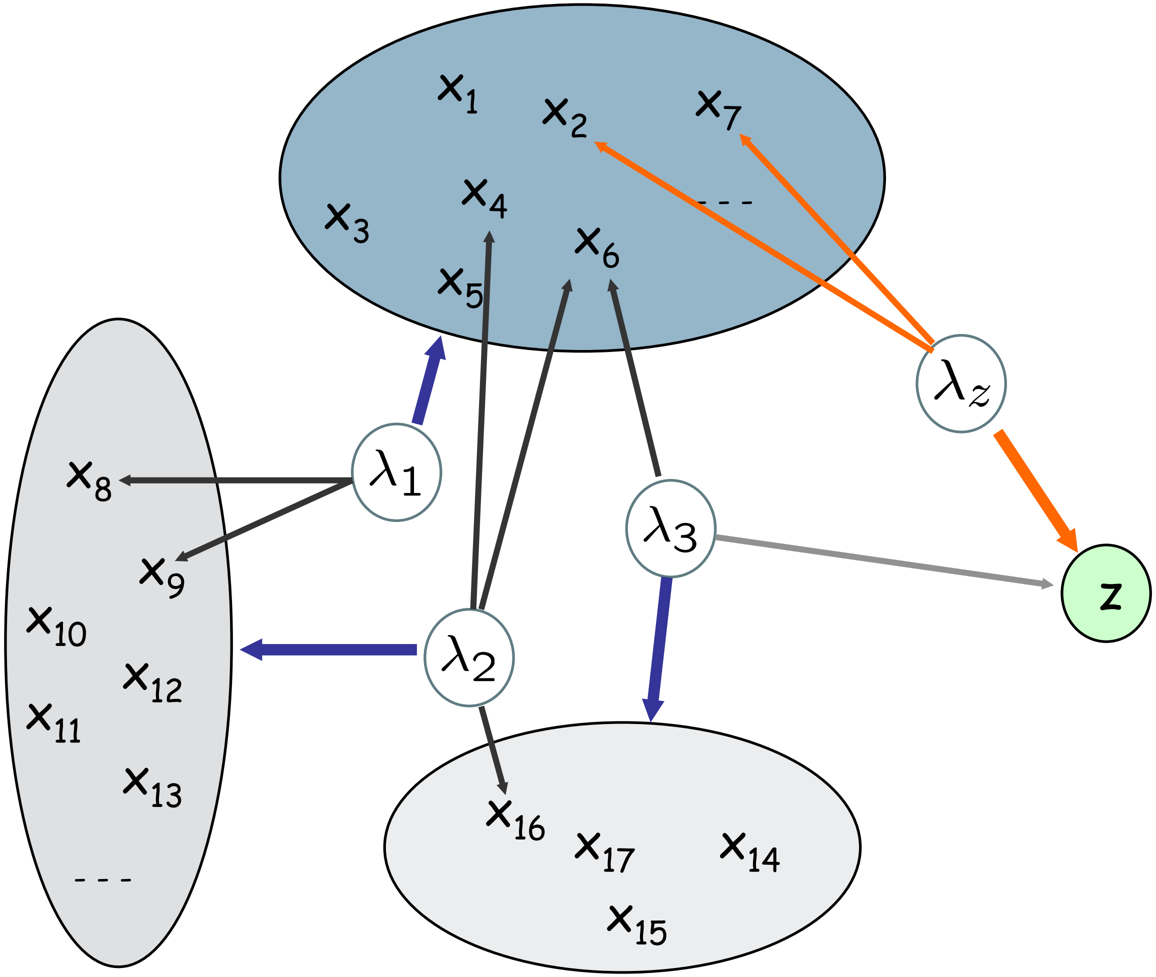
Isolation of non –associated genes

… noise/idiosyncratic

Factors
=
Patterns,
Signatures



Sparse Latent Factor Regression



Response variables z

Evaluate p(z,x)

Predict z

z linked to some latent factors in x space?

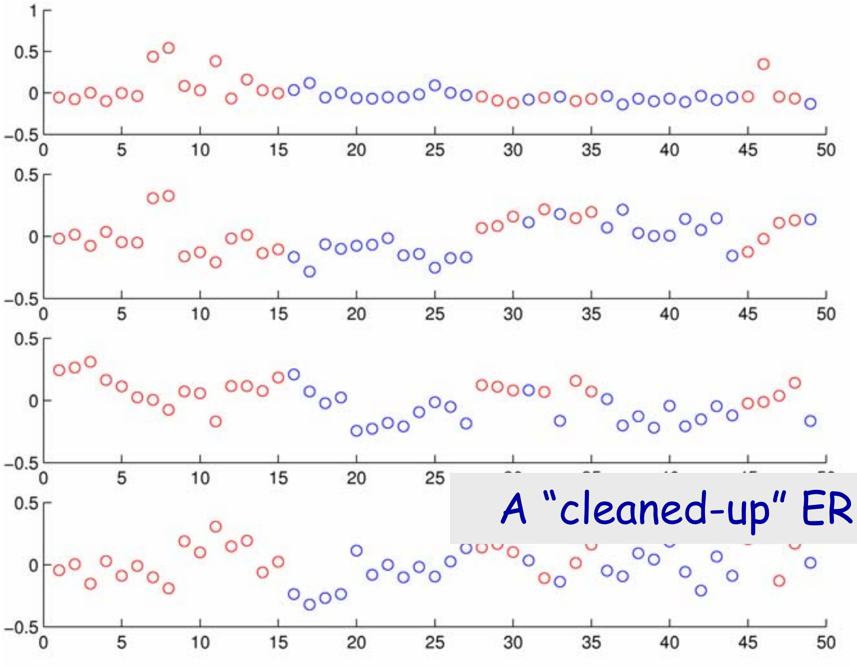
… and to some individual x variables?

Response factors λ_z

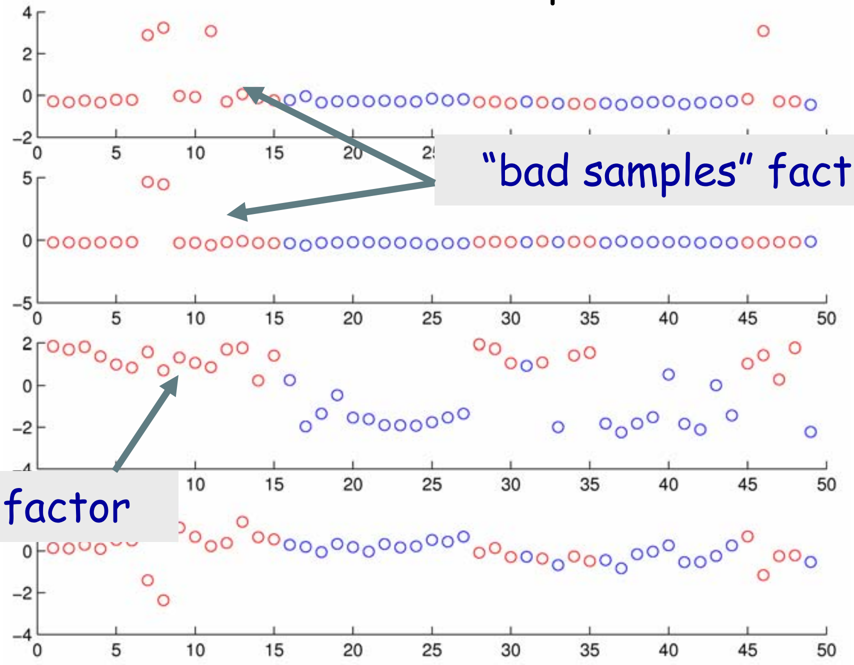
(West 2003, Valencia 7; Carvalho et al, 2005)

Sparse Factor Regression: Predictive Signatures

SVD



Sparse Model



"bad samples" factors

A "cleaned-up" ER factor

Factor models "clean up" SVD

SVD "noisy" factors

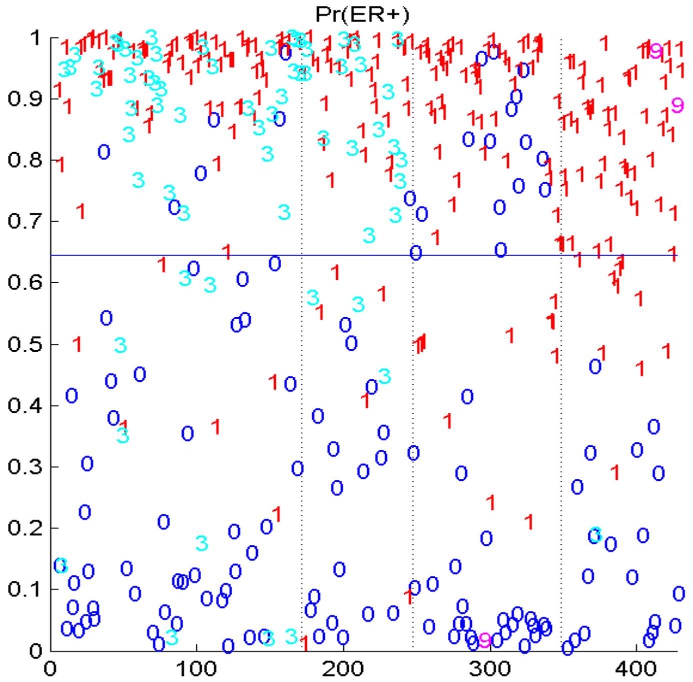
(West 2003, Valencia 7)

Sparse Factor Regression: Predictive Signatures

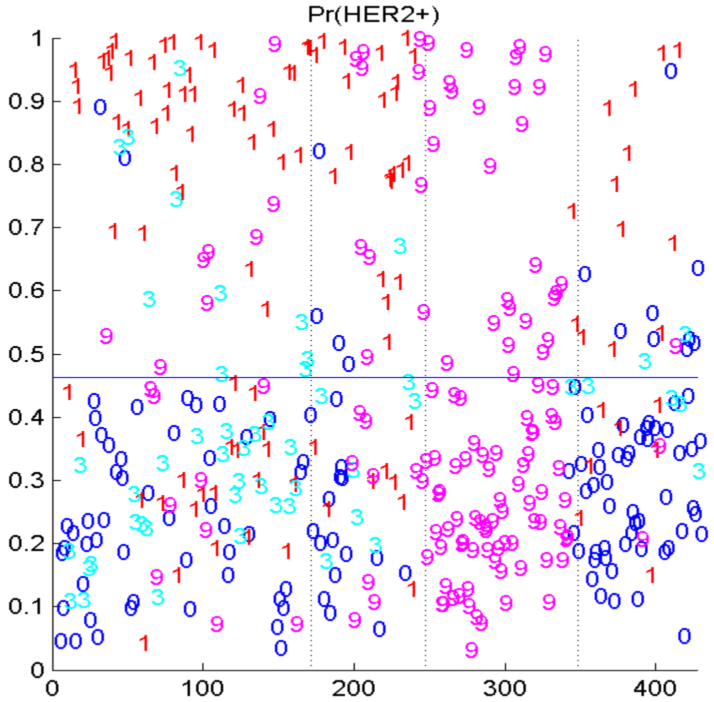
Expression assays of hormonal status

Jointly predicted: Multiple factors

4+ studies

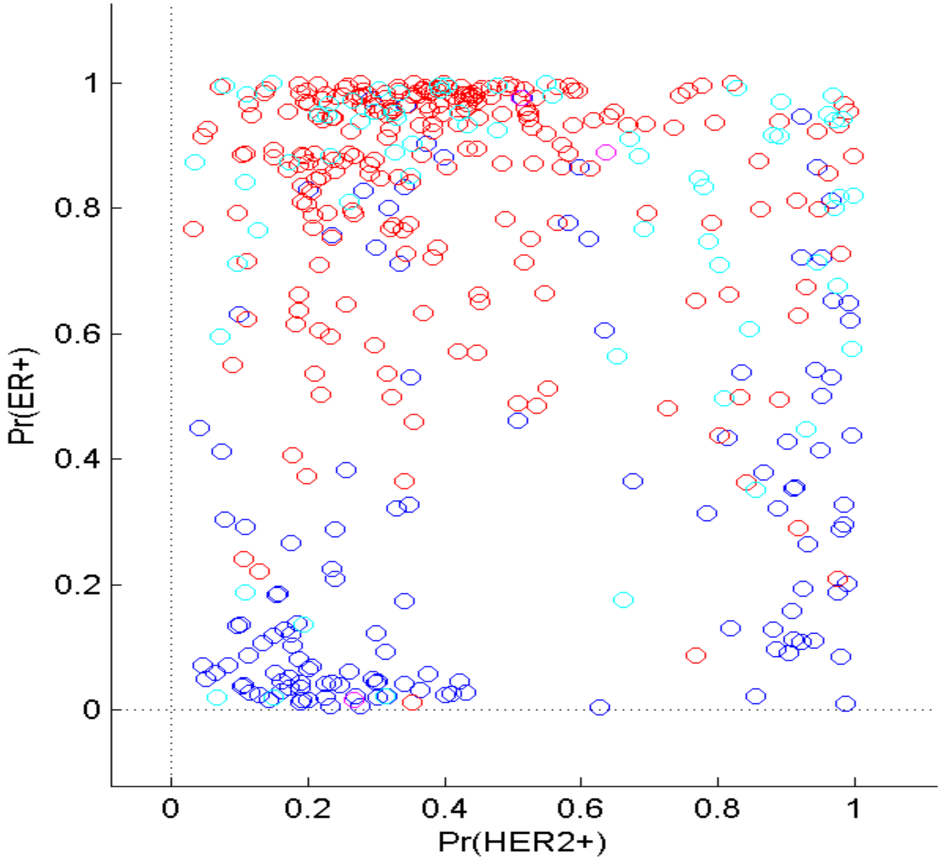


ER



HER2

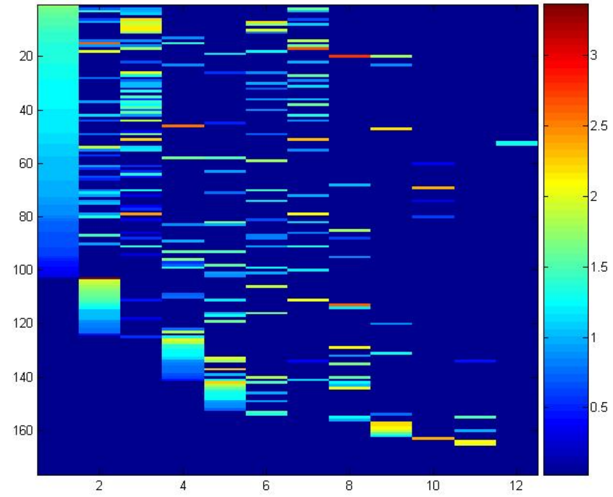
Sparse Factor Regression: Predictive Signatures



Jointly predicted:

Multiple tests/assays

Evolutionary Exploration and Enrichment of Gene Set



E2F Pathway:

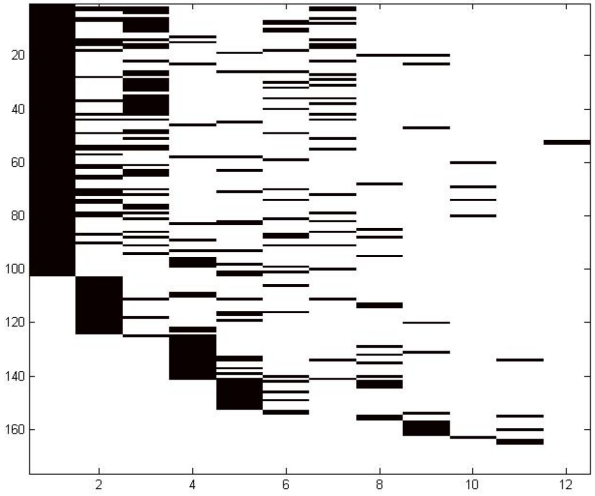
40 initial genes

- Replication
- Proliferation
- Cell cycle:check point
- Apoptosis
- "Structure"

[illegible] genes

[illegible] factors

Factor model fit to breast cancer



Thresholded gene-factor probabilities & effects

Evolutionary analysis: Add/delete genes, factors

Self-organising factor analysis

"Greedy" Stochastic Search methods – MCMC cousins

(Carvalho et al 05; Hans et al 05)

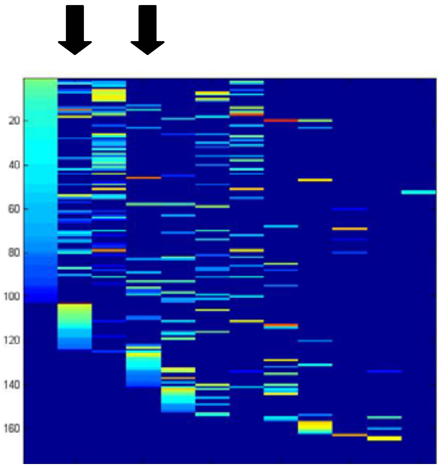
Exploratory/Discovery Analyses

2
(apoptosis)

- MDM2
- GSN
- AGC1
- RBM8A
- MTHFS
- CUGBP2
- PKN2
- SSR4
- FANCG
- NME3
- POU4F1
- AGC1
- MDM2
- CRHR1
- H1FX
- RPS3A
- ABCB8
- RGS12
- GLG1
- DOC-1R
- TNPO3
- MDM2
- MAPT
- LOR
- GUCA1A
- GRIA1
- CDC34
- COL11A2
- MYC
- TBL3
- BTF3
- UCP3
- LBA1
- CDKN2C
- HTR6
- CDC6
- CYP2A13
- KHDRBS1
- KIAA0284
- PEX5
- CYP2A6
- LTK
- SSTR3
- MDM2

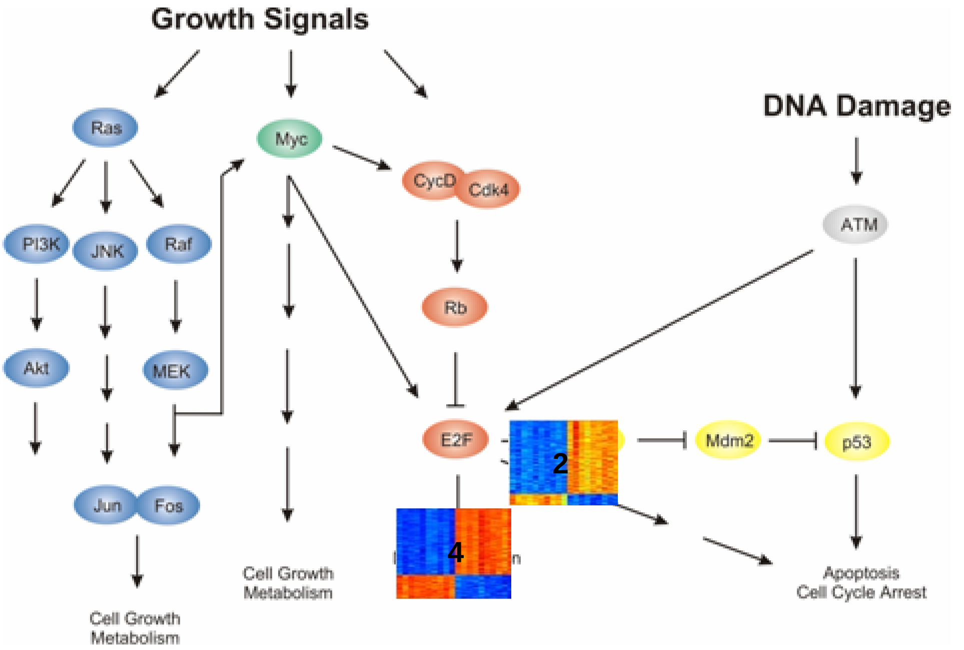
4
(DNA replication)

- MCM6
- VCAM1
- CYP2A13
- PITX1
- MCM2
- DDX39
- MCM7
- GSN
- GSTM1
- CCNE1
- MCM3
- MCM4
- MFGE8
- ABCA3
- CDKN2A
- MCM5
- KIAA1026
- TOMM70A
- CDC2
- SAS
- POLR2H
- CSNK1D
- NME3
- CDC6
- CHC1
- TNFRSF14
- BACH
- MVD
- FANCG
- LIG1
- SF3B4
- CCNE1
- ABCF3

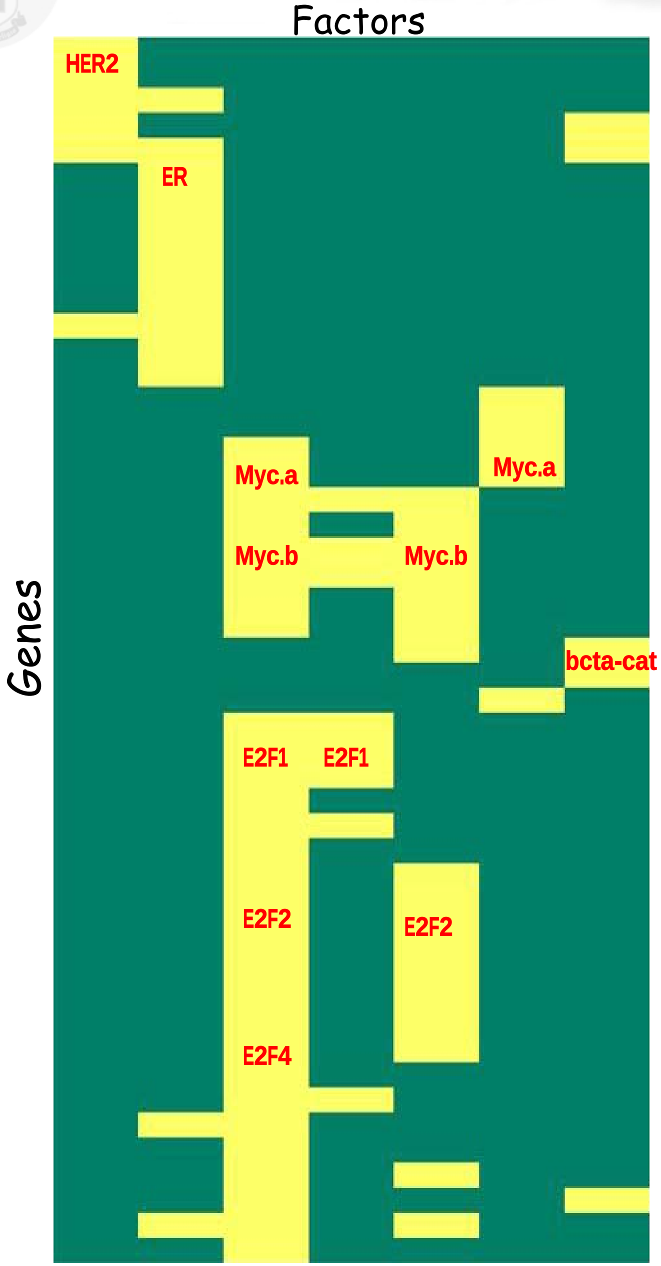


Evolutionary gene set enrichment
Experiments: gene set responses
to perturbations?
Promotor TF interactions

Duke/NCI Systems Biology Center



Structure and Skeletons of Variable-Factor Wiring



Biologically "Hard" vs "Soft" interactions

TF binding sites

P-P interactions

Expression expts

Priors over sparsity probabilities

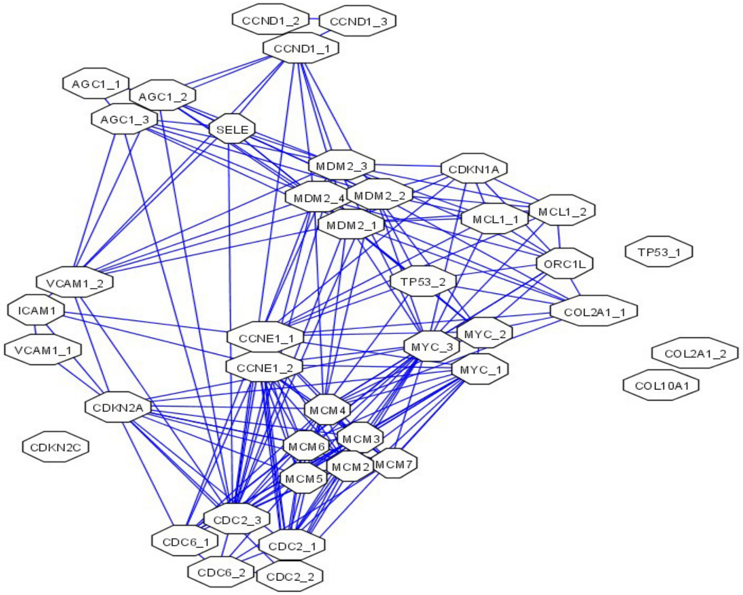
Status, relative "activation", deregulation

Priors over non-zero effects

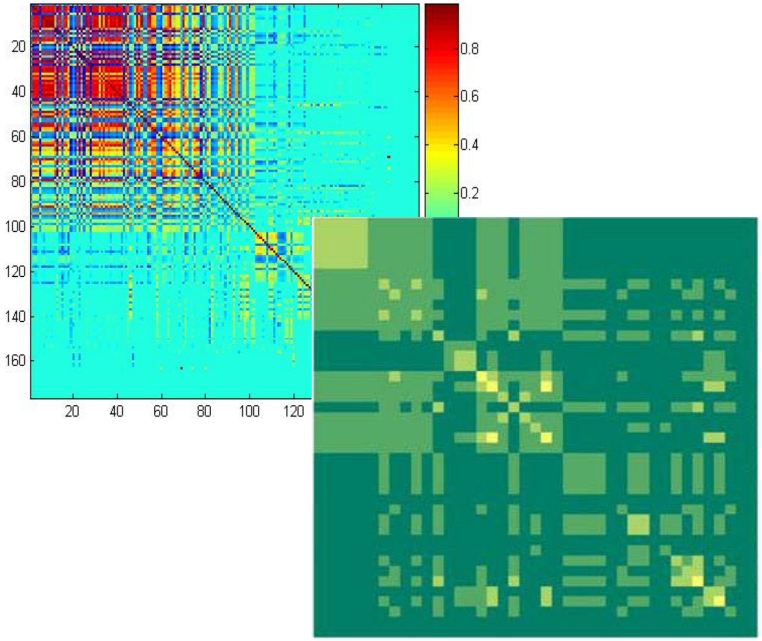
Graphs, Networks and Factor Models

Graphs of sparse factor models – covariance or association graphs

Graphical models – precision (inverse covariance) graphs



(Dobra et al 04, J Multivariate Analysis;
Jones et al 05, Statistical Science;
Rich et al 05, Cancer Research)



Multiple cancer data sets – "Meta-analysis"

Constant/consonant structure

- network "motifs"

(Zhou et al USC/PNAS 05)

Comprehensive Sparse Models for Gene Expression

Sparsity models for:

- Design effects
- Control/correction factor effects
- Latent structure: Complex associations
- Response prediction

Combined model

(Carvalho et al 05; Lucas et al 05)

Computation, Model Search: MCMC and MCMC-inspired Stochastic Search

These slides:
www.isds.duke.edu/~mw/downloads/SemStat05

Papers, software, many links:
www.isds.duke.edu/~mw

ABS04 web site: Lecture slides, stats notes, papers, data, links:
www.isds.duke.edu/~mw/ABS04

Integrated Cancer Biology Program
icbp.genome.duke.edu

Genome Institute @ Duke
www.genome.duke.edu

End 6f-Flights

Chris Hans

Joe Nevins

Carlos Carvalho

Beatrix Jones

Quanli Wang

Joe Lucas

www.isds.duke.edu/~mw