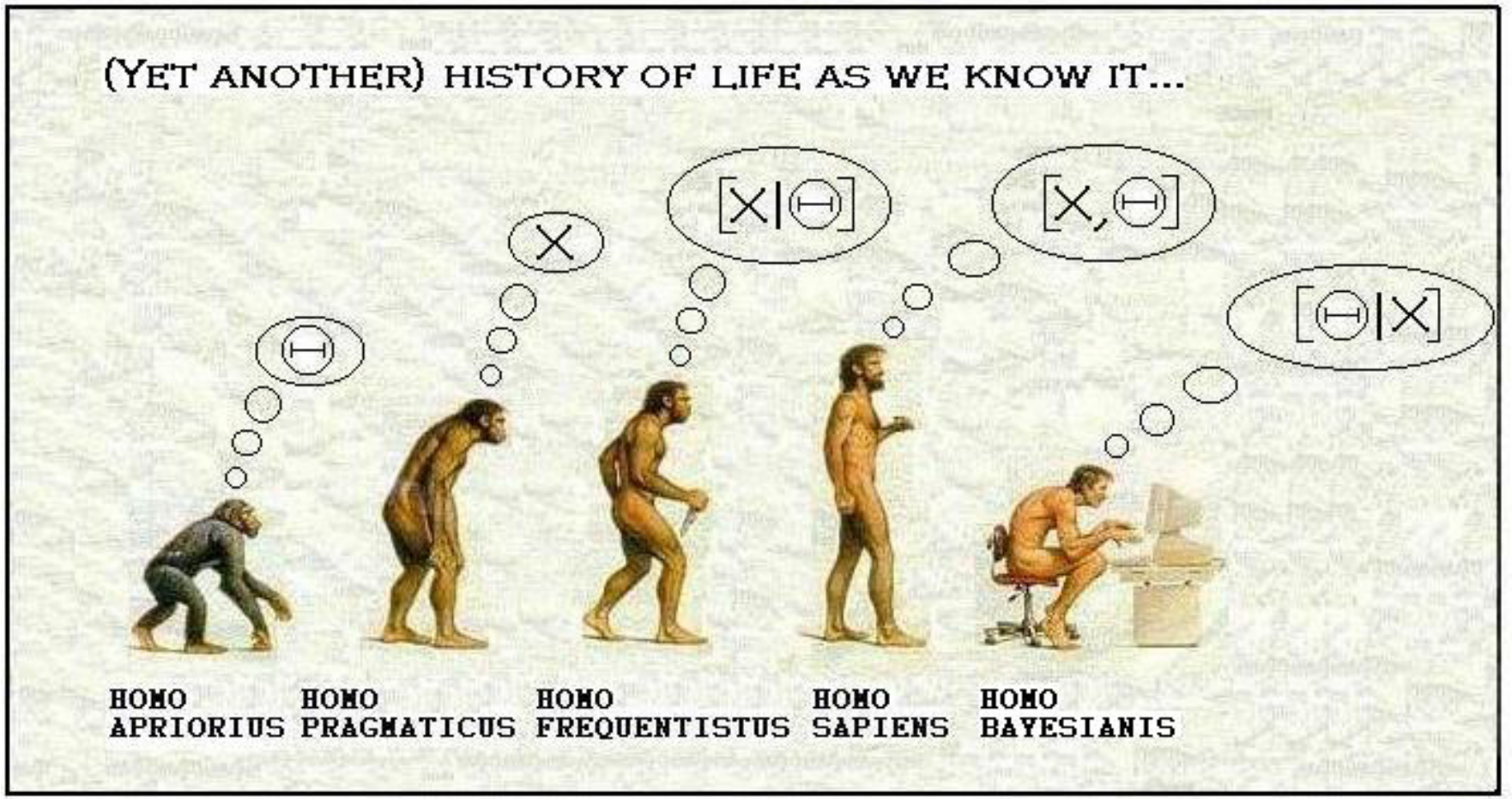
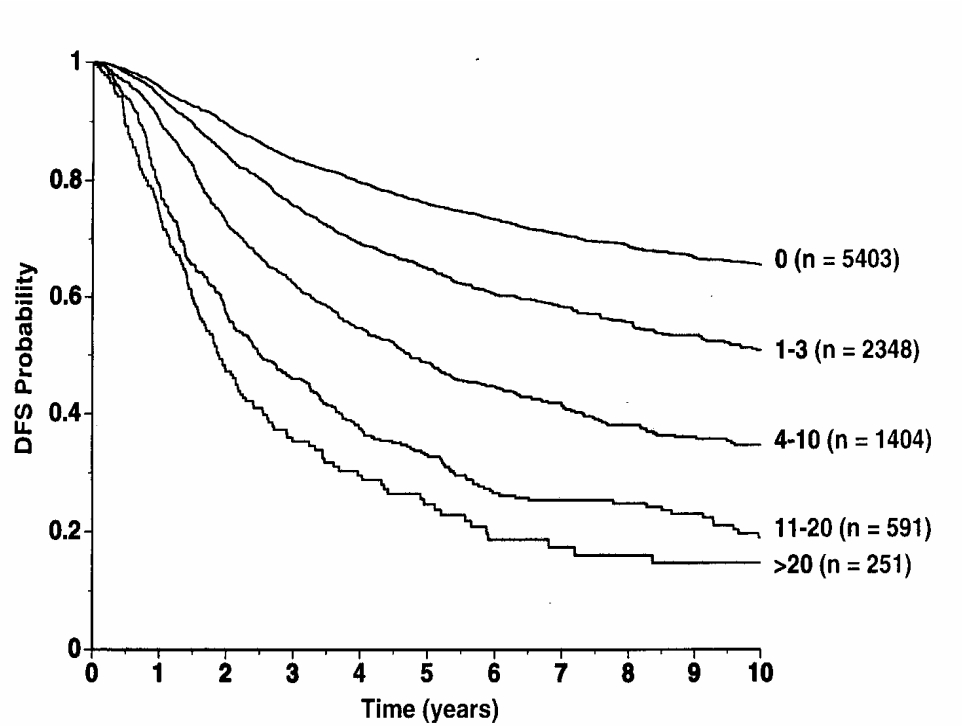
Bayesian Prediction Tree Models


(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

Statistical Prediction Tree Modelling for Clinico-Genomics

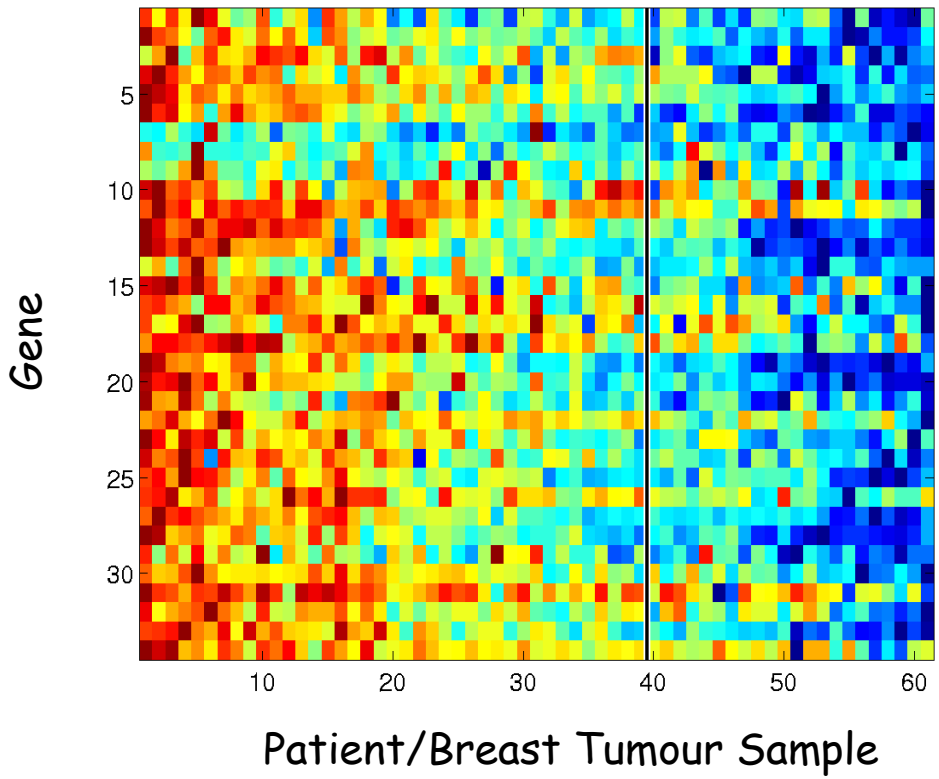
- Clinical gene expression data
  - expression signatures, profiling
- Tree models for predictive sub-typing
- Combining clinical + molecular data
- Gene discovery and prioritisation
- Breast cancer prognosis - clinical testing

Lymph Node Metastasis
- Key Breast Cancer Risk Factor -



But … lymph node dissection carries morbidity, inaccuracy

Gene Subsets Associated with Axillary Lymph Node Status



Expression Signatures - Metagenes

Multiple related genes:

- multiplicities, redundancies
- idiosyncratic noise, variability

Aggregate "Metagene" summaries:

- characterize patterns
- noise reduction in estimating common "signal"
- improve statistical efficiency, robustness
- multiple "small" clusters of genes

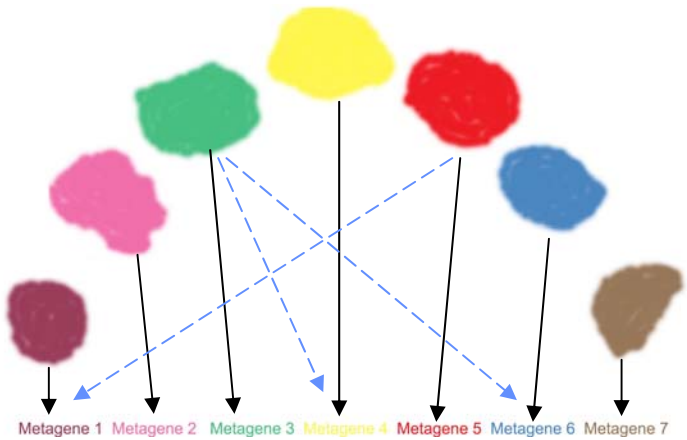
Metagenes – Characterising Factors

$Y = A(M) + E$

- Dimension reduction: Signal improvement
- Clustering, k-means
- Empirical or model-based factor analysis

- (Latent) factors in graphical models
- Related/intersecting clusters

West, 2003, Bayesian Statistics 7

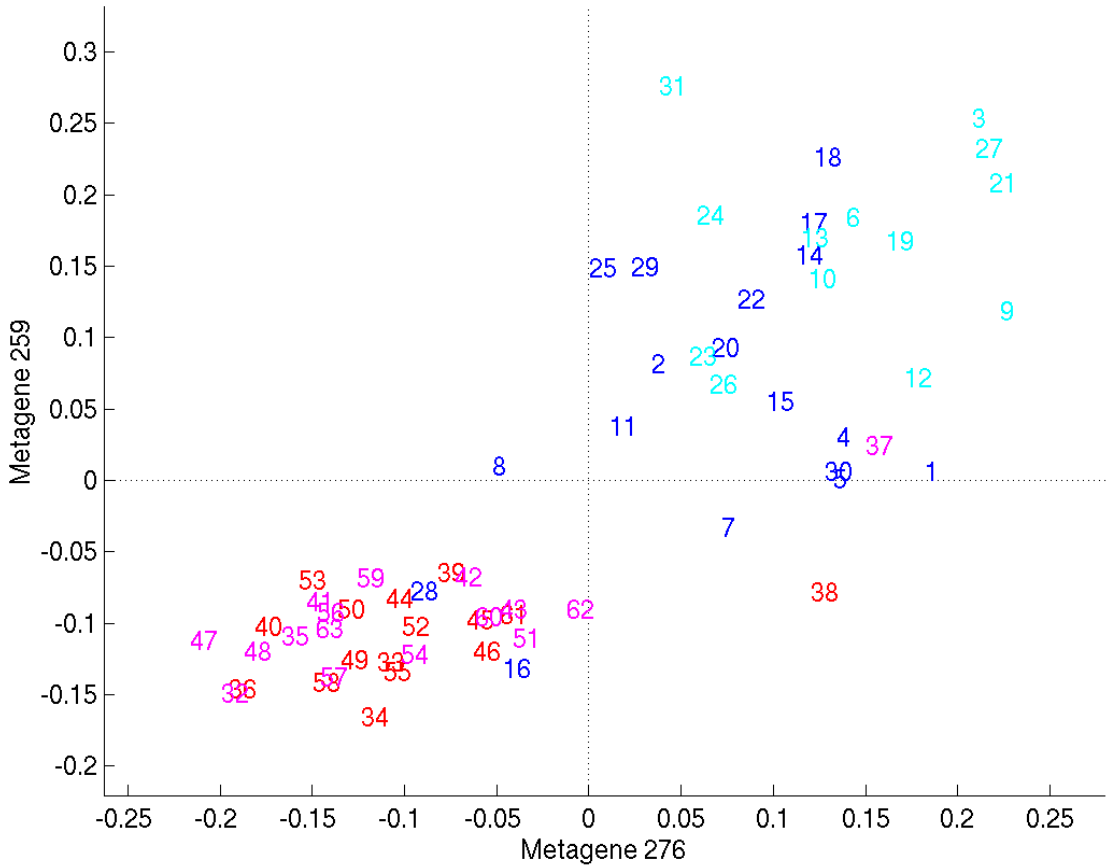


Reclustering, Covariance selection & Graphical Models – software tools ..

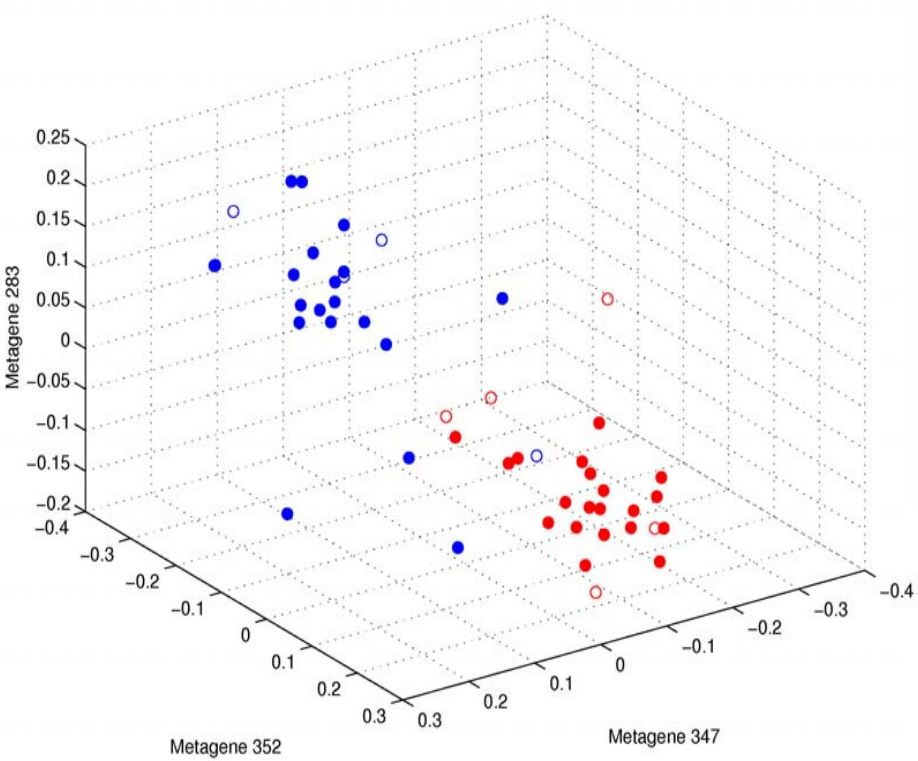
MetageneCreator

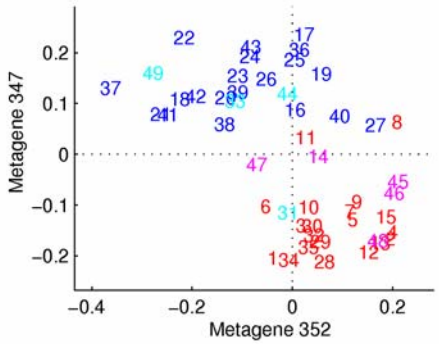
(Dobra, West, 2004, forthcoming)

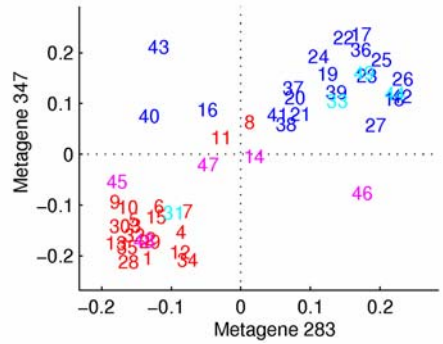
Atherosclerosis Metagenes
- Disease Susceptibility -

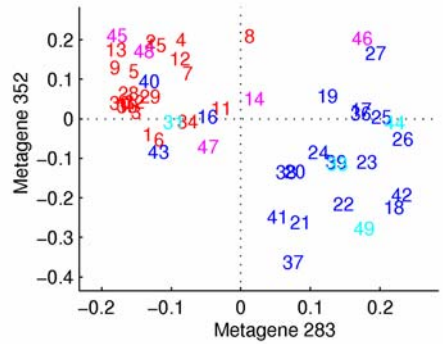


ER Metagenes in Breast Cancers





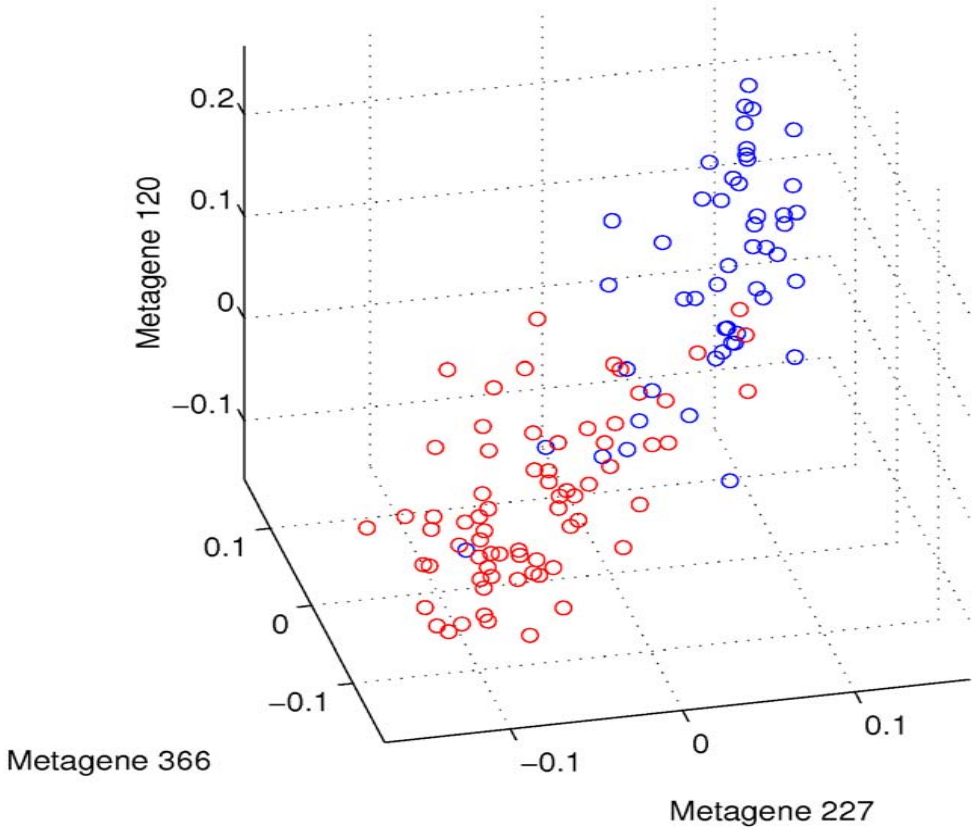




79 genes: ER regulated: Tff1, bcl-1,2, ..
27 genes – ESR1
39 genes – other co-regulated, androgens, ..

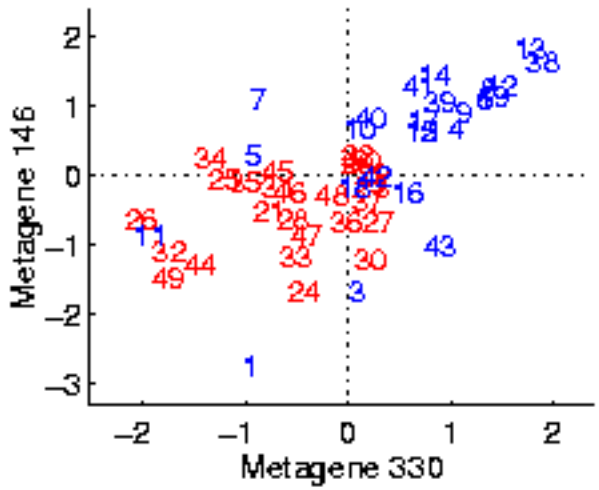
49 samples - Huang et al, Duke PNAS 2001

ER Metagenes in Breast Cancers



158 samples - Pittman et al, Duke/KFSYS PNAS 2004

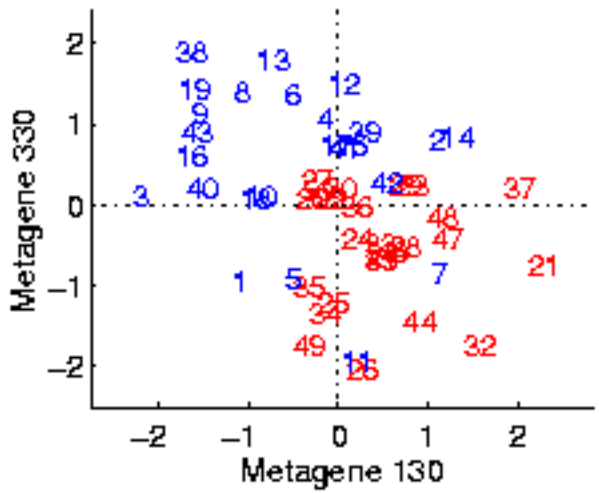
Lymph Node Metastasis Metagenes



Classification Trees Models:

- *Lymph node outcomes*
- *Recurrence outcomes*

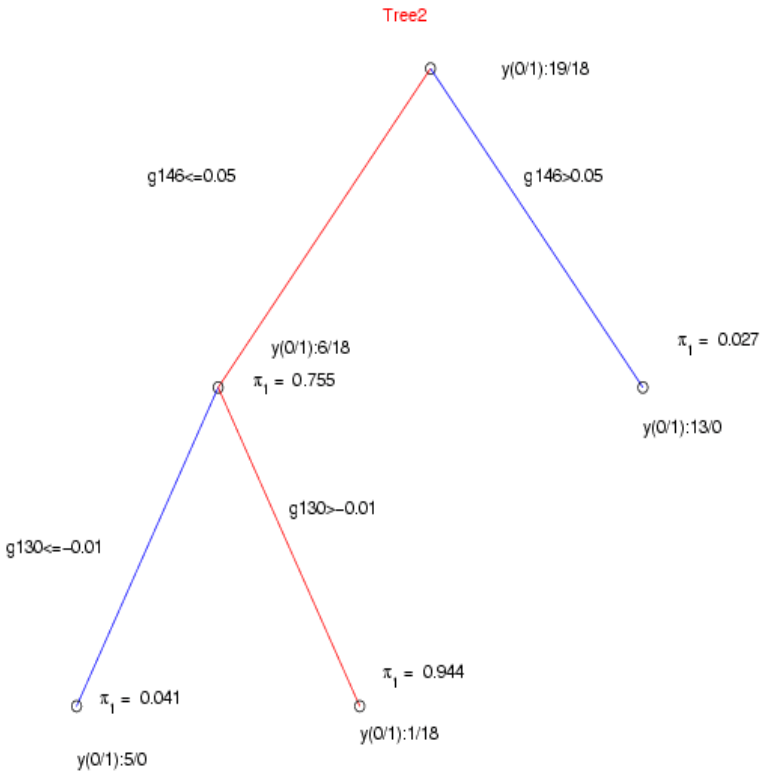
- Cross-validation assessment
- Predictive classification



Huang et al, Lancet 2003

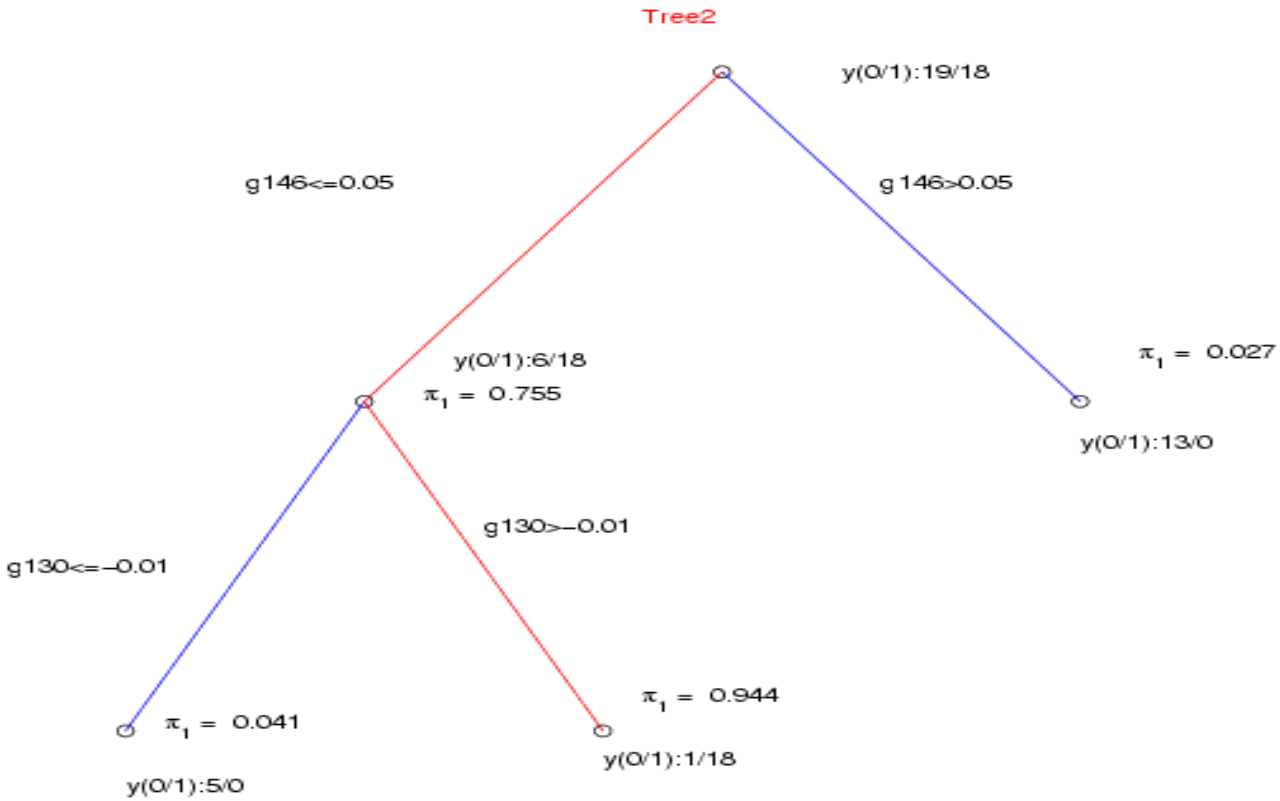
Binary Tree Models

Tree Models: Classification via Recursive Partitioning



- Non-linear: interactions of covariates
- "Natural" in contexts of clinical prognosis
- Many trees:
  Model & prediction uncertainty
- Prospective or
  retrospective studies

Binary Outcomes: *Y=0/1*
- Retrospective Sampling (Case-Control studies) -



$$\pi = \Pr(Y = 1 | x_1 \le \tau_1, \ldots, x_k \le \tau_k)$$

Binary Outcomes: Prospective Inference from Retrospective Model

$$\pi = \Pr(Y = 1 | x_1 \le \tau_1, \ldots, x_k \le \tau_k)$$

$$a_{1y} = \Pr(x_1 \le \tau_1 | Y = y), \qquad y = 0,1$$

$$a_{2y} = \Pr(x_2 \le \tau_2 | x_1 \le \tau_1, Y = y), \qquad y = 0,1$$

$$\frac{\pi}{1-\pi} = \frac{\Pr(Y=1)}{\Pr(Y=0)} \frac{a_{1,1} a_{2,1}}{a_{1,0} a_{2,0}}$$

Binary Outcomes:
- Retrospective Model -

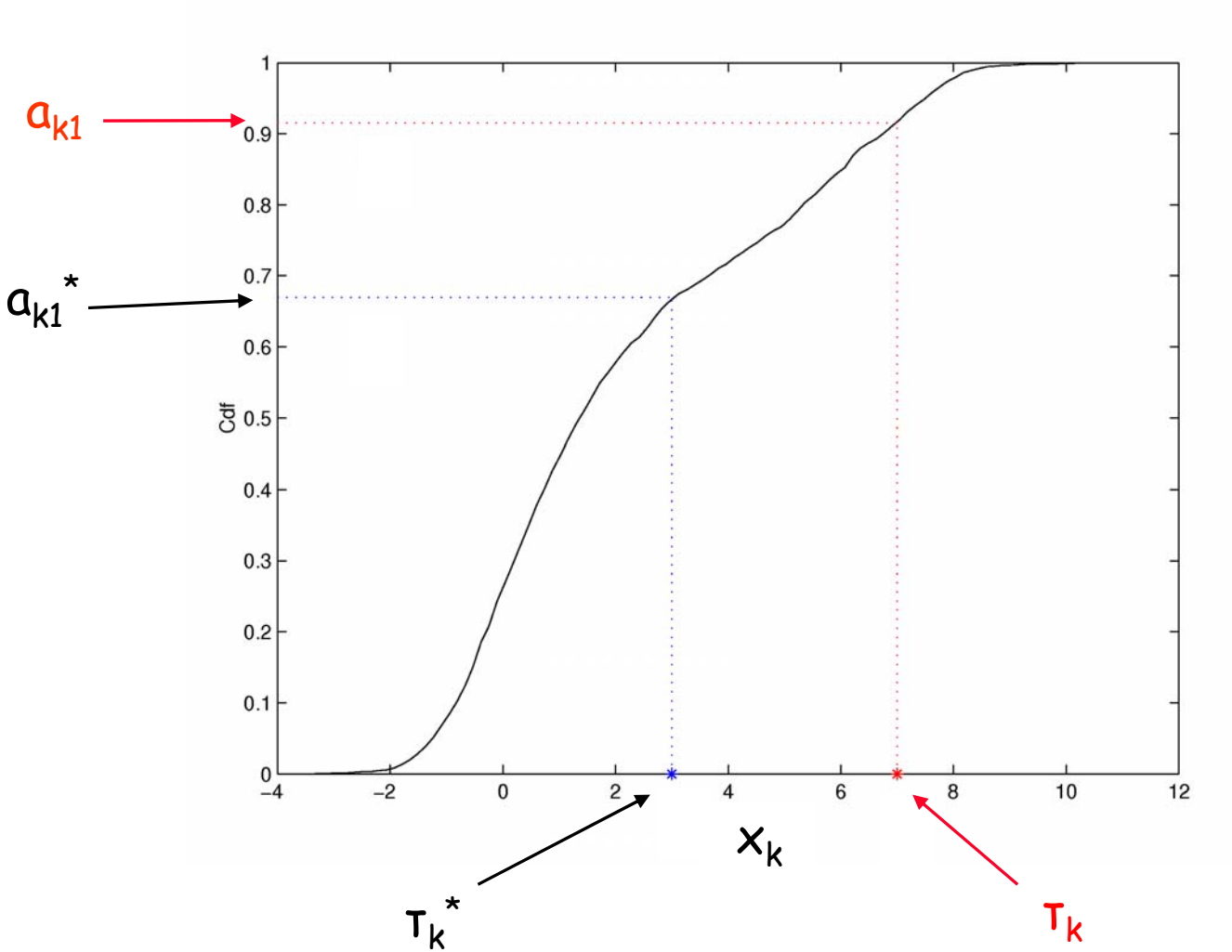
- Model conditional CDFs for predictors

$$F(x_k | x_1 \leq \tau_1, \ldots, x_k \leq \tau_k, Y = y)$$

- Want consistency as thresholds vary
- Nonparametric Bayes: Dirichlet model
- Modelling in x space – joint structure
- Implies Beta priors on a_{ky}

Dirichlet Process: Beta Margins

$F(x_k|-,y=1)$



Assessing Predictor:Threshold Pairs for Association with Binary Outcome

Any predictor (metagene) x, threshold τ

Sample arranged in 2x2 table

Retrospective: column totals fixed

Two Bernoulli sequences: columns with "success" probabilities a_0 & a_1

Beta priors: Conjugate

	$y=0$	$y=1$	
$X \le \tau$	n_{00}	n_{01}	N_0
$X > \tau$	n_{10}	n_{11}	N_1
	M_0	M_1	

Assess/test: Bayes' factor for $a_0 \neq a_1$ *versus* $a_0 = a_1$

Nonlinear association measure: Function of threshold

(Pittman et al, Biostatistics 2004)

Growing & Using Binary Trees

Node splits:

- assess candidate predictor:threshold pairs
- 2x2 table: conservative Bayesian tests

Multiple trees:

- Multiple splits at any node

Inference: 1 tree

- beta posteriors for a_{iy}
- simulate: impute Pr*(Y=1|leaf)*

Prediction with multiple trees:

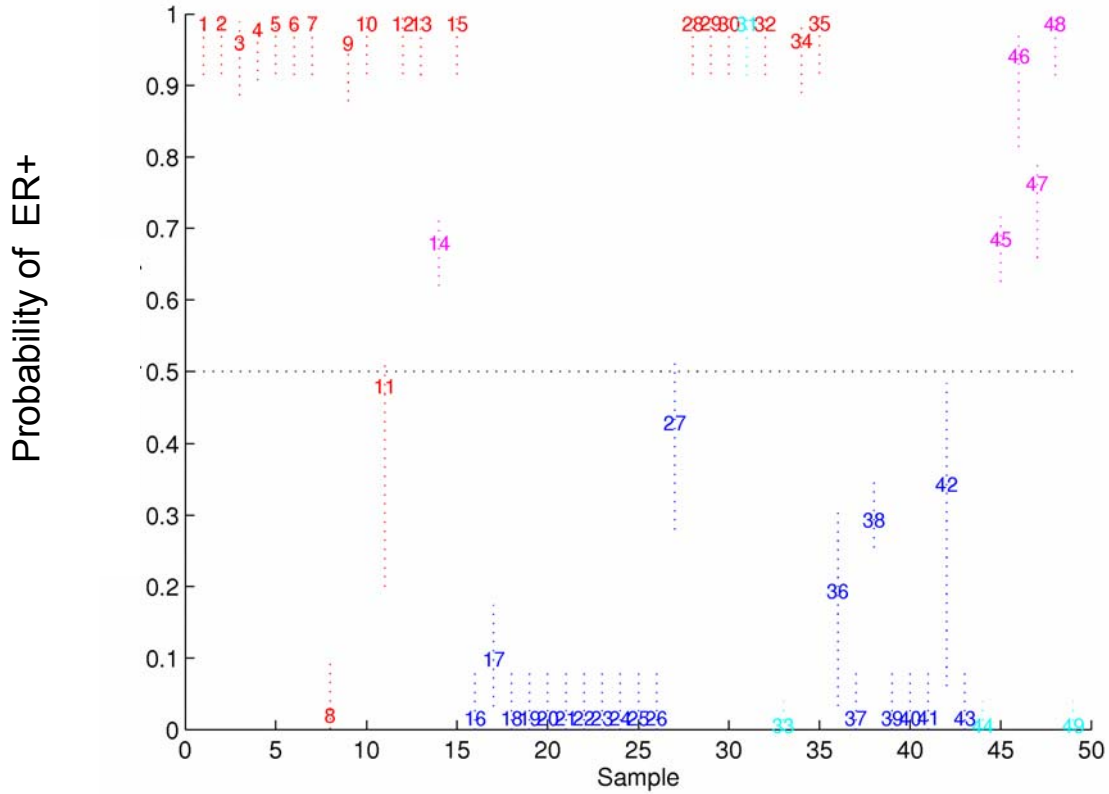
- Average across trees
- Model uncertainty
- "Smoothing" partitions

(Pittman et al, Biostatistics 2004)

Predicting ER Status With Metagene Trees

Duke PNAS 2001 data –

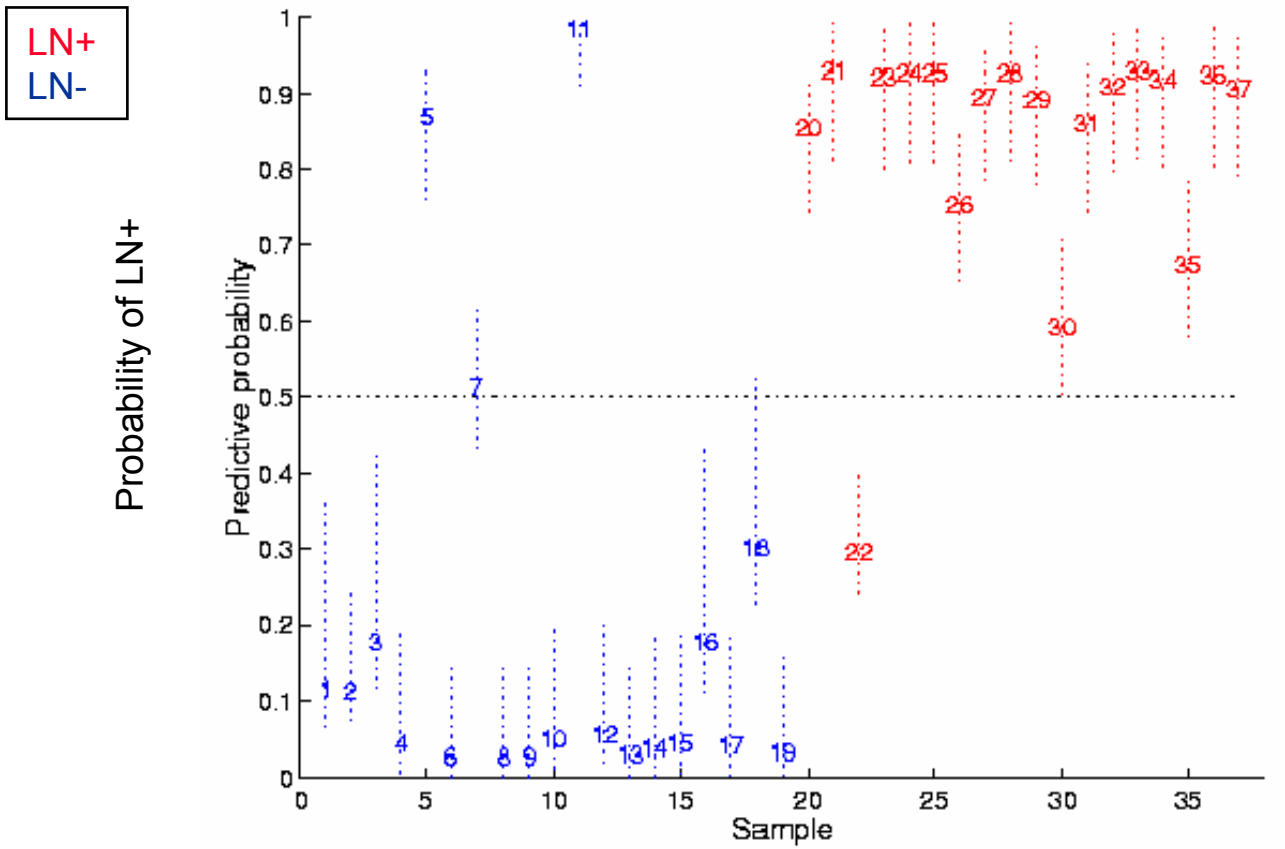
Out-of-sample cross validation



(Pittman et al, Biostatistics 2004)

Predicting Lymph Node Status With Metagenes

Out-of-sample cross validation



Huang et al, Lancet 2003

Gene Identification

Implicated metagenes – gene subsets

Genes correlated with key metagenes

Breast Cancer – nodal metastasis:

- Interferon pathway/inducible gene subset
- Interferons mediate anti-tumour response

Evidence of disfunction of normal anti-tumour response?

- BRCA1, p27

Duke PNAS 2004 Breast Cancer Data

Examples (see Matlab code explorations)

Explore ER (0/1) regression using metagene predictors

Explore tree models for cancer recurrence

Look at nonlinear association with status - for each metagene across the expression (threshold) range

Fit forests of binary trees, look at 'top' trees, predictions of samples held-out

Tree Models for Survival Analysis

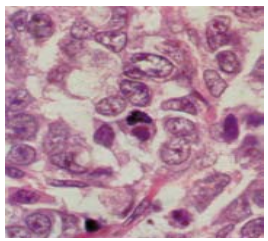
Prospective Tree Models

- Recurrence in Breast Cancer -

Survival tree models – Prospective

- Clinical + metagene predictors
- Survival distributions in subgroups

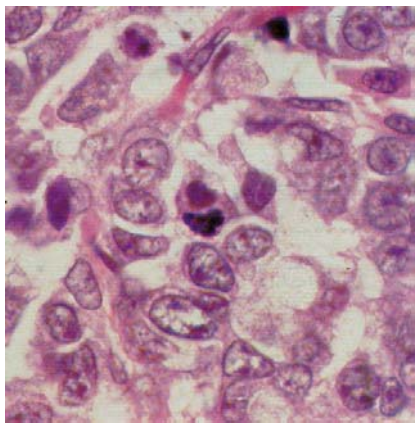
Challenge: Dissect Heterogeneity



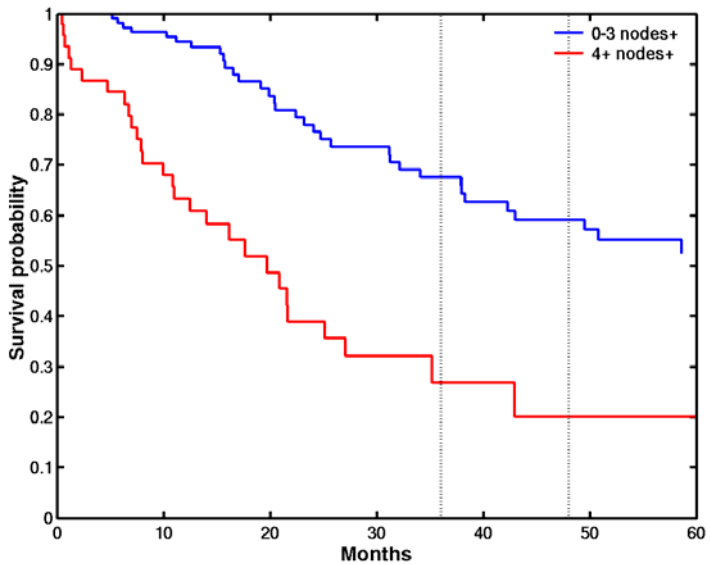
Cancer is fundamentally a **heterogeneous** disease

Understanding and dissecting the heterogeneity is key to prognosis and treatment

Clinical Predictors: "Low Resolution" Phenotypes



- Lymph node involvement



Refining Phenotype: *Adding* Molecular Information

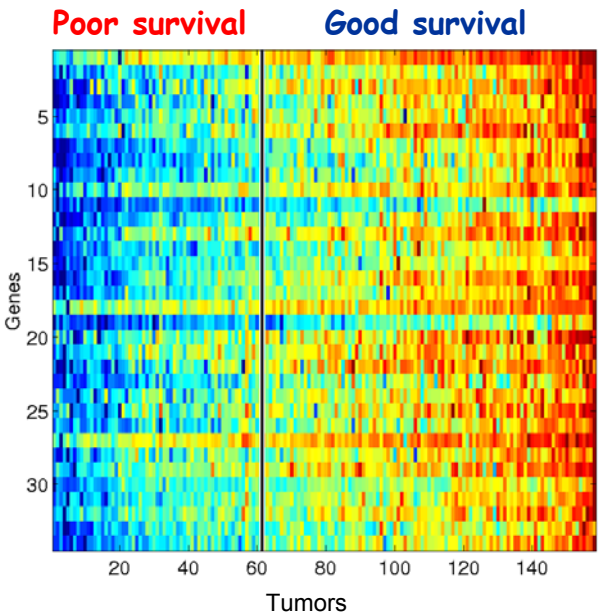
Analysis of the tumor

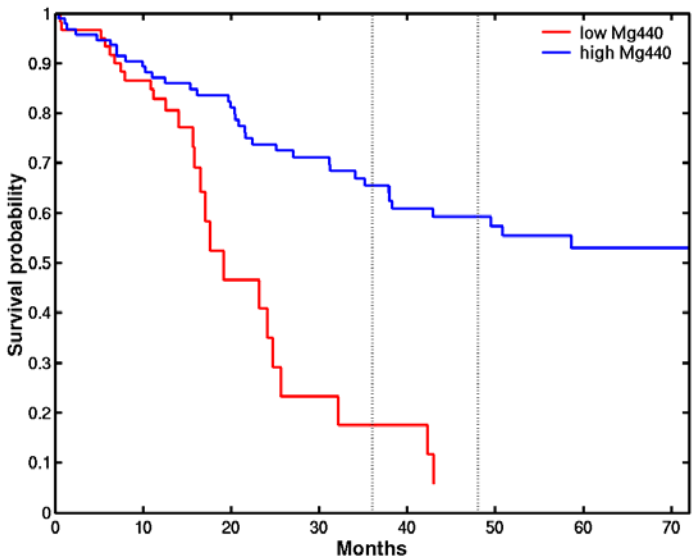
- Gene expression
- DNA copy status (CGH)
- DNA methylation
- Protein profiles
- Metabolic profiles

Analysis of the 'host'

- Genotypes
- Serum protein profiles
- Serum metabolic profiles
- Serum gene expression profiles?

Metagenes are Predictive of Breast Cancer Recurrence

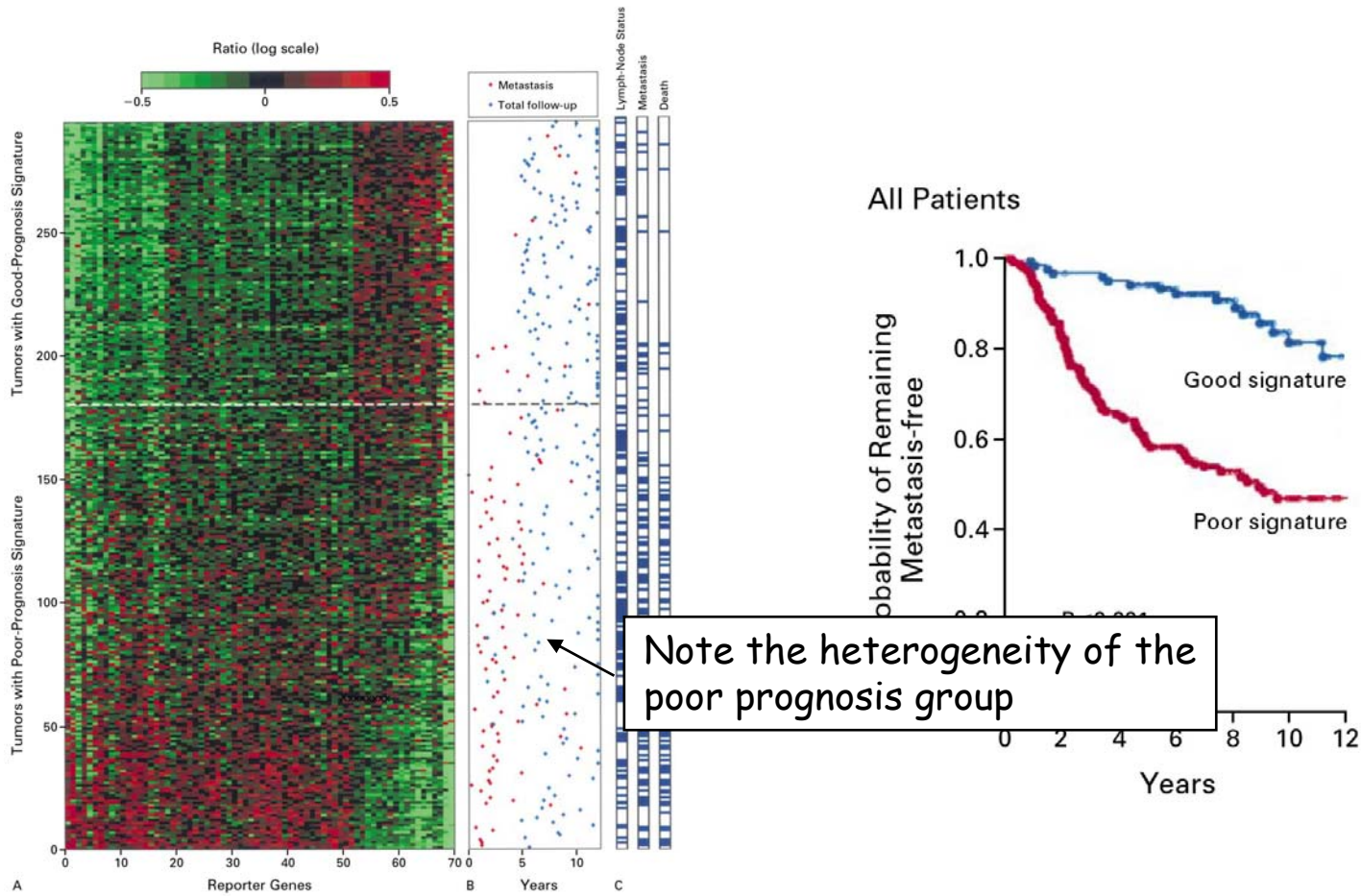




It's genomic, but 'group' prediction, yet another biomarker

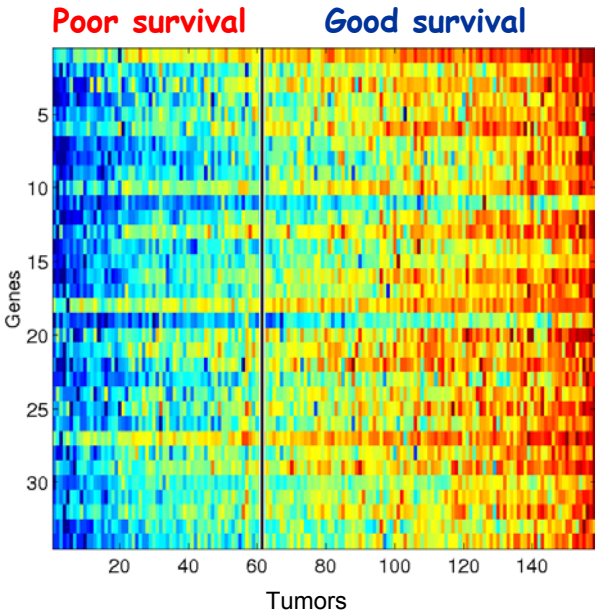
A 70 Gene Predictor

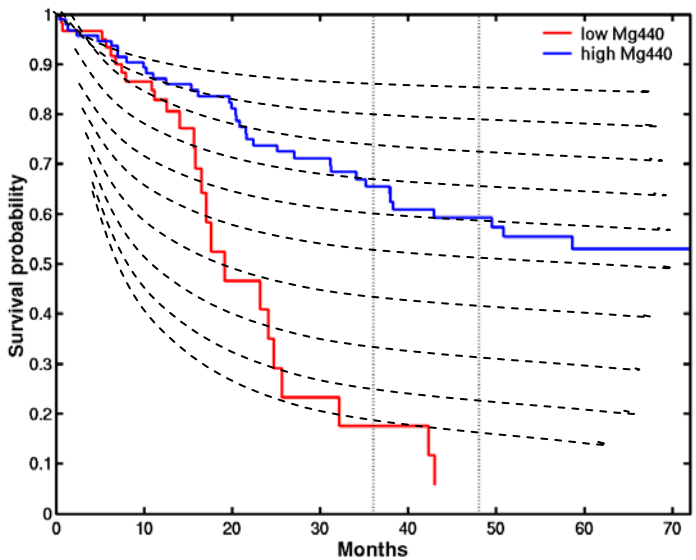
van de Vijver et al., N. Engl. J. Med. 347: 1999 (2002)



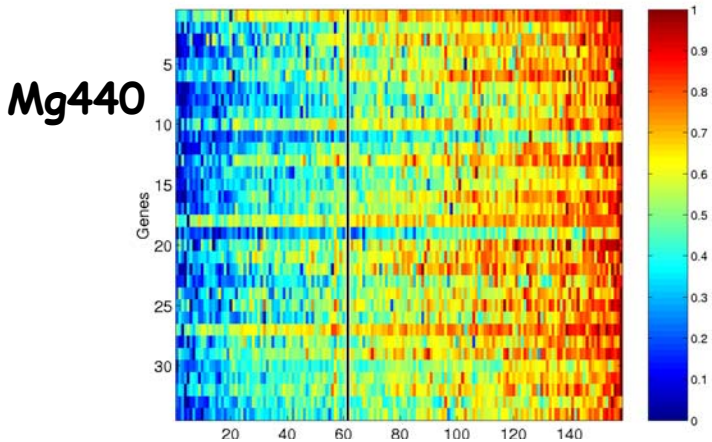
Moving Gene Expression Profiling to the Clinic

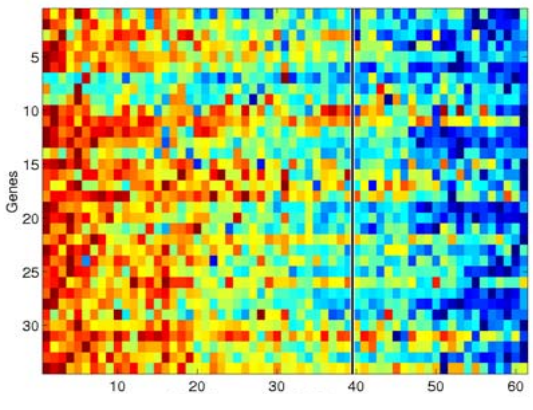
Personalised Prediction
- Dissecting Heterogeneity -

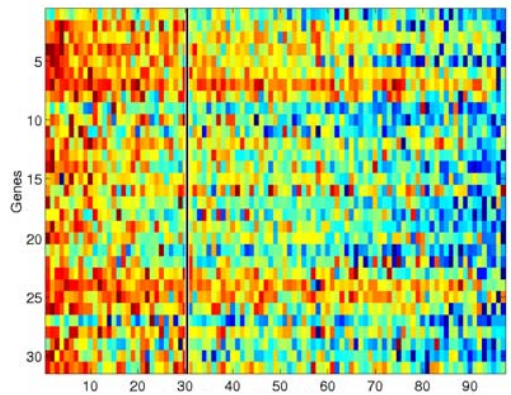




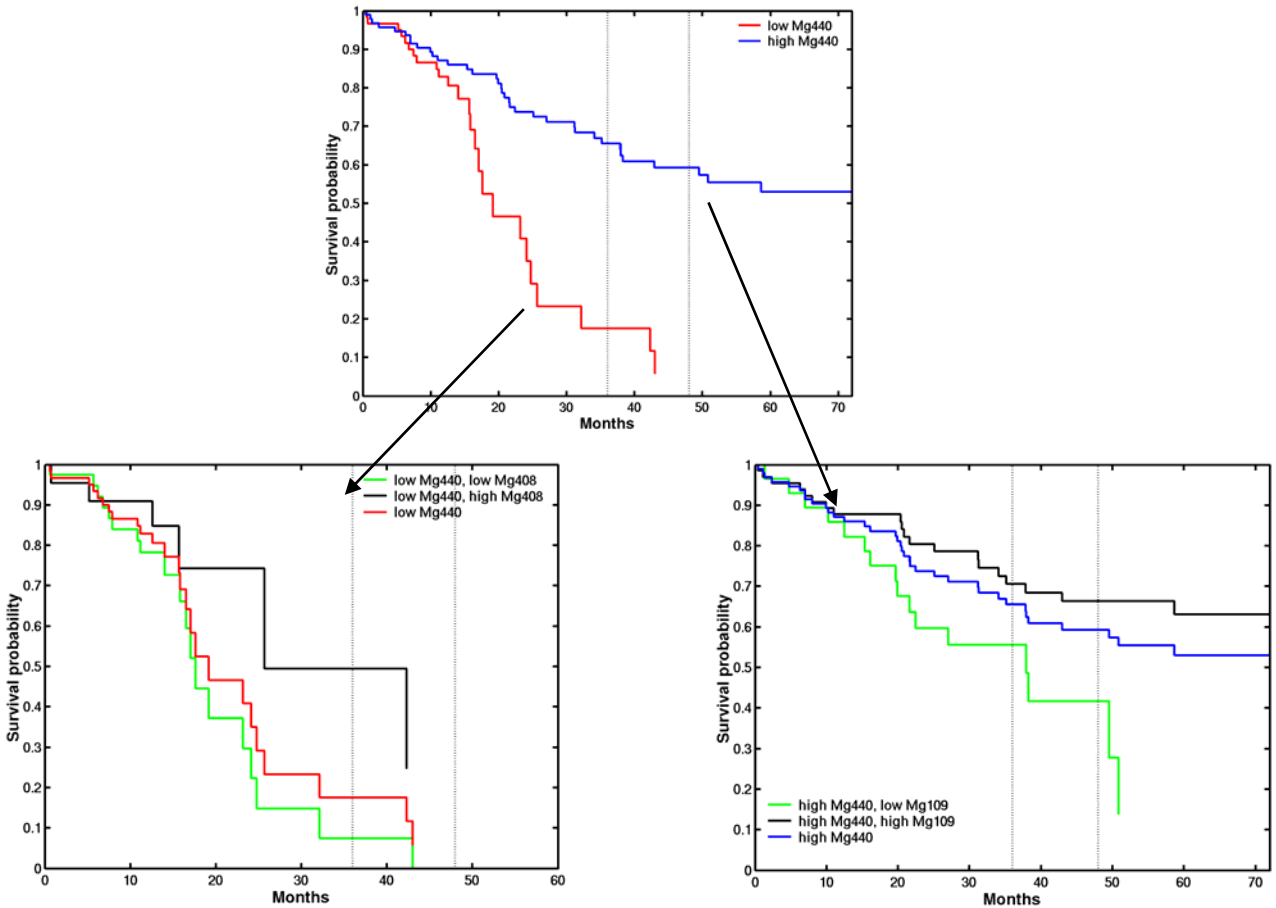
Dissecting Heterogeneity Using Multiple Metagenes



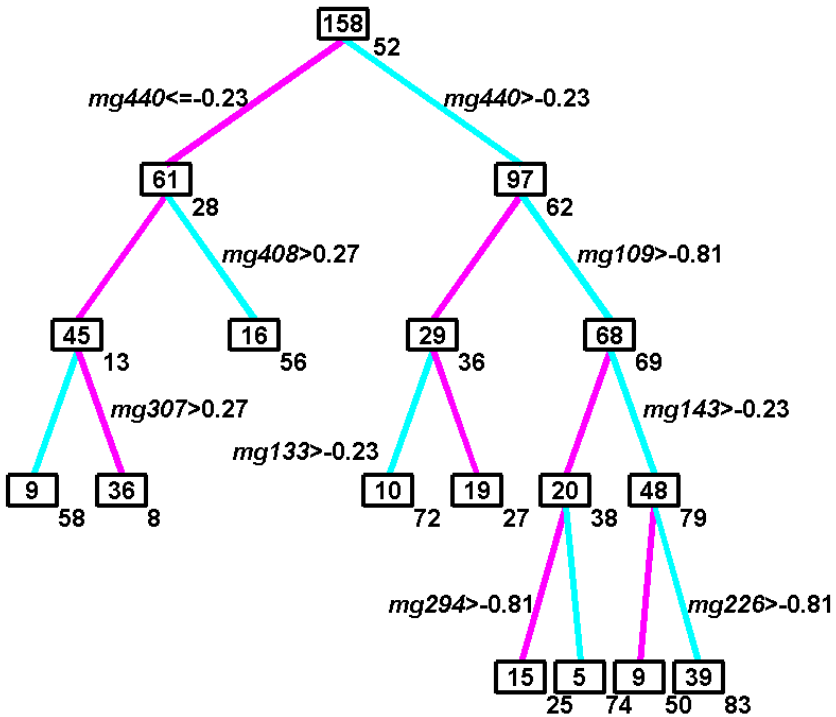




Subtyping by Multiple Metagenes
- Can Improve Risk Discrimination -



Predictive Survival Tree Models



(Pittman et al, PNAS 2004)

Leaf j: $Y \sim \text{Weib}(\mu_j, \alpha)$

Hierarchical (gamma) priors:

$p(\mu_j | m) = Ga(\mu_j | c, c/m_j)$

- *mean m_j has hyperprior*
- *m_j "similar" for sibling leaves*
- *shrinkage/data sharing*

Marginal likelihood of tree:

$p(Y | Tree, \alpha) = \Pi_{\text{leaves } j}\, p(Y_j, \alpha)$

Growing & Using Survival Trees

Node splits:

- assess candidate predictor:threshold pairs
- 2 Weibulls vs 1?
- Bayes' test

Multiple trees:

- Multiple splits at any node

Inference: 1 tree

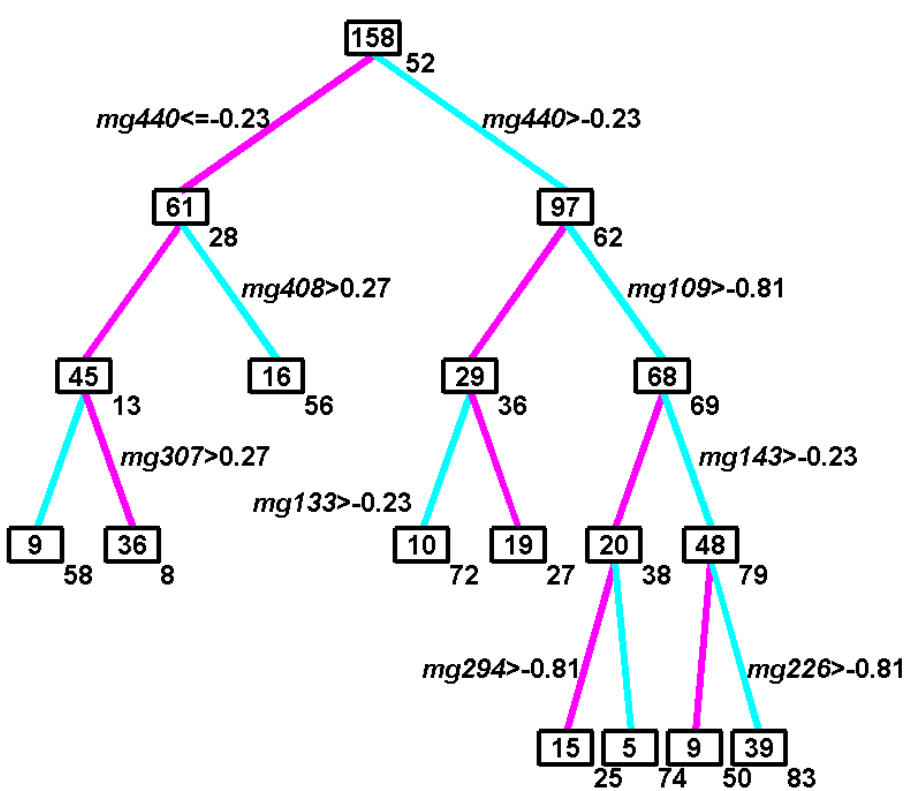
- Empirical Bayes for m_j, a
- Pareto predictions in leaf

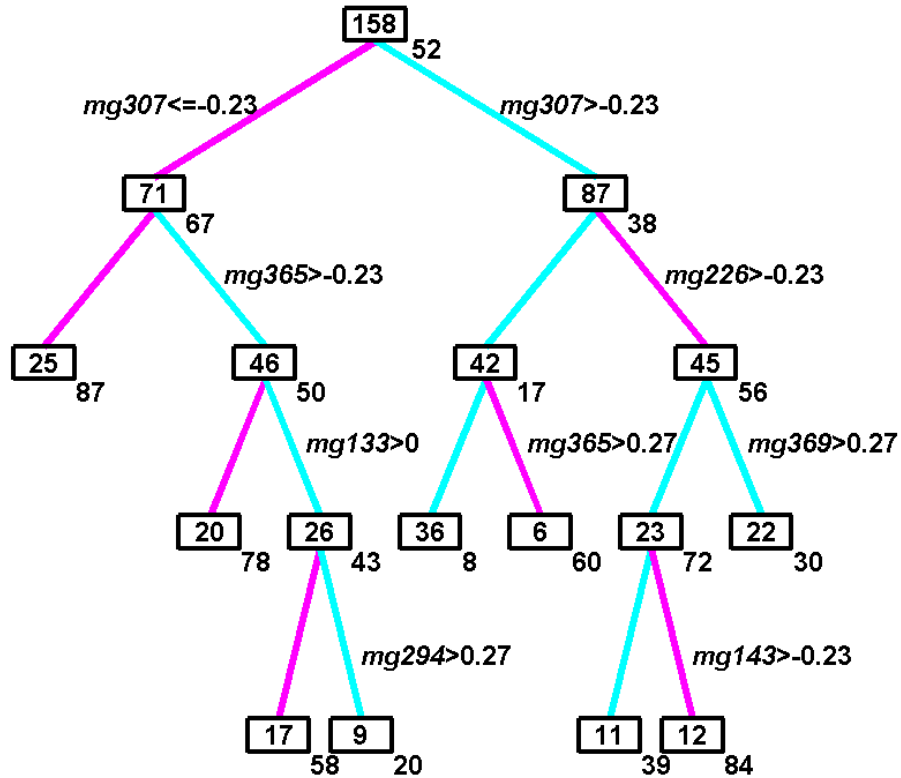
Prediction with multiple trees:

- Average across trees
- Model uncertainty
- Simulation

(Pittman et al, PNAS 2004)

Forests of Metagene Trees

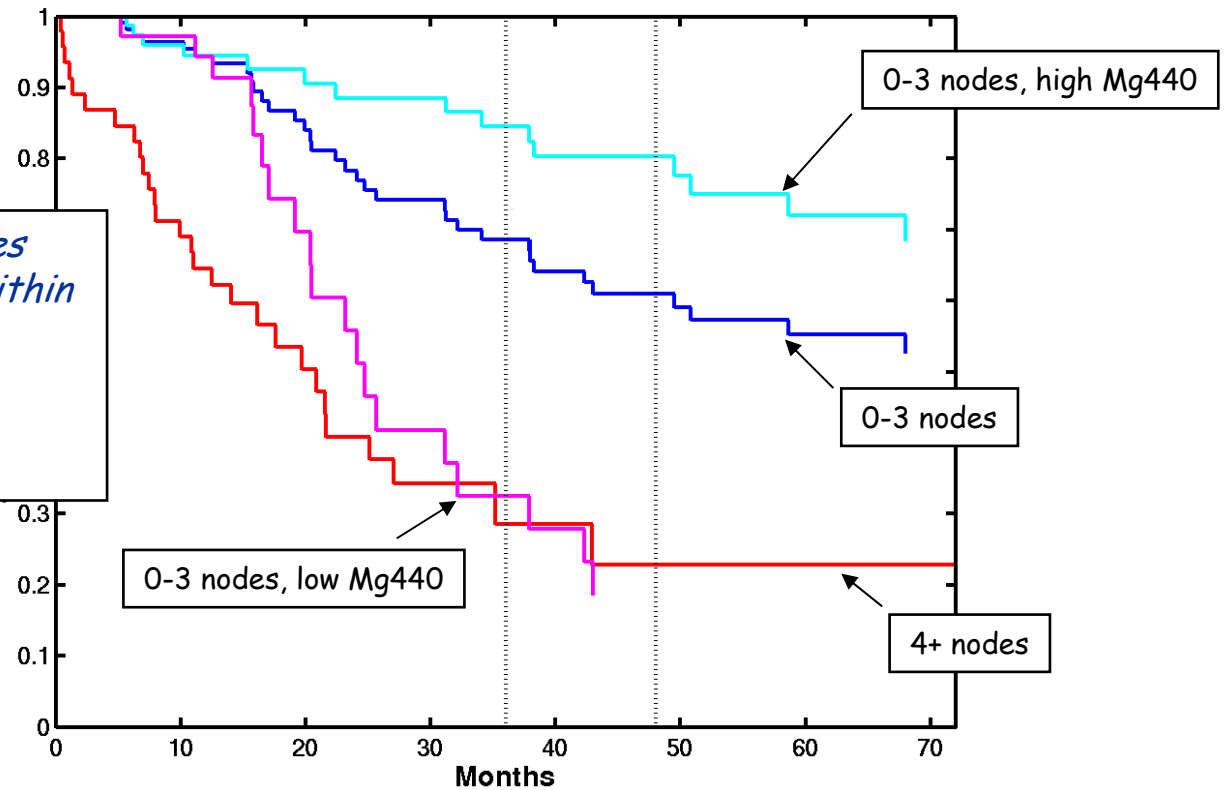




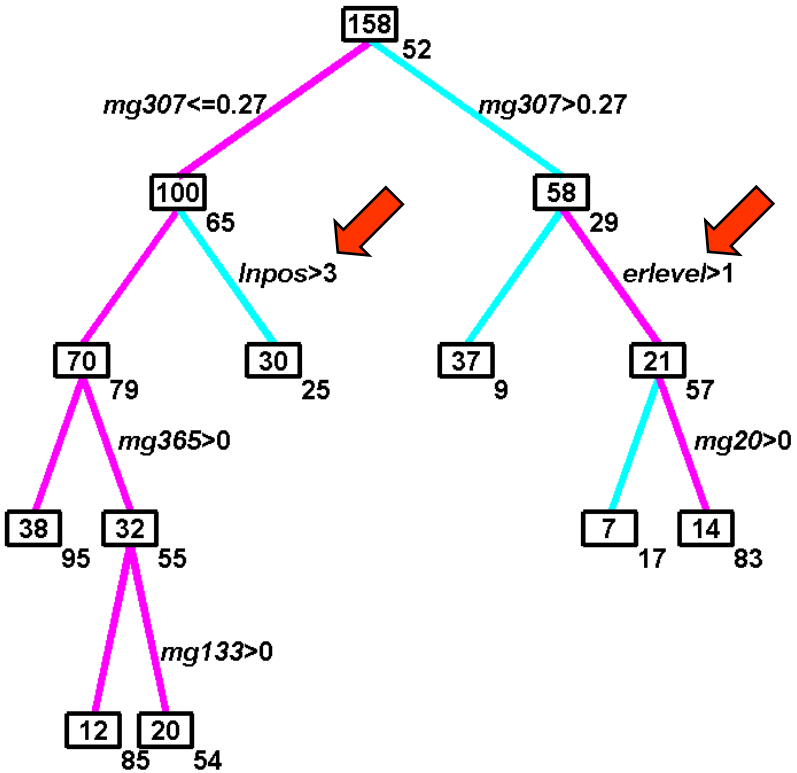
Clinico-Genomics for Modifying and Refining Risk Stratification

Genomic data identifies higher risk patients within clinical low risk group

Treatment decisions? Clinical trials?



Forests of Clinico-Genomic Trees



Select from full sets of clinical and genomics potential predictor variables

multiple trees

variable subset combinations - co-occurrence

Personalised Prognosis: Prediction of Individual Outcomes

Patient 158:

Tumor size 1.5cm

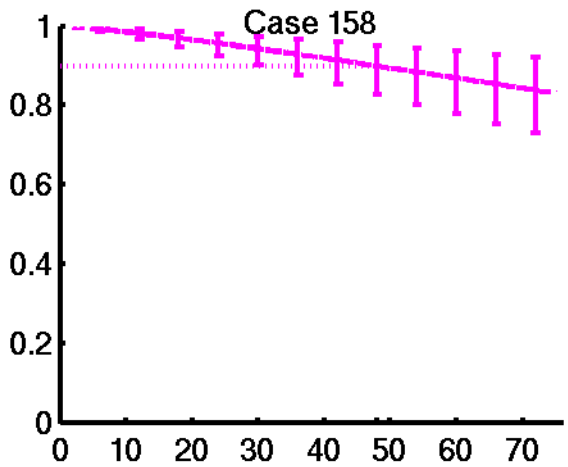
LN=1

ER-

Mg440= 0.25

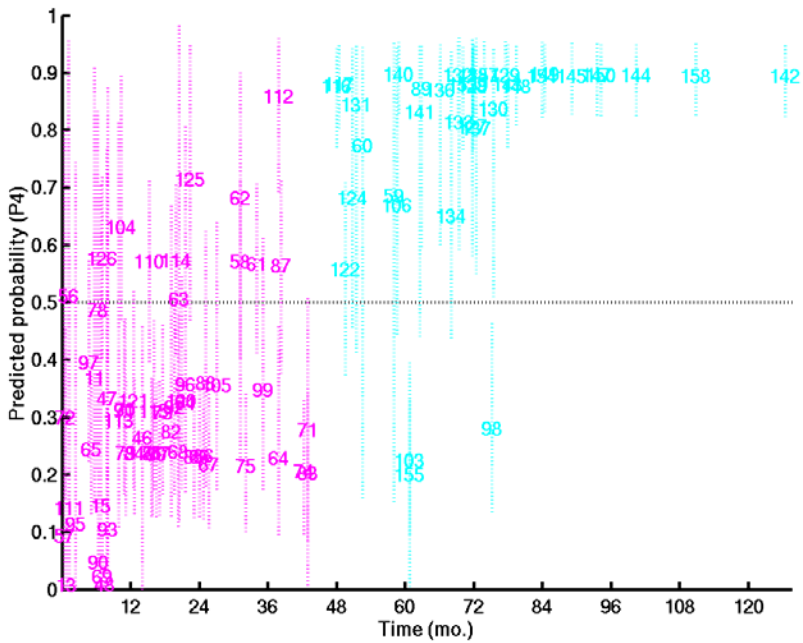
Mg20= -.04

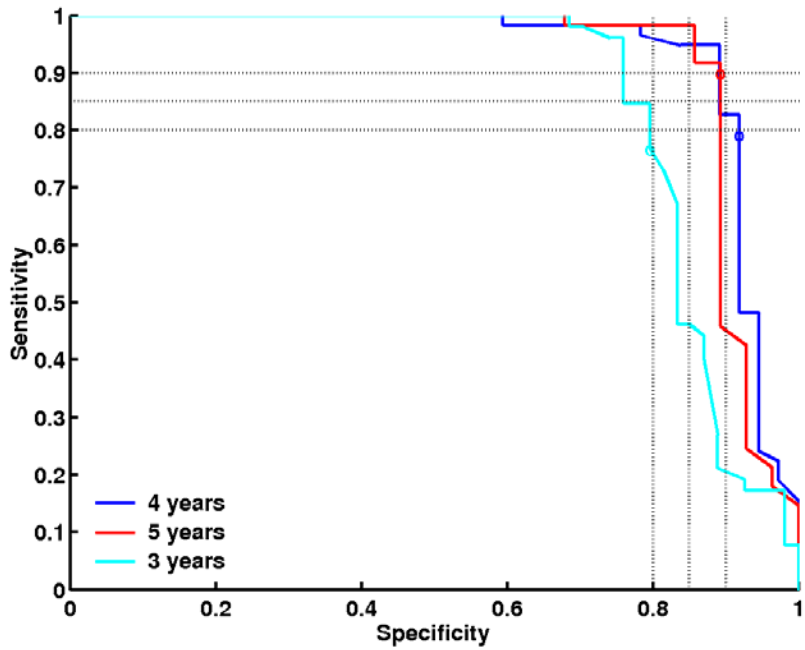
… etc



Predictions: Uncertainty & Assessment

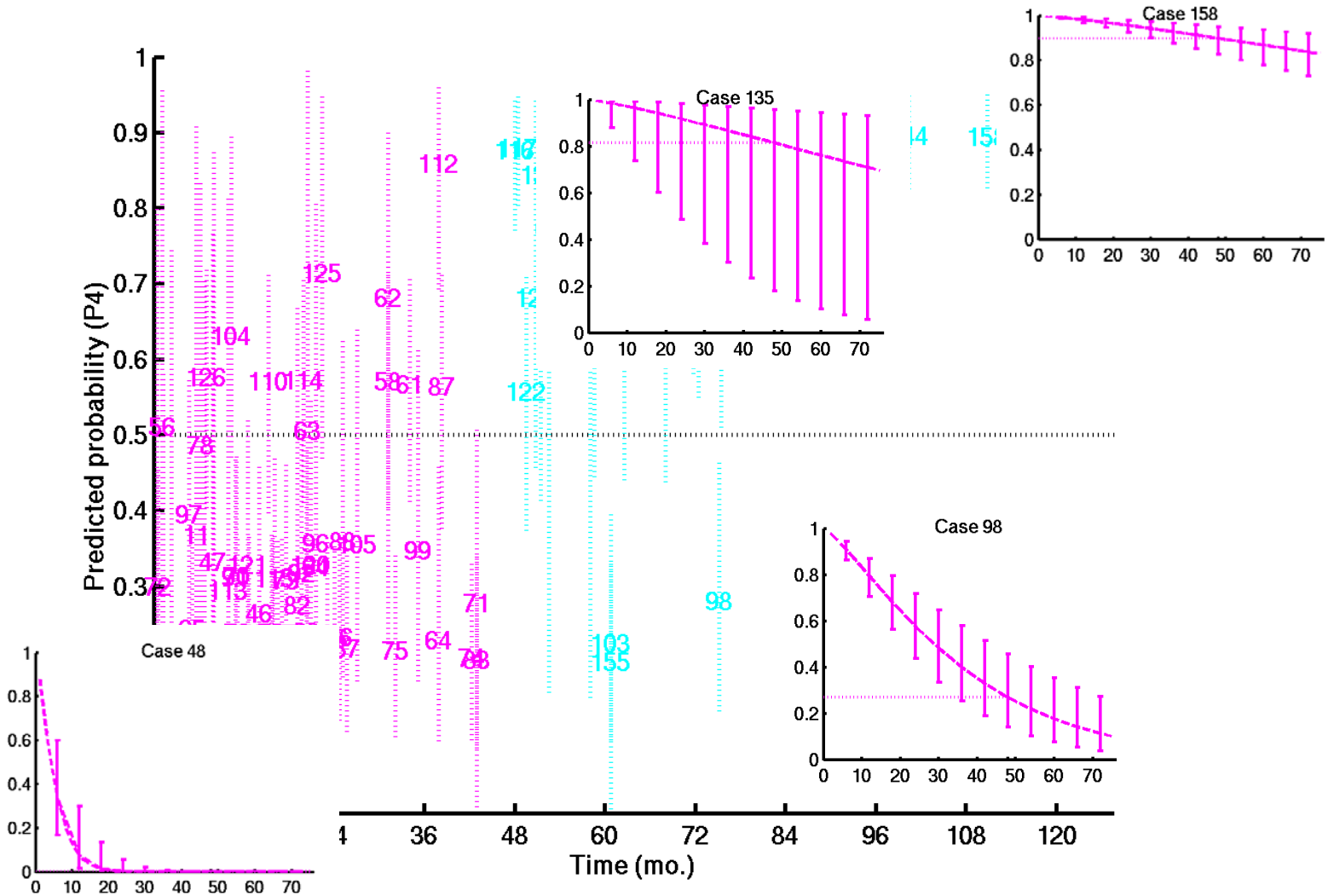
Out-of-sample cross validation



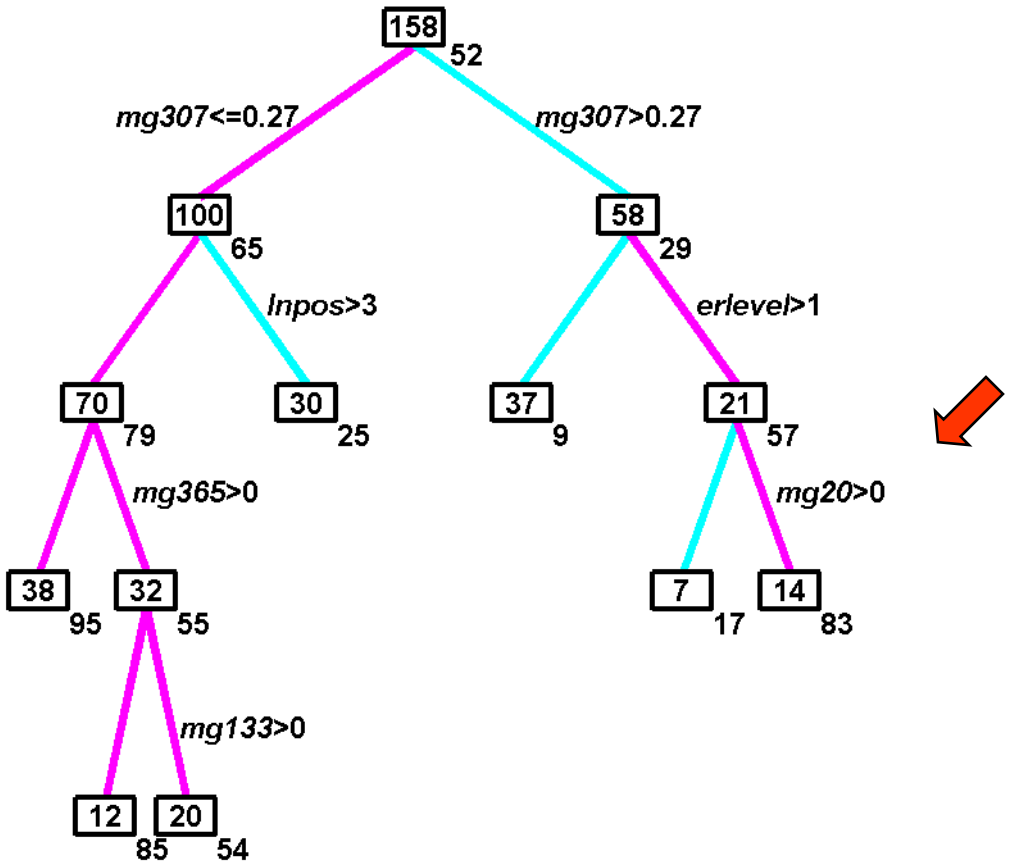


Recur+
Recur-

Personalised Prognosis



Biology behind Metagenes



Erb-b2/Her-2 nu metagene

Gene Identification
- Biological Signatures -

Erb-B2/Her2-nu metagene

Several ER metagenes

Recurrence metagenes: Myc oncogene, E2Fs

Lymph node & recurrence predictors

- metagenes predictive of lymph node status

… "surrogate" for node counts …

- BRCA1, p27

- Interferon inducible genes

Some Current Related Studies

- Validation studies
- Combining data - Multi-center studies
- Ethnic variation in genomic patterns
- Treatment trials - Improved risk stratification
- Treatment response predictors

Patterns related to genes involved in cell proliferation -

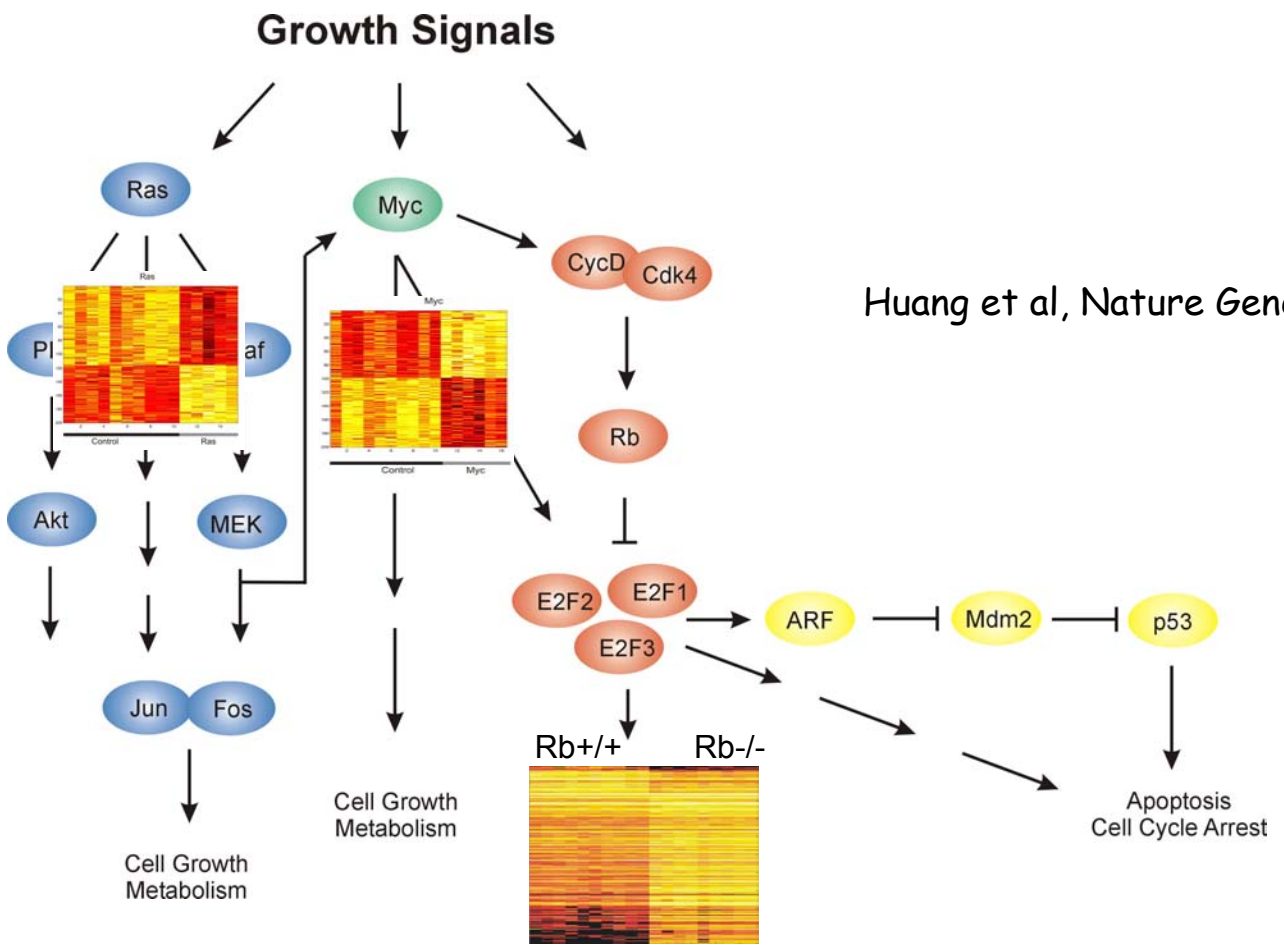
> improved biological metagene characterisations

> refined "subtyping":

Myc, Ras, Rb/E2F, …, metagenes

(Nature Genetics 2003, Cancer Research 2003)

Signaling Pathways That Control Cell Proliferation



Huang et al, Nature Genetics 2003

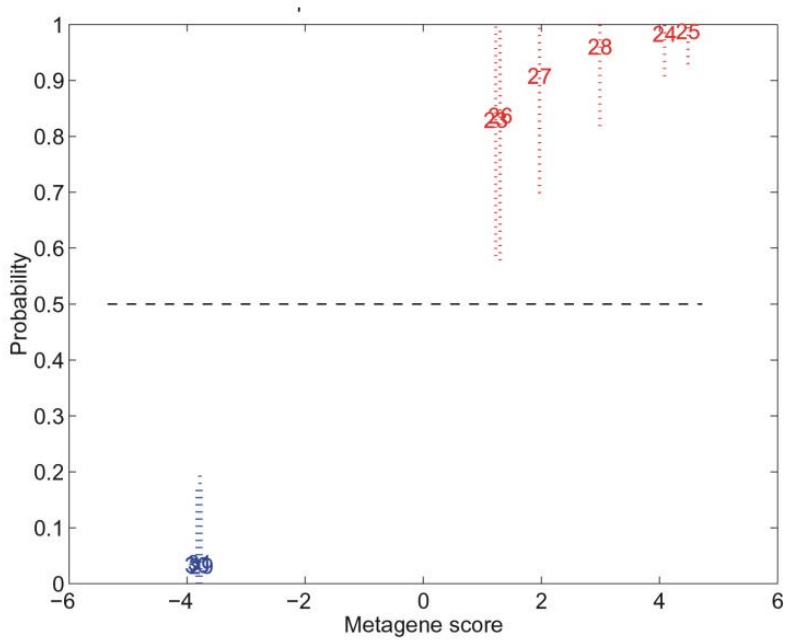
An Rb Metagene Predicts Loss of Rb Function in Tumors

Human

Rb+/- ⟶ Rb-/- | Retinoblastoma

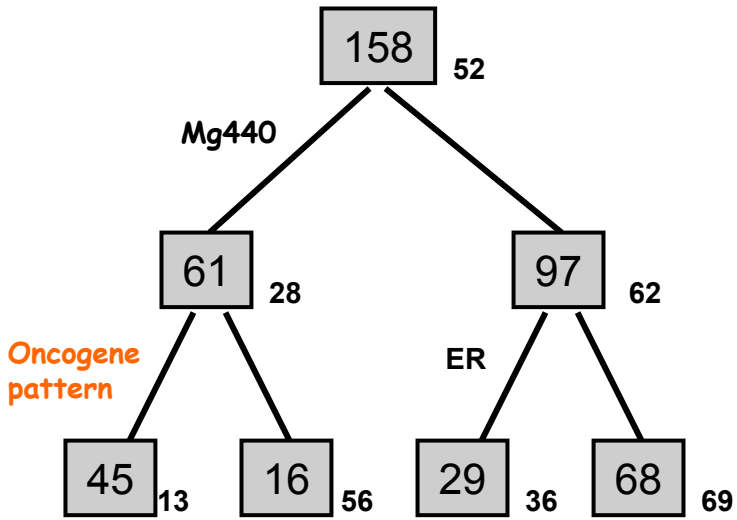
Mouse

Rb+/- ⟶ Rb-/- | Pituitary tumors, Thyroid tumors



Black et al, Cancer Research 2003

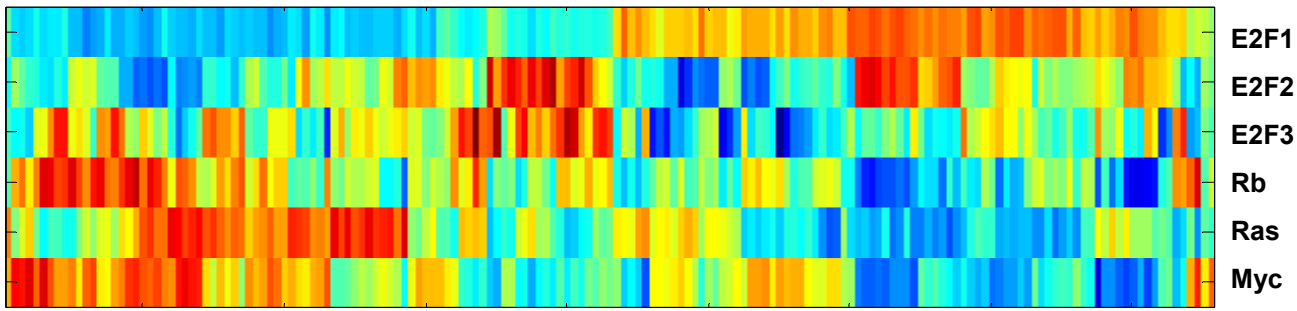
Potential for Improved Characterisation and Prediction



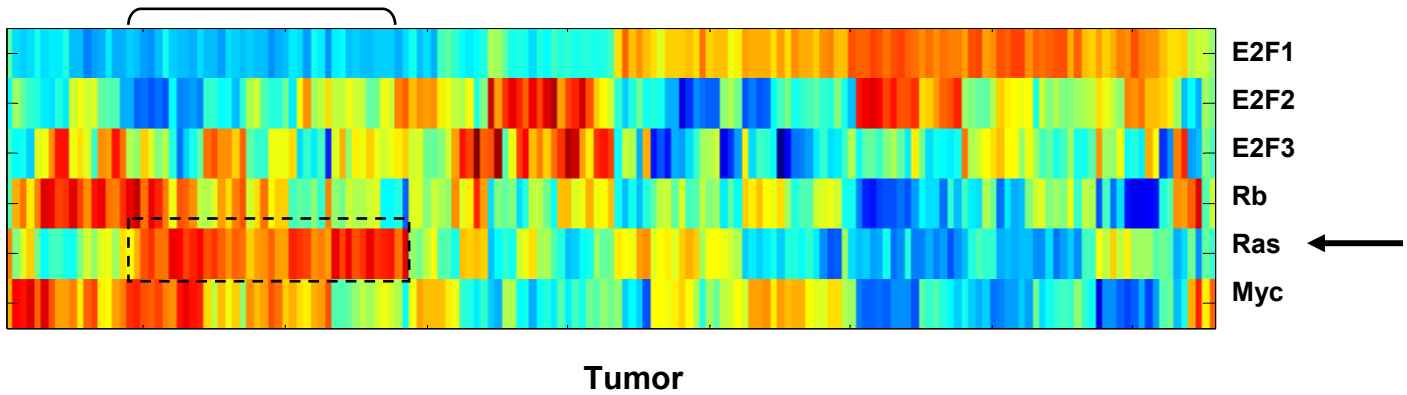
… reflecting multiple interacting aspects of tumour genomics

… and also potential to aid in identification of therapeutic targets

Pathway Predictions and MetaPathways



Making Use of Pathway Predictions



Using Pathway Information to Guide Treatment?

Metastatic Breast Cancer Patients

↓

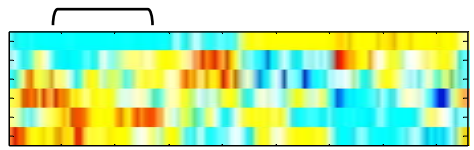
Biopsy of metastatic tumour

↓

Treat with Drug A

Response ↙ ↘ Non-response

Pathway Prediction



↓

Pathway Specific Drug

Current Foci
- Statistics & Computation -

- Model refinements

  smooth partition trees; aggregation of predictors

- Stochastic search

  MCMC analysis, and search & annealing to explore space of candidate trees, and posterior over trees

- Cluster/distributed computing
- Defining and evaluating patterns

  improved metagene models: factor methods, large-scale graphical models, association network

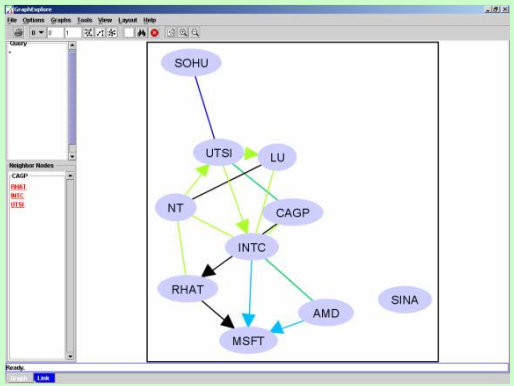
http://graphexplore.cgt.duke.edu/

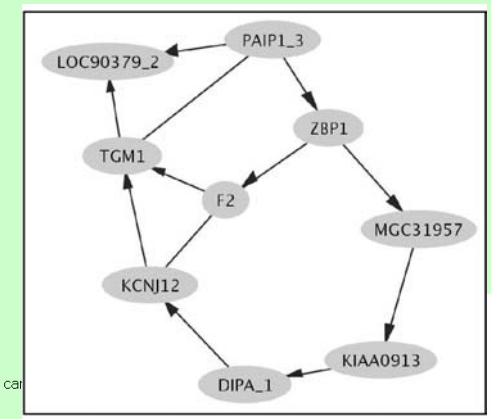
GraphExplore

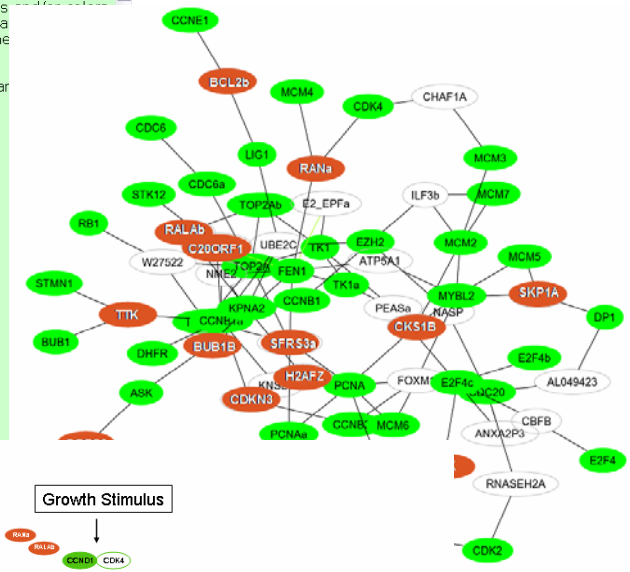
- News
- Overview
- Tutorial
- Examples
- Download
- Web Start
- About CAGP
- Contact us

A very important and attractive feature of **GraphExplore** is that it lets you choose specific colors and shapes for each object in the network. You can access these functions from the menu by going to **Options** > **Node** > **Color** or **Options** > **Node** > **Shape**. Moreover, **GraphExplore** lets you assign the same shapes to groups/clusters of objects. These groups can be different than the clusters you loaded with a new project. Therefore objects with the same shape ca clustering, objects with the same color can identify another clustering and both of these clusterings can be different than the clusters loaded with the degree of flexibility is necessary to create meaningful displays of objects having different functions.

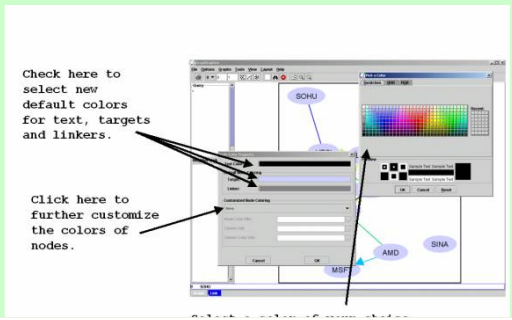
You can begin by typing "*" in the **Query** box and select **Graphs** > **Subgraph** to create a display of the entire network. Remark that all the objects ar have the same default color (light blue).

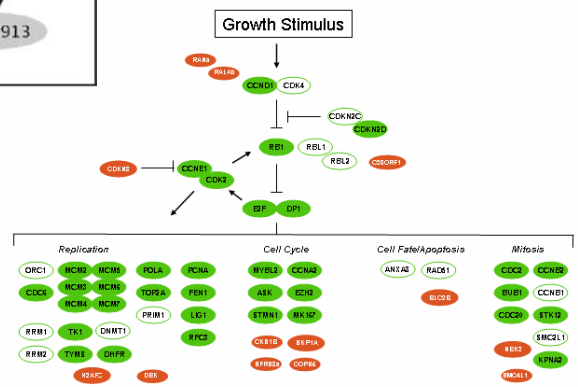






Bring up the colors dialog box by going to **Options** > **Node** > **Color**. You ca format.





Large-scale graphical model search and evaluation

Inference on large, sparse inverse covariance matrix

(Dobra et al, JMVA 2004)

Duke University

Joseph Nevins

Erich Huang
Holly Dressman
Penni Black
Guang Yao

Jennifer Pitman
Adrian Dobra
Quanli Wang
Chris Hans
Carlos Carvalho

Molecular Genetics
Cancer Genomics

Statistics

Computational & Applied Genomics Program

Koo Foundation-Sun Yat Sen Cancer Center

Andrew Huang, Skye Cheng, Mei-Hua Tsou