ROSETTA

BIOSOFTWARE

www.rosettabio.com

Computational Biology:

An Industrial Perspective

Lee Weng, Ph.D.

ACM Symposium on Applied Computing

March 10, 2003

Overview

Molecular Profiling

Gene Expression and Microarray Technology

Computational Challenges

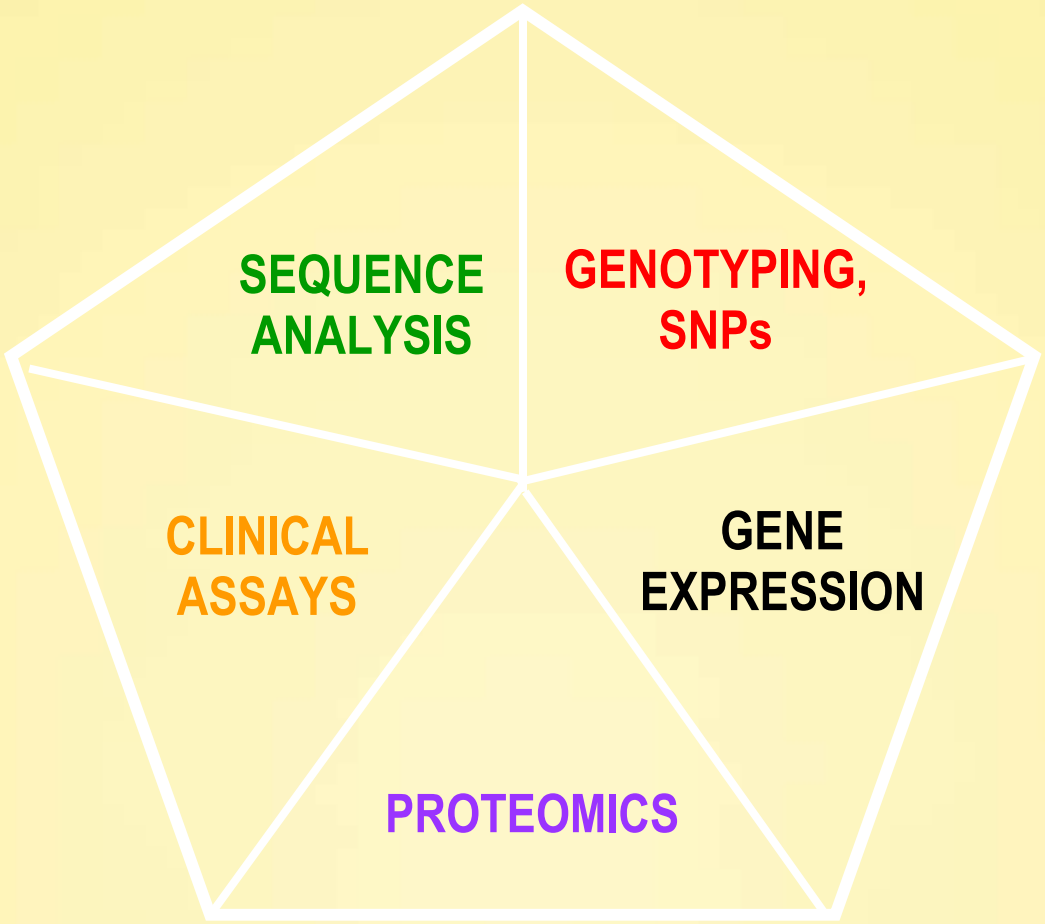
The Rosetta Resolver System

Beyond Textbooks

Data-Driven Knowledge Discoveries

Bring Data back to Biology

Molecular Profiling - A System Biology Approach



Industrialization of Molecular Biology

New high-throughput data-acquisition technologies have fundamentally changed the way to study biology

» Genome-wide DNA sequencing
» Microarrays for gene expression
» LC/MS-based protein expression analysis
» Large-scale molecular profiling becomes possible

Numerical computing is the corner stone of molecular profiling

» More powerful data storage, retrieval and management tools
» New computing methods and environments for data analysis and knowledge discovery
» New standards for information exchanges

Knowledge discovery ≠ Data integration

Opportunities in Molecular Profiling

Study Biology as a System

» Genes are not isolated. Groups of genes work together to serve a biological function. A stimulus will affect many genes.

» Gene expression profiles provide insights to the function of the biological system.

Study Biology as a Collaborative Team Effort

» It has become less and less practical to study a complex biological system by one or two biologists.

» Systematic data collection and information sharing have become inevitable.

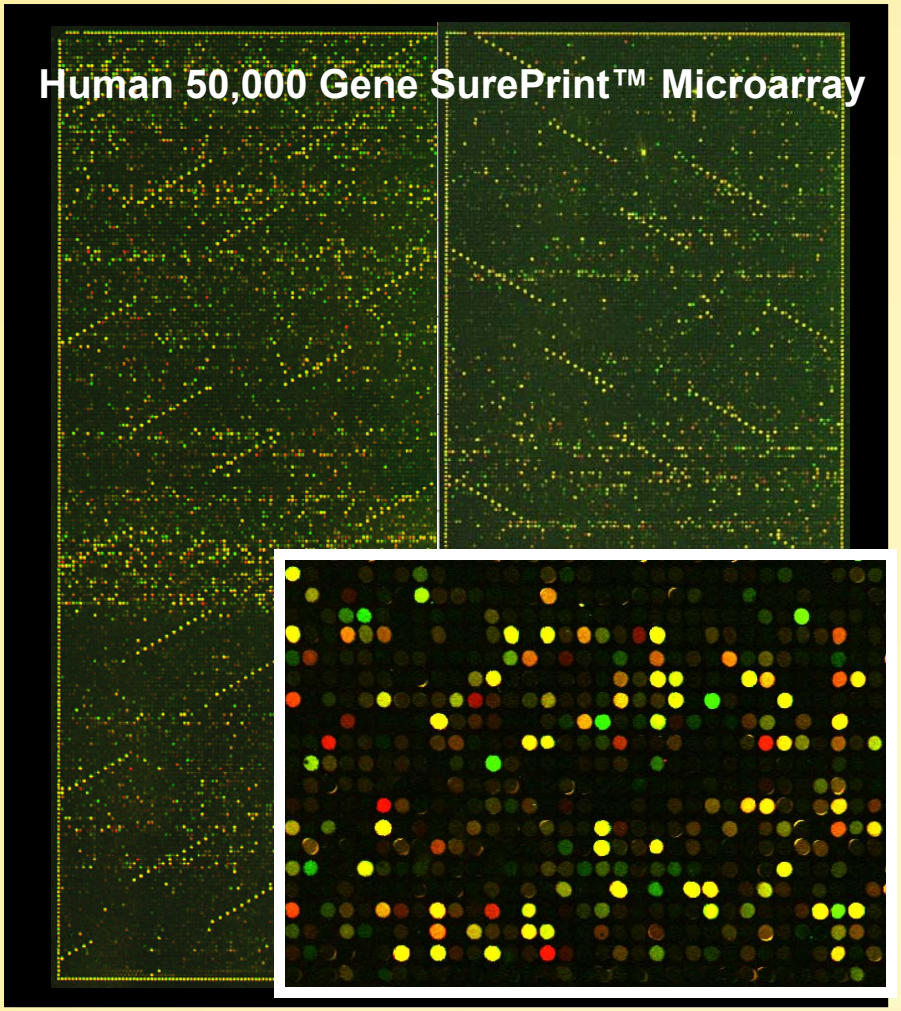
Study Biology as an Industry

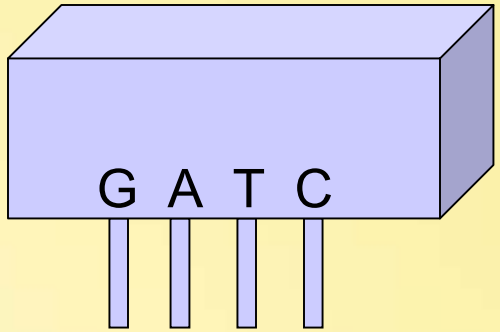
» Drug discovery becomes more difficult and costly. It is an enterprise-level effort.

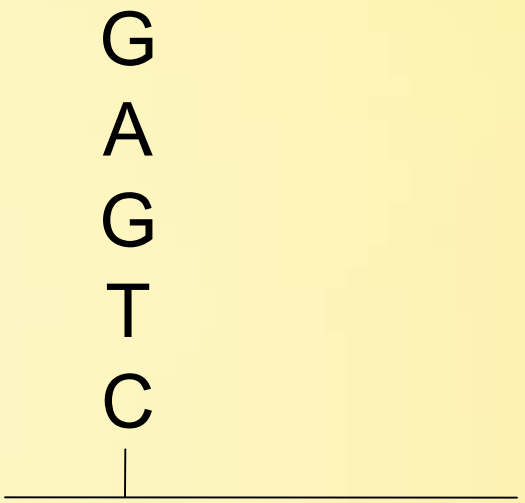
» Drug development requires enterprise solutions in data management, data analysis, and knowledge discovery.

» These solutions should meet various regulatory requirements, such as FDA 21 CFR Part 11.

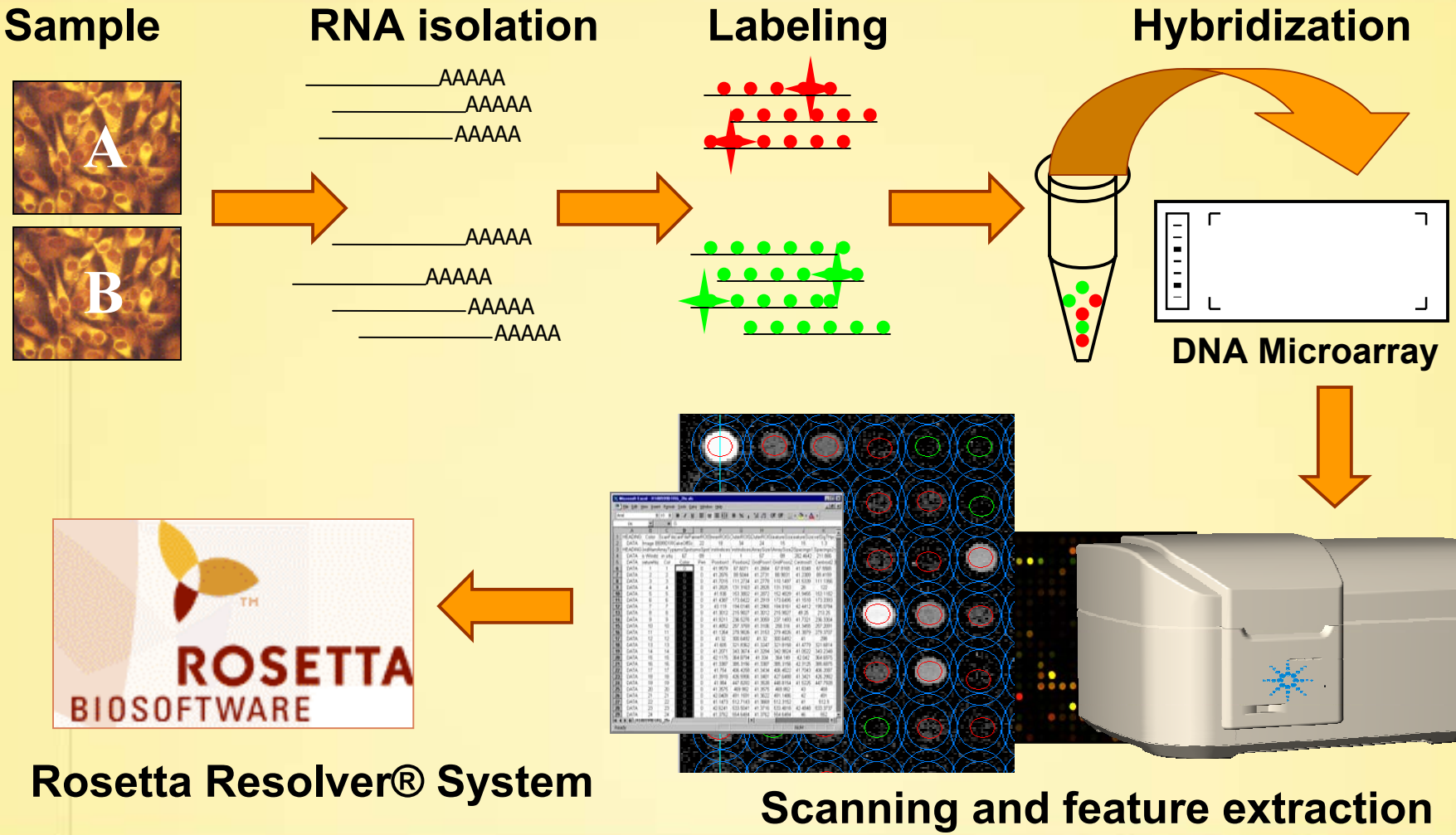
Example: Microarray Technology for Gene Expression Study



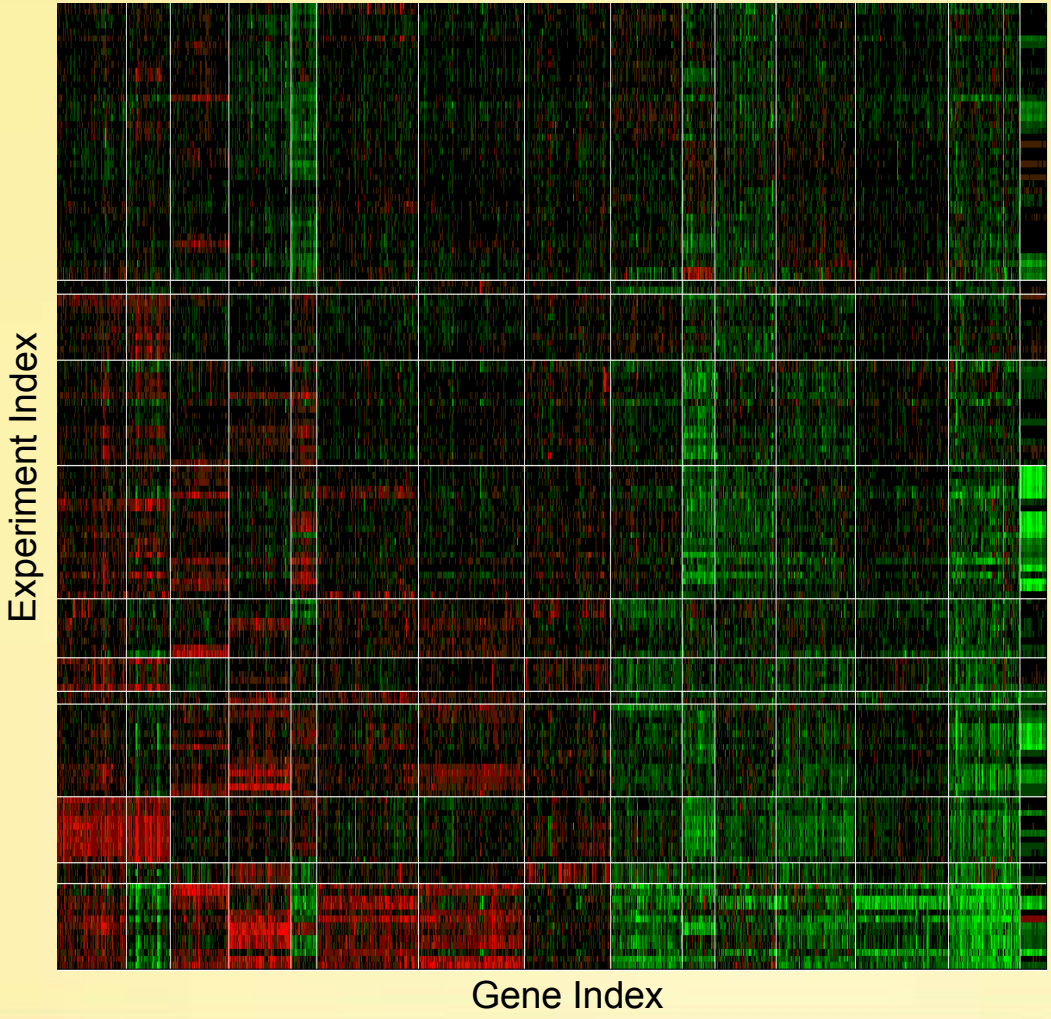




Processes in Analyzing Microarray Gene-Expression Data



Example: Compendium Approach to Hepatotoxicity



50+ Compounds act as Reference Standards

146 experiments
~1,800 genes

Down-regulated gene
Up-regulated gene

Challenges in Molecular Profiling

Large Volume of Data

» Microarray technology generates terabytes of gene expression data.

» How to effectively manage the data? How to make the data available and the information searchable in an enterprise environment? How to efficiently leverage available analysis tools on the large data source?

Small Number of Replications

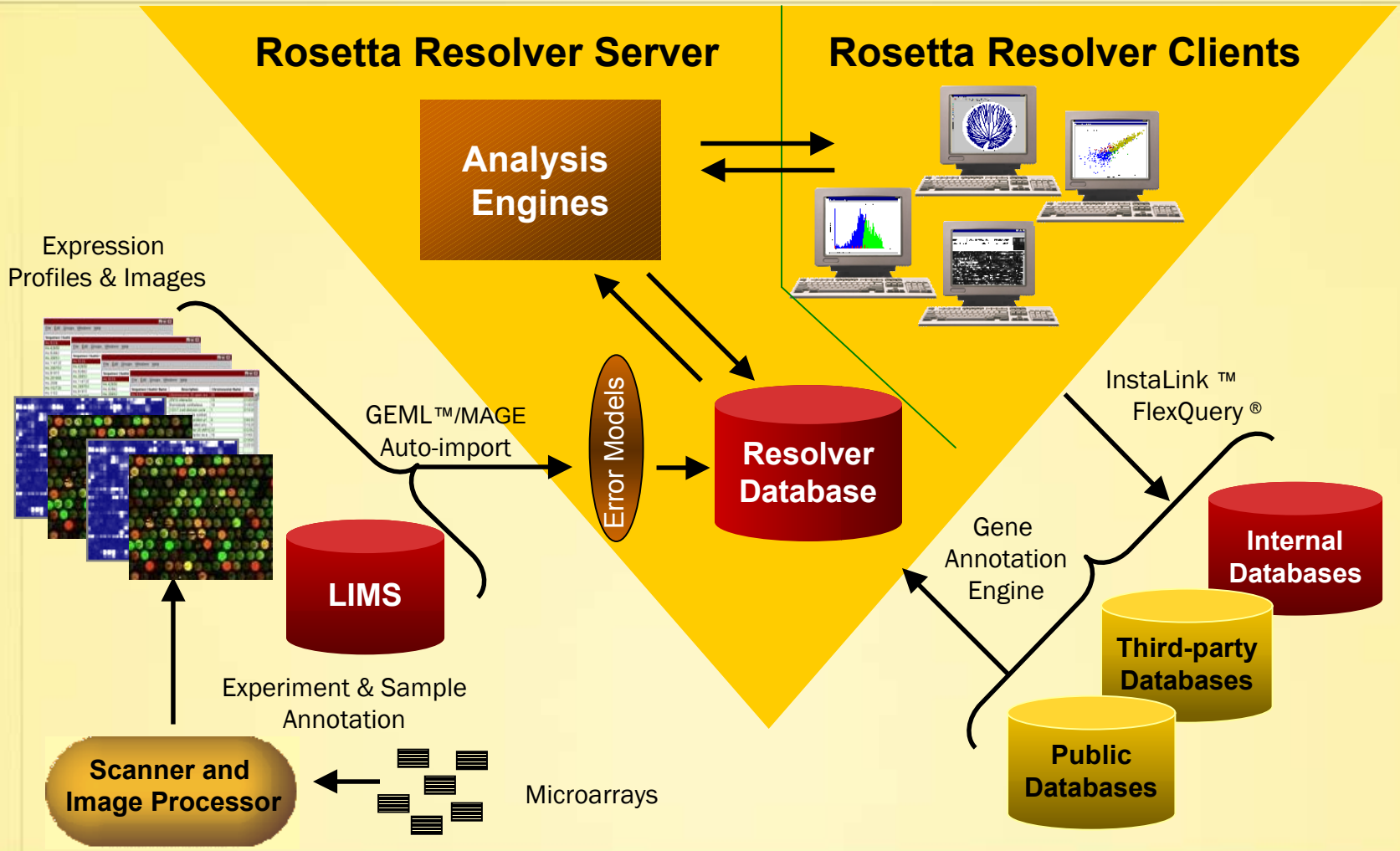
» Limited by available materials and high experiment costs, the number of replicates in molecular profiling is always very small, if any. Traditional statistical methods do not work well in this case.

Limited Knowledge

» Need to associate profiles of cellular constituents to help interpret the biological functions.

» Need to leverage existing biological knowledge in discovering new knowledge from the data.

Rosetta Resolver® System: An Enterprise Solution for Gene Expressions

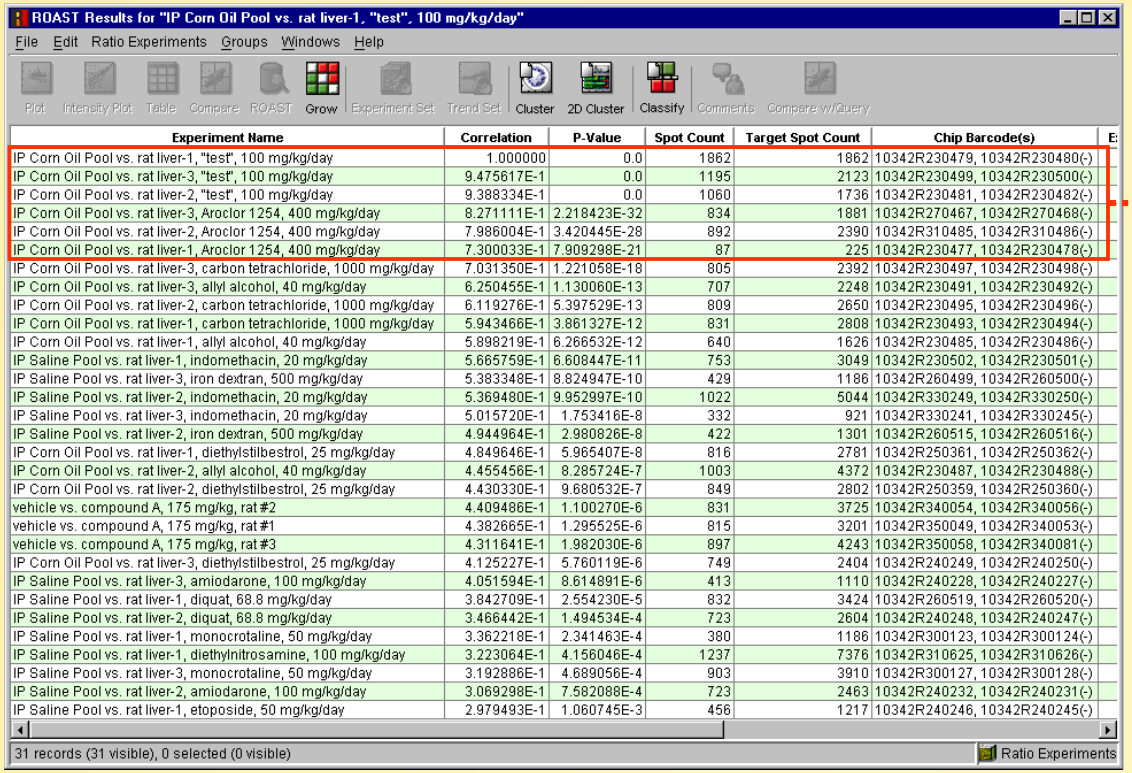


Enterprise solution to:

- Compile gene expression information from a wide variety of technologies in a central repository.
- Share information and analysis results collaboratively throughout the organization.
- Perform large-scale intensity- and ratio-based analyses that leverage entire gene expression databases.
- Publish and exchange data with collaborators in GEML™ and MAGE formats for visualization and analysis in the Rosetta Resolver system and other software applications.

Example: Rosetta Array Search Tool - ROAST

ROAST is analogous to BLAST for expression profile similarity searching. Use ROAST to search for co-regulated sequences, reporters, exons or UniGenes, as well as similar experiments.



Experiment Name	Correlation	P-Value	Spot Count	Target Spot Count	Chip Barcode(s)	E:
IP Corn Oil Pool vs. rat liver-1, "test", 100 mg/kg/day	1.000000	0.0	1862	1862	10342R230479, 10342R230480(-)	
IP Corn Oil Pool vs. rat liver-3, "test", 100 mg/kg/day	9.475617E-1	0.0	1195	2123	10342R230499, 10342R230500(-)	
IP Corn Oil Pool vs. rat liver-2, "test", 100 mg/kg/day	9.388334E-1	0.0	1060	1736	10342R230481, 10342R230482(-)	
IP Corn Oil Pool vs. rat liver-3, Aroclor 1254, 400 mg/kg/day	8.271111E-1	2.218423E-32	834	1881	10342R270467, 10342R270468(-)	
IP Corn Oil Pool vs. rat liver-2, Aroclor 1254, 400 mg/kg/day	7.986004E-1	3.420445E-28	892	2390	10342R310485, 10342R310486(-)	
IP Corn Oil Pool vs. rat liver-1, Aroclor 1254, 400 mg/kg/day	7.300033E-1	7.909298E-21	87	225	10342R230477, 10342R230478(-)	
IP Corn Oil Pool vs. rat liver-3, carbon tetrachloride, 1000 mg/kg/day	7.031350E-1	1.221058E-18	805	2392	10342R230497, 10342R230498(-)	
IP Corn Oil Pool vs. rat liver-3, allyl alcohol, 40 mg/kg/day	6.250455E-1	1.130060E-13	707	2248	10342R230491, 10342R230492(-)	
IP Corn Oil Pool vs. rat liver-2, carbon tetrachloride, 1000 mg/kg/day	6.119276E-1	5.397529E-13	809	2650	10342R230495, 10342R230496(-)	
IP Corn Oil Pool vs. rat liver-1, carbon tetrachloride, 1000 mg/kg/day	5.943466E-1	3.861327E-12	831	2808	10342R230493, 10342R230494(-)	
IP Corn Oil Pool vs. rat liver-1, allyl alcohol, 40 mg/kg/day	5.898219E-1	6.266532E-12	640	1626	10342R230485, 10342R230486(-)	
IP Saline Pool vs. rat liver-1, indomethacin, 20 mg/kg/day	5.665759E-1	6.608447E-11	753	3049	10342R230502, 10342R230501(-)	
IP Saline Pool vs. rat liver-3, iron dextran, 500 mg/kg/day	5.383348E-1	8.824947E-10	429	1186	10342R260499, 10342R260500(-)	
IP Saline Pool vs. rat liver-2, indomethacin, 20 mg/kg/day	5.369480E-1	9.952997E-10	1022	5044	10342R330249, 10342R330250(-)	
IP Saline Pool vs. rat liver-3, indomethacin, 20 mg/kg/day	5.015720E-1	1.753416E-8	332	921	10342R330241, 10342R330245(-)	
IP Saline Pool vs. rat liver-2, iron dextran, 500 mg/kg/day	4.944964E-1	2.980826E-8	422	1301	10342R260515, 10342R260516(-)	
IP Corn Oil Pool vs. rat liver-1, diethylstilbestrol, 25 mg/kg/day	4.849646E-1	5.965407E-8	816	2781	10342R250361, 10342R250362(-)	
IP Corn Oil Pool vs. rat liver-2, allyl alcohol, 40 mg/kg/day	4.455456E-1	8.285724E-7	1003	4372	10342R230487, 10342R230488(-)	
IP Corn Oil Pool vs. rat liver-2, diethylstilbestrol, 25 mg/kg/day	4.430330E-1	9.680532E-7	849	2802	10342R250359, 10342R250360(-)	
vehicle vs. compound A, 175 mg/kg, rat #2	4.409486E-1	1.100270E-6	831	3725	10342R340054, 10342R340056(-)	
vehicle vs. compound A, 175 mg/kg, rat #1	4.382665E-1	1.295525E-6	815	3201	10342R350049, 10342R340053(-)	
vehicle vs. compound A, 175 mg/kg, rat #3	4.311641E-1	1.982030E-6	897	4243	10342R350058, 10342R340081(-)	
IP Corn Oil Pool vs. rat liver-3, diethylstilbestrol, 25 mg/kg/day	4.125227E-1	5.760119E-6	749	2404	10342R240249, 10342R240250(-)	
IP Saline Pool vs. rat liver-3, amiodarone, 100 mg/kg/day	4.051594E-1	8.614891E-6	413	1110	10342R240228, 10342R240227(-)	
IP Saline Pool vs. rat liver-1, diquat, 68.8 mg/kg/day	3.842709E-1	2.554230E-5	832	3424	10342R260519, 10342R260520(-)	
IP Saline Pool vs. rat liver-2, diquat, 68.8 mg/kg/day	3.466442E-1	1.494534E-4	723	2604	10342R240248, 10342R240247(-)	
IP Saline Pool vs. rat liver-1, monocrotaline, 50 mg/kg/day	3.362218E-1	2.341463E-4	380	1186	10342R300123, 10342R300124(-)	
IP Saline Pool vs. rat liver-1, diethylnitrosamine, 100 mg/kg/day	3.223064E-1	4.156046E-4	1237	7376	10342R310625, 10342R310626(-)	
IP Saline Pool vs. rat liver-3, monocrotaline, 50 mg/kg/day	3.192886E-1	4.689056E-4	903	3910	10342R300127, 10342R300128(-)	
IP Saline Pool vs. rat liver-2, amiodarone, 100 mg/kg/day	3.069298E-1	7.582088E-4	723	2463	10342R240232, 10342R240231(-)	
IP Saline Pool vs. rat liver-1, etoposide, 50 mg/kg/day	2.979493E-1	1.060745E-3	456	1217	10342R240246, 10342R240245(-)	

Correlation between gene expression experiments may indicate similar mechanisms of toxicity between known and unknown compounds.

A Million-Dollar Question: Which Genes are Differentially Expressed?

Limitations of Fold-Change Method

» Biologists often use 2-fold change as the threshold to detect differential expressions.

» It suffers in low sensitivity and low specificity because it does not consider the error of the measurement.

Limitations of Traditional Statistical Tests

» Textbook t-test or ANOVA test do not work well when the number of replicated experiments is small.

Solutions: Rosetta Resolver Error Models

Beyond Textbooks: Error Models in the Rosetta Resolver® System

Error Models

» Technology-specific error models are used to estimate microarray measurement errors.

» The Rosetta Resolver system leverages the error models together with experimental replicates to compute accurate P-values and error bars for every gene expression measurement.

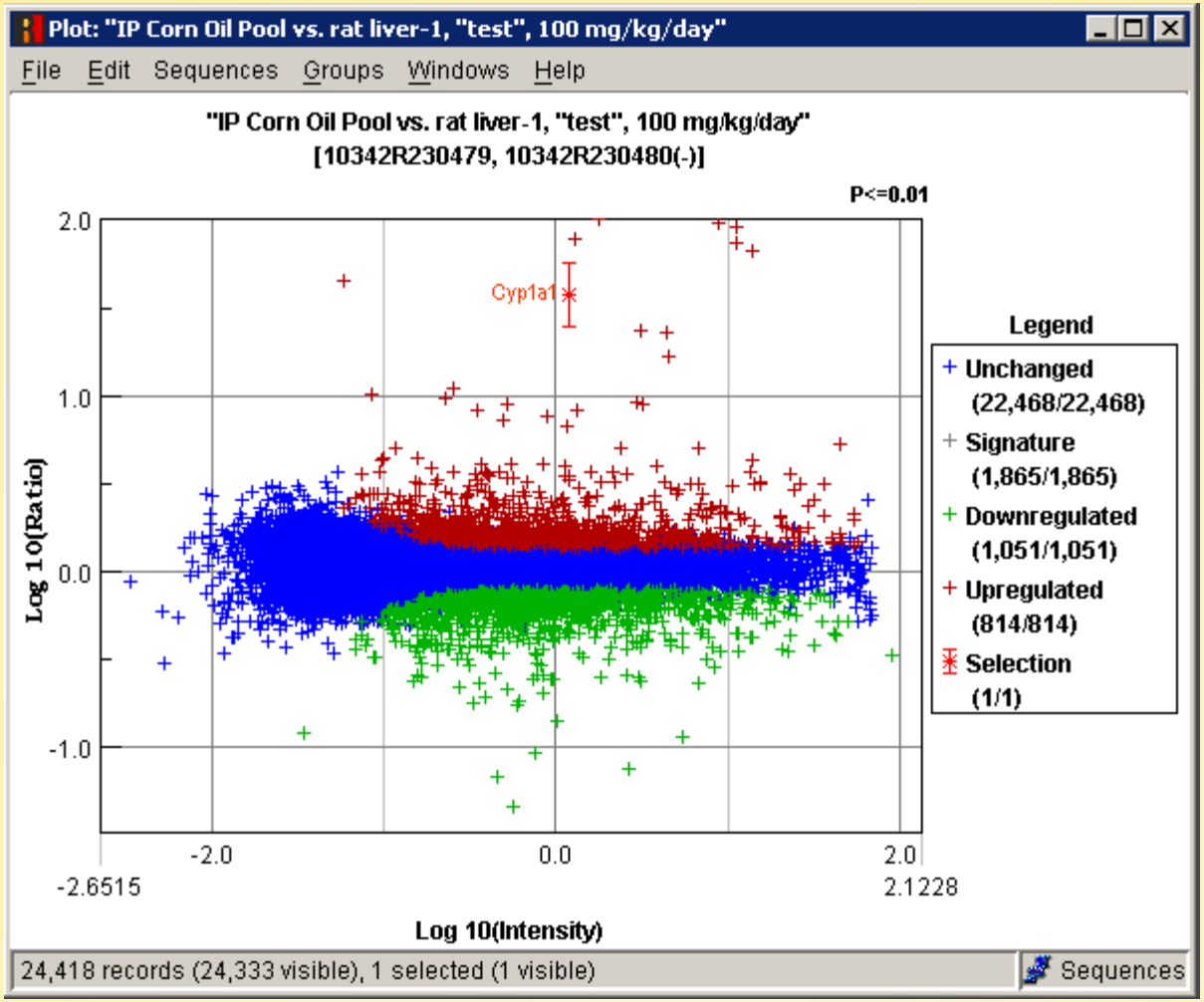
Benefits

» Error models help to reliably estimate variance when the number of replicates is small.

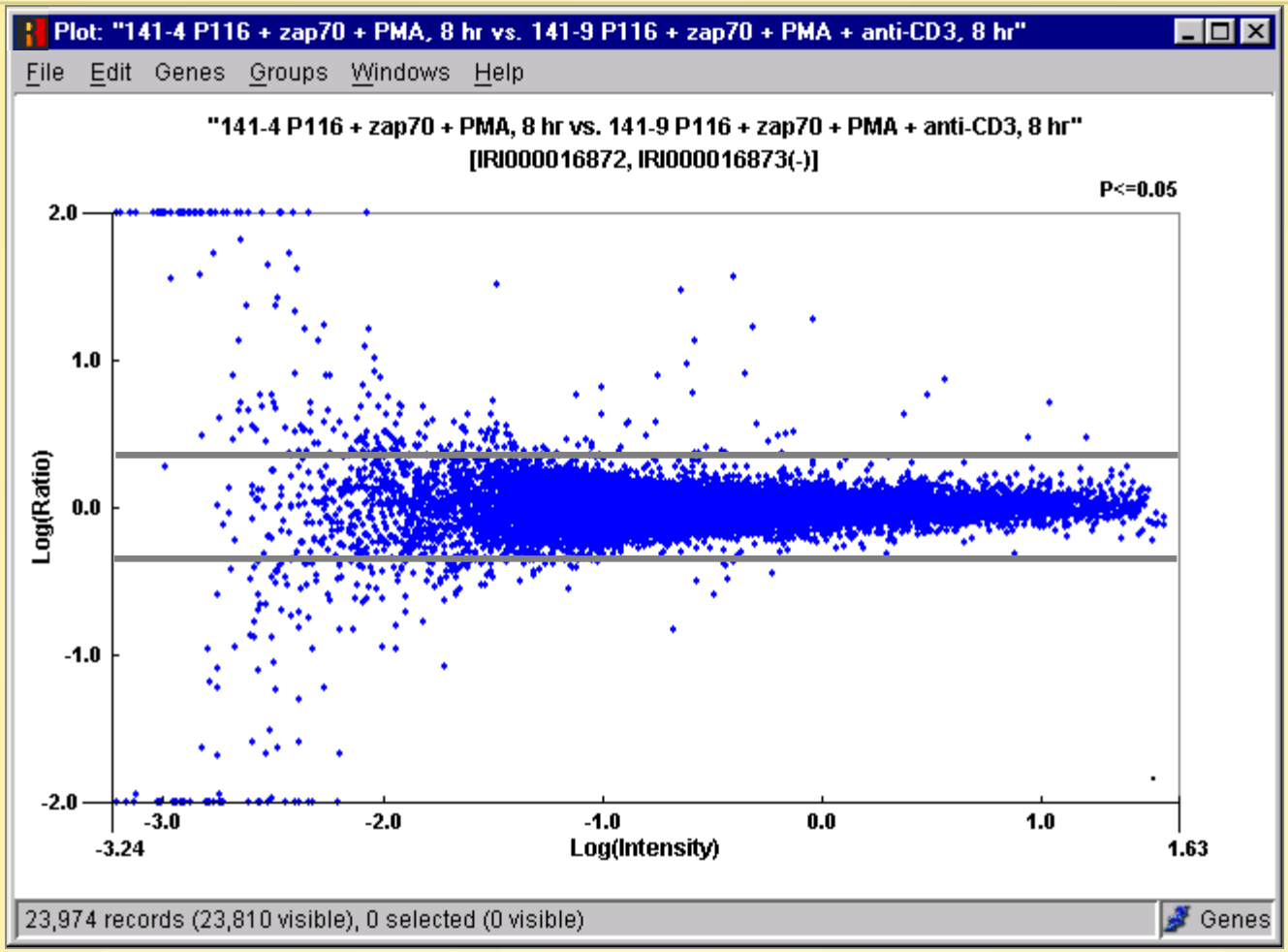
» Error models provide continued quality control in data analysis.

» P-values and error bars are propagated and leveraged throughout the analysis, adding additional predictive power to cluster analysis, similarity searching, trend analysis, etc.

Example: Error Model Estimates Expression Measurement Error

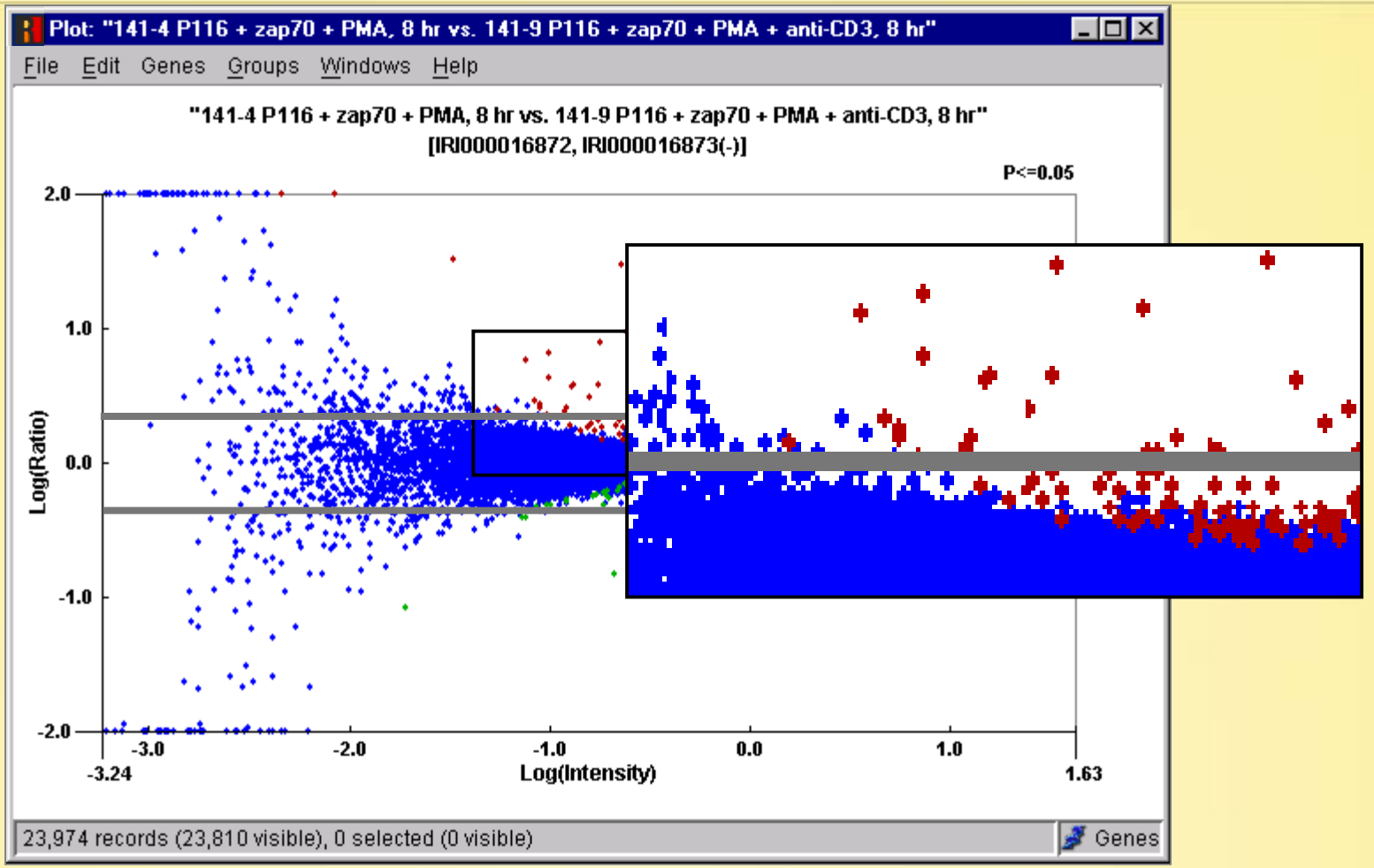


Example: Error Model Benefits in Ratio Analysis



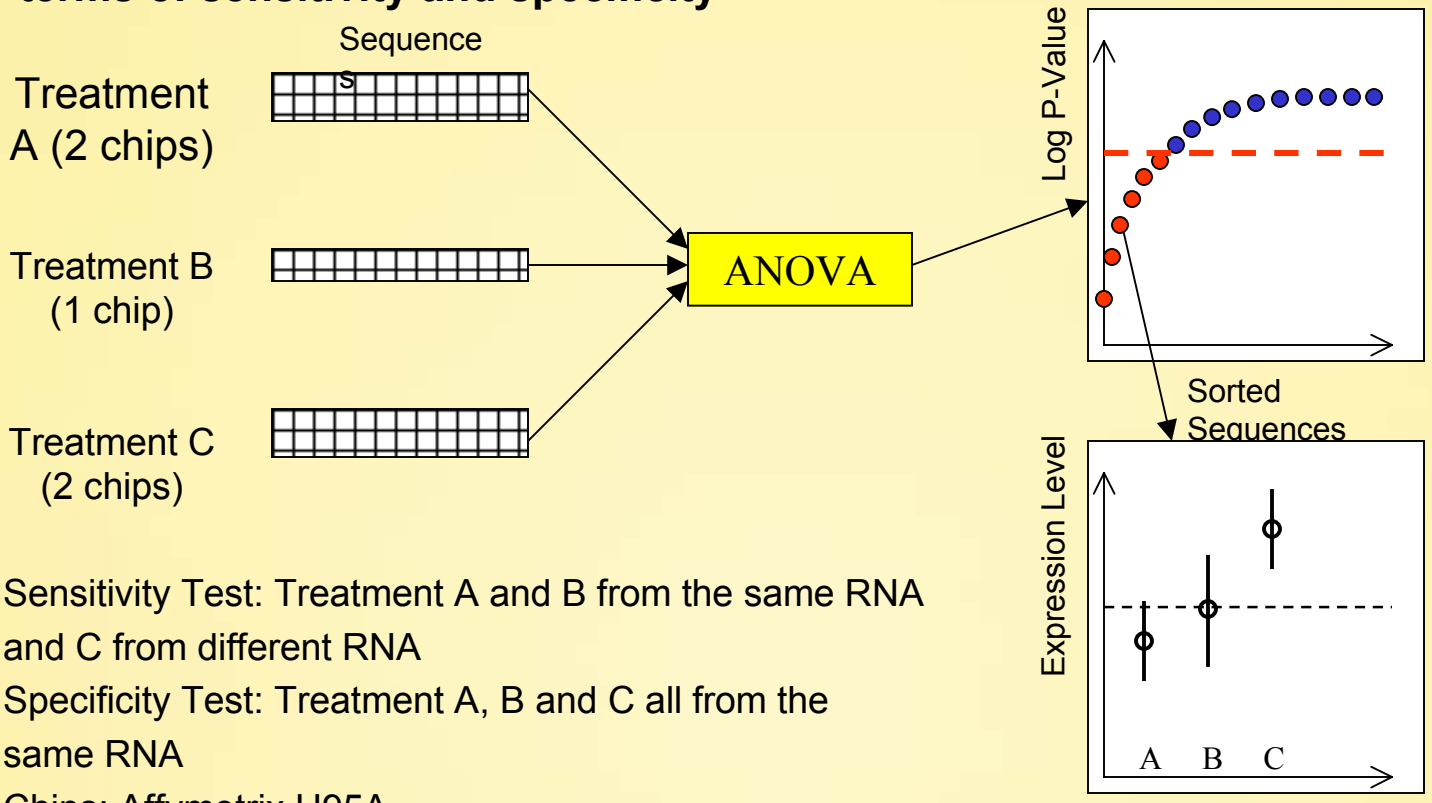
Which up/downregulations are statistically significant?

Example: Error Model Benefits in Ratio Analysis



Example: Error Model Benefits in ANOVA Tests

Compare the power of the textbook ANOVA and the improved ANOVA in terms of sensitivity and specificity



Sensitivity Test: Treatment A and B from the same RNA and C from different RNA

Specificity Test: Treatment A, B and C all from the same RNA

Chips: Affymetrix U95A

Example: Error Model Benefits in ANOVA Tests

For given threshold P-value<0.01, the improved ANOVA in the Rosetta Resolver System provides much lower false positive rate (better specificity) and much higher detection rate (better sensitivity) than the textbook ANOVA.

Analysis methods	False Positive rate	Detection rate
Textbook ANOVA	0.0093	0.17
Improved ANOVA	0.00048	0.30

Data-Driven Knowledge Discoveries

Unsupervised Learning Methods

» Discover new knowledge by data associations based on similarities.

» Clustering tools in the Rosetta Resolver system: agglomerative, divisive, k-mean, SOM, pattern grow, et al.

Supervised Learning Methods

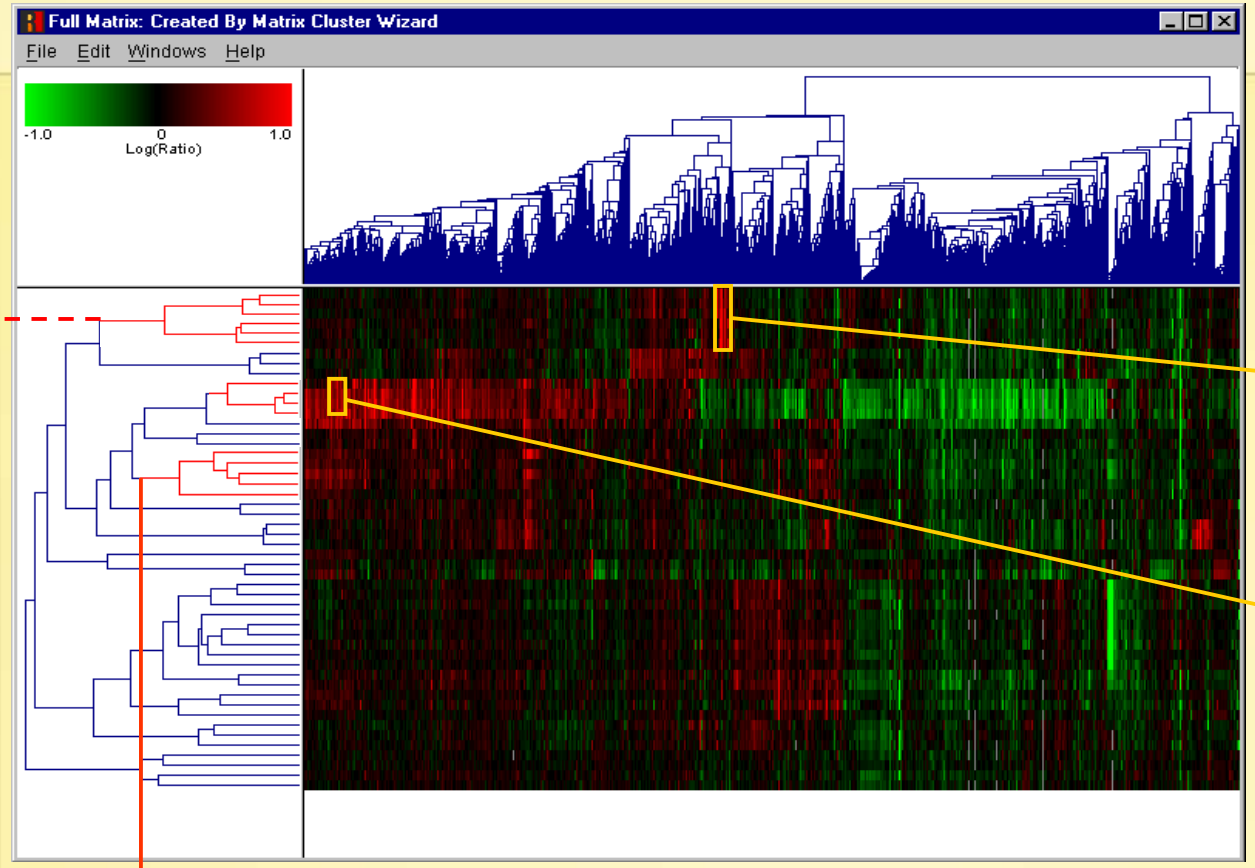
» Discover new knowledge by training.

» Classification tools in the Rosetta Resolver system: Bayesian classifier, et al.

Example: An Unsupervised Learning Case Study

» **Rosetta and Abbott Laboratories collaboration: Toxicogenomics**
Waring, et al. *Toxicology & Applied Pharmacology* 175, 28-42 (2001)

Toxicogenomics Signatures



CYP1A1 & CYP1B1 - regulated signature genes by treatment with Aroclor1254 and 3-methylcholanthrene.

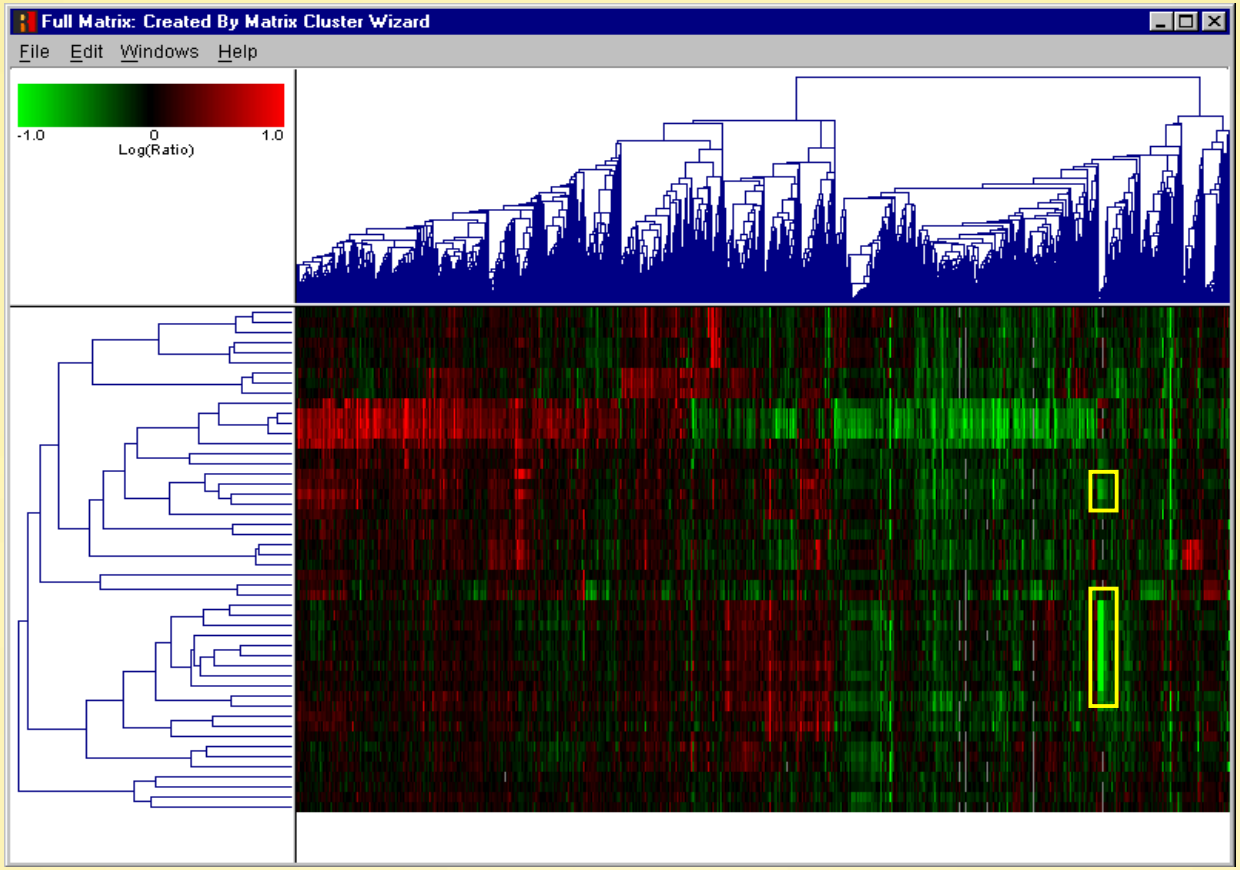
Cyclin D1 & Bax - up regulated signature genes by treatment with DNA-damaging diethylnitrosamine and monocrotaline.

Cluster: Allyl Alcohol and Carbon tetrachloride - Both compounds cause cell injury by free radical formation; co-regulated signature genes include genes involved in oxidation.

Cluster: Aroclor1254 and 3-methylcholanthrene - Both compounds are aromatic hydrocarbons that have similar mechanisms of toxicity.

Global vs. Local Similarities

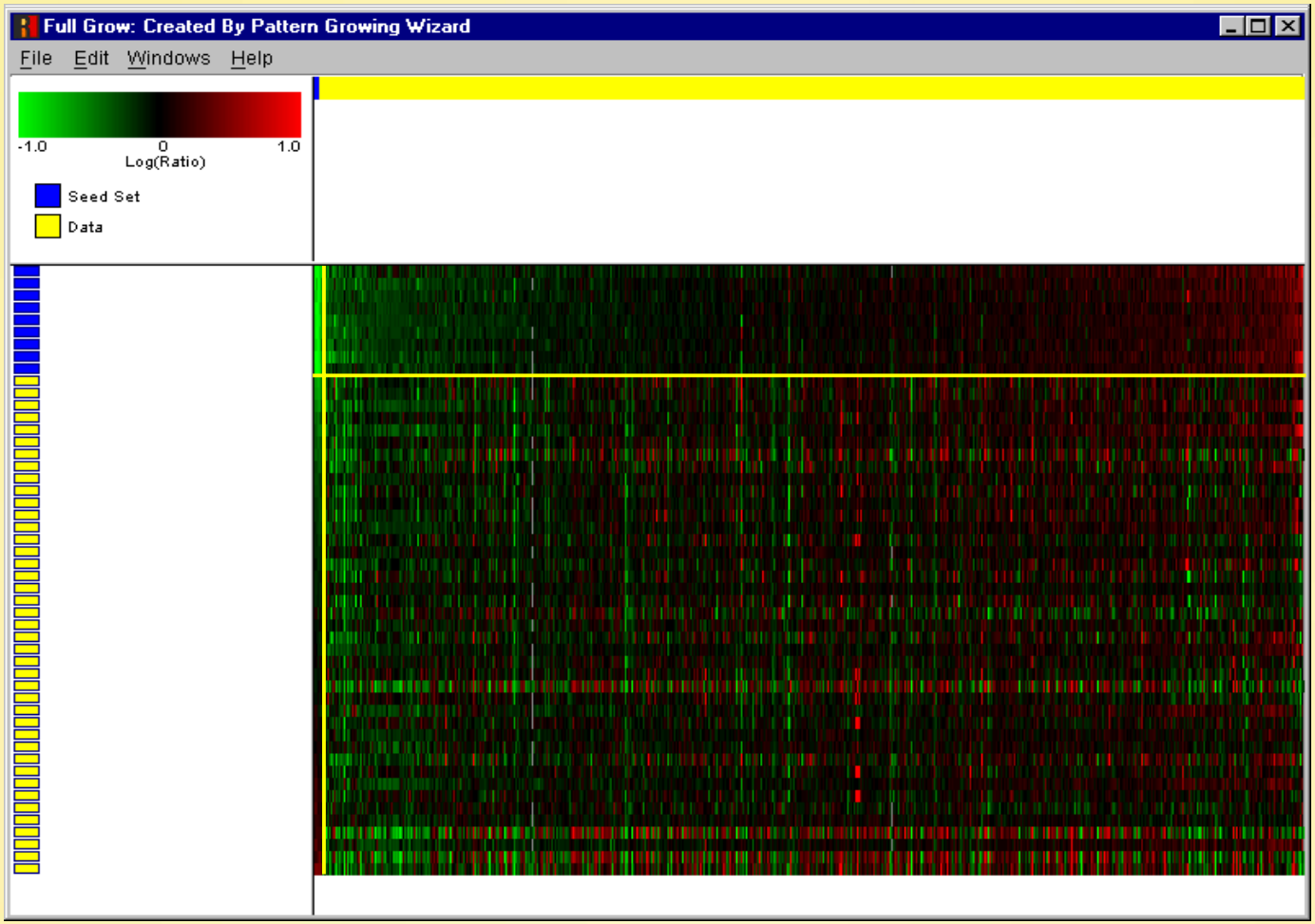
These experiments share a certain geneset response



47 Experiments

5,000 Responsive Genes

Global vs. Local Similarities : The GROW Algorithm



Example: A Supervised Learning Case Study

» **Rosetta and Netherlands Cancer Institute (NKI) collaboration: Breast Cancer**

van 't Veer, et al., *Nature* 415, 530-536 (2002)

Can Gene Expression Profiling be Used to Predict Clinical Outcome in Breast Cancer Patients?

Profile RNA samples from 98 breast tumors

- 44 with >5 years disease-free survival
- 34 with <5 years disease-free survival
- 20 with BRCA1 germ-line mutations

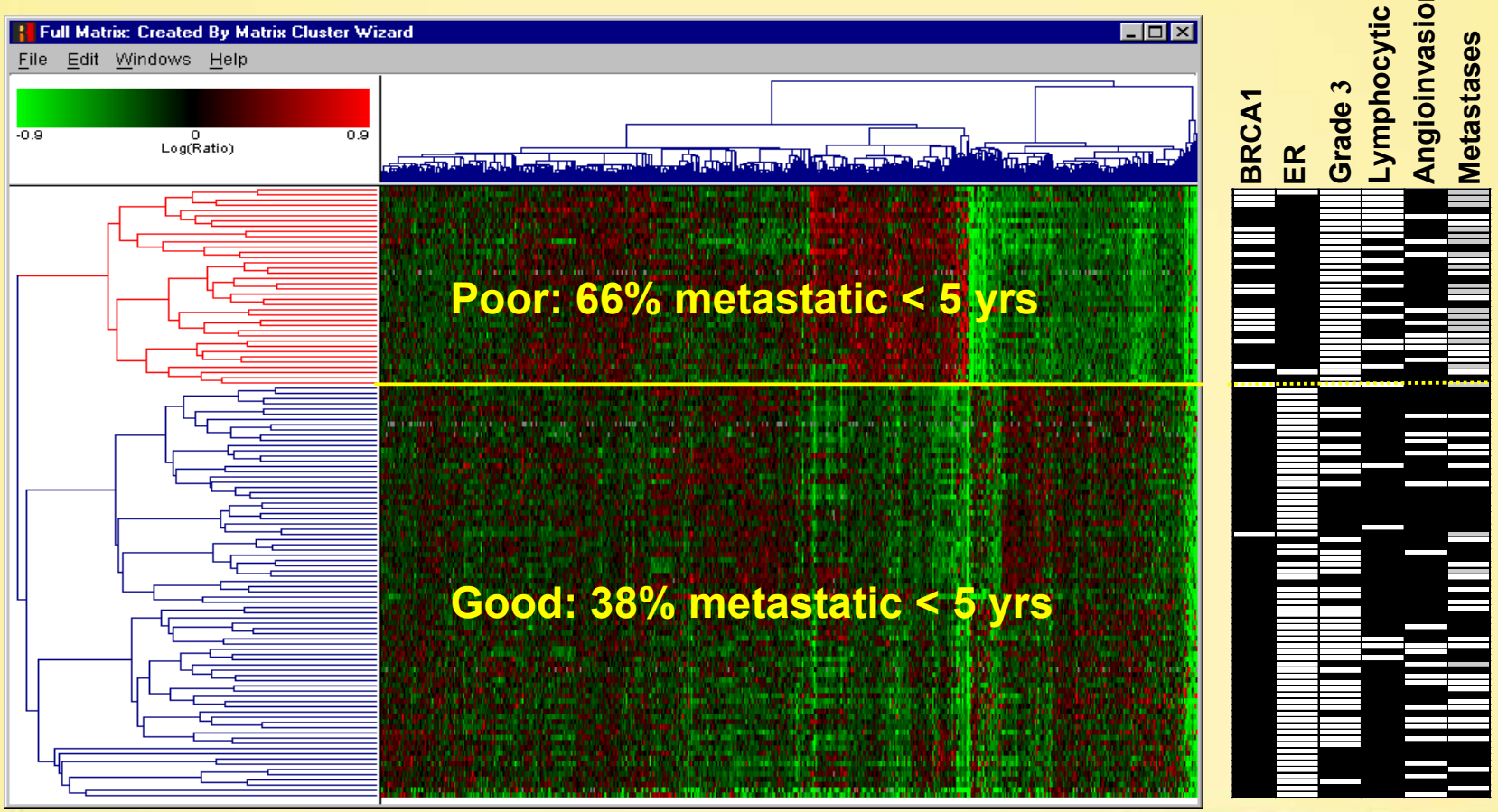
All samples carefully selected:

- Patient age < 55 years
- Tumor size < 5 cm; no lymph node involvement
- Samples well annotated

Objectives:

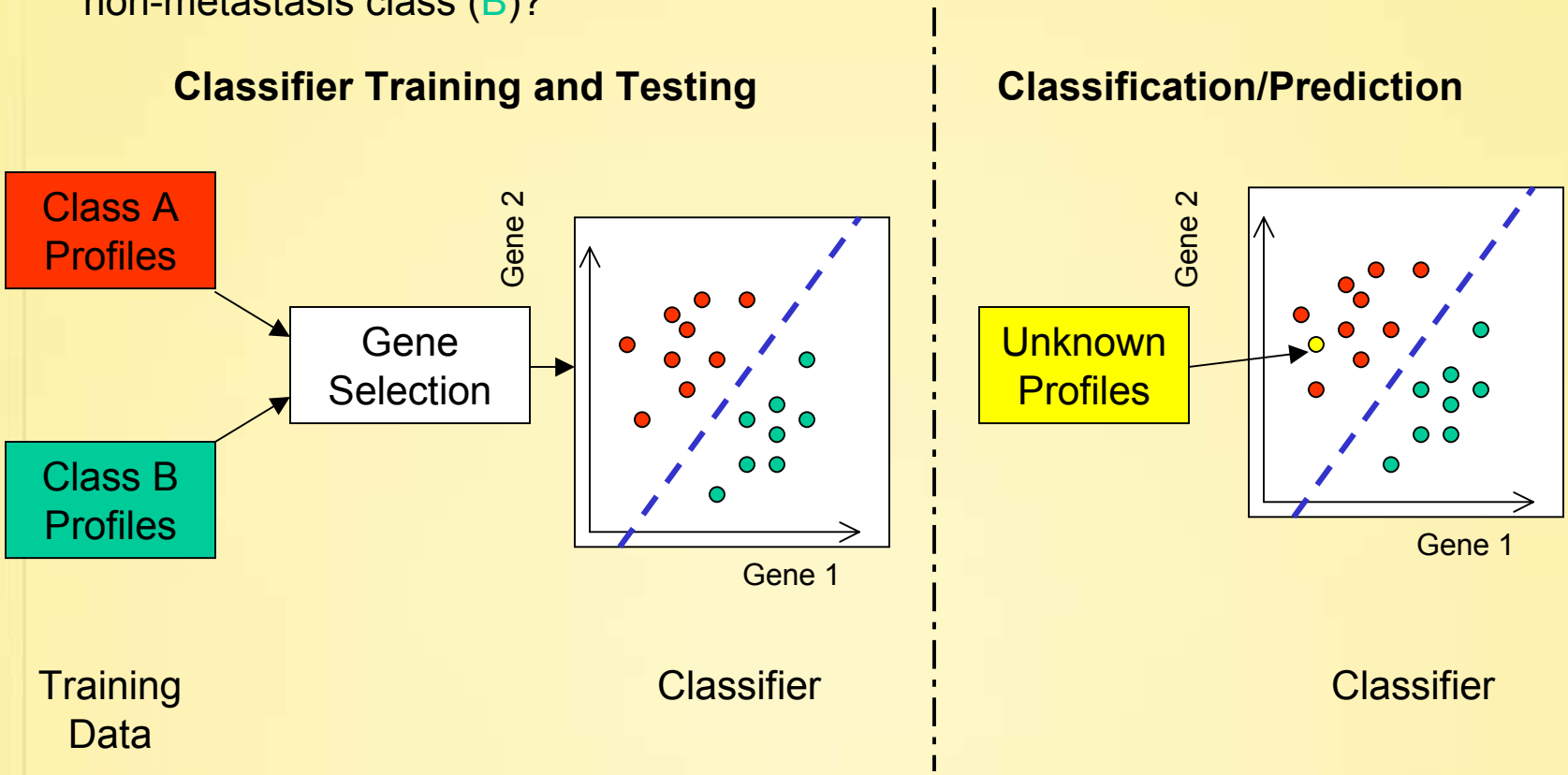
- Identify patterns of gene expression that correlate with prognosis
- Confirm results with a new set of 19 similar breast tumors

Unsupervised Clustering Divides Tumors into “Good Prognosis” and “Poor Prognosis” Tumors

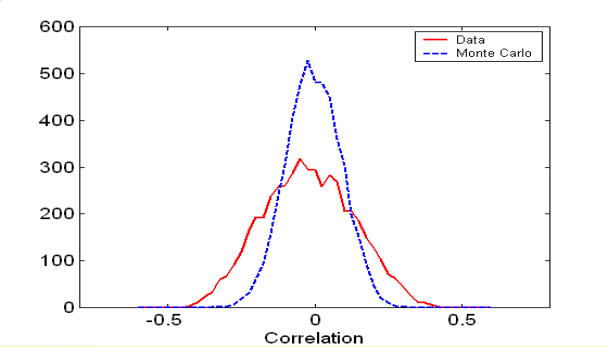


Classifier Concept

"Is this unknown profile most likely a member of the metastasis class (A) or the non-metastasis class (B)?"

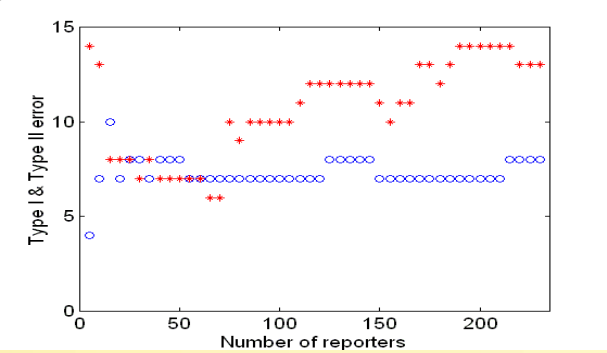


Steps for Developing a Classifier



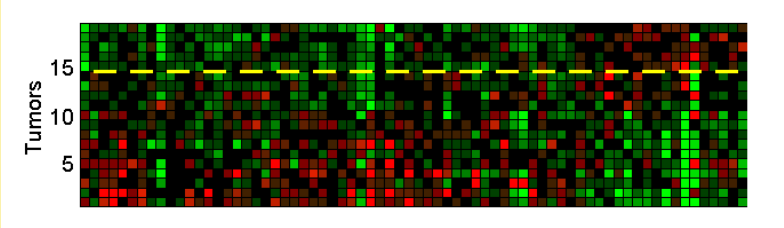
» **Select and rank features**

Define reporter genes that predict distant metastases



» **Optimize the number of reporter genes**

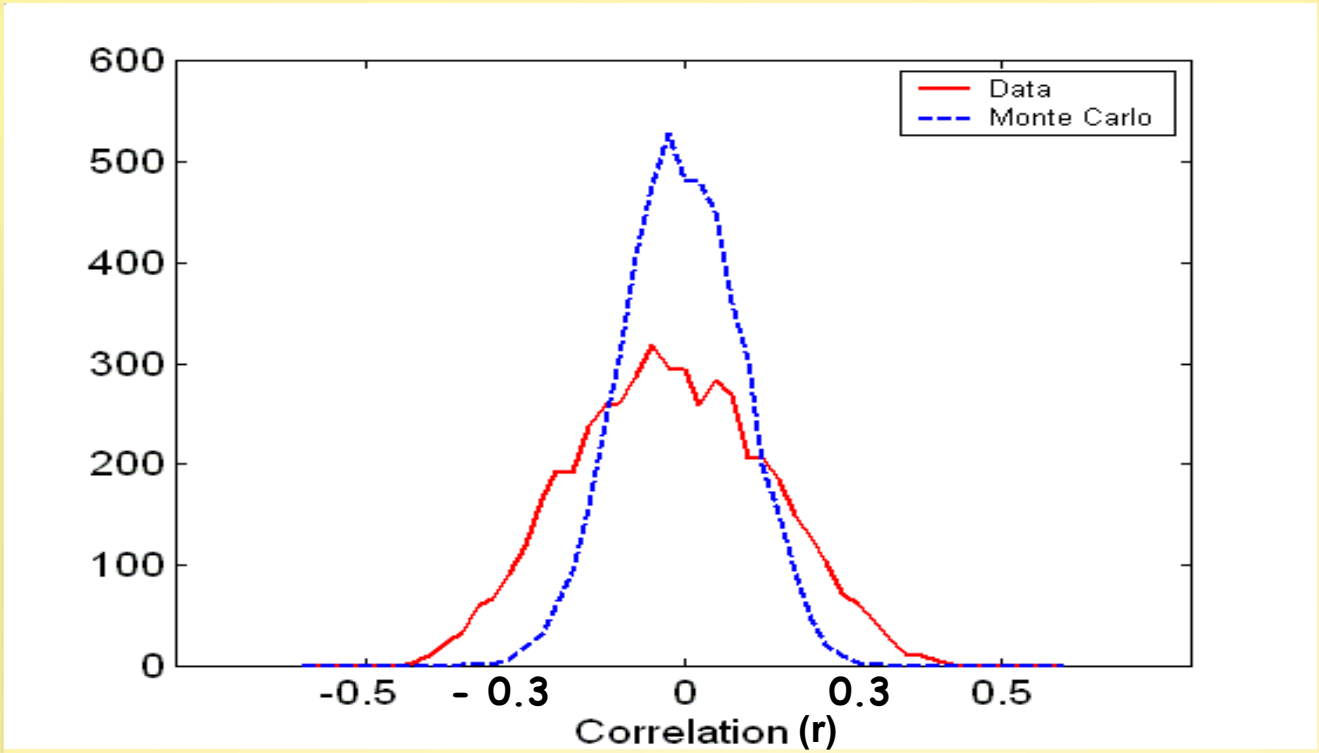
Leave-one-out cross-validation for defining optimal Classifier



» **Evaluate the power of classifier**

Independent validation set of 19 tumors

Define Reporter Genes that Predict Distant Metastases

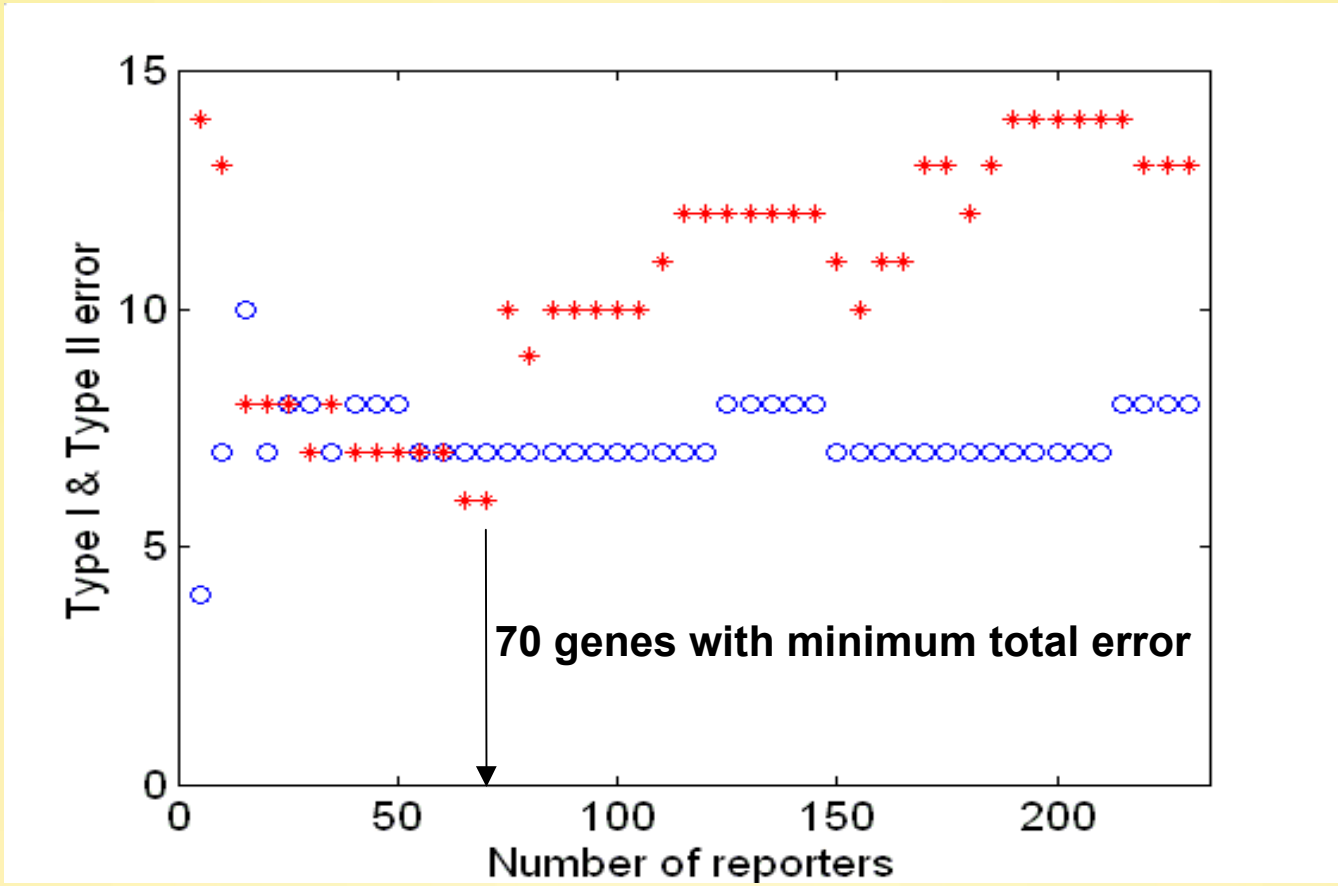


231 genes with prognostic category
red curve to r <- 0.3 &
red curve to r > 0.3

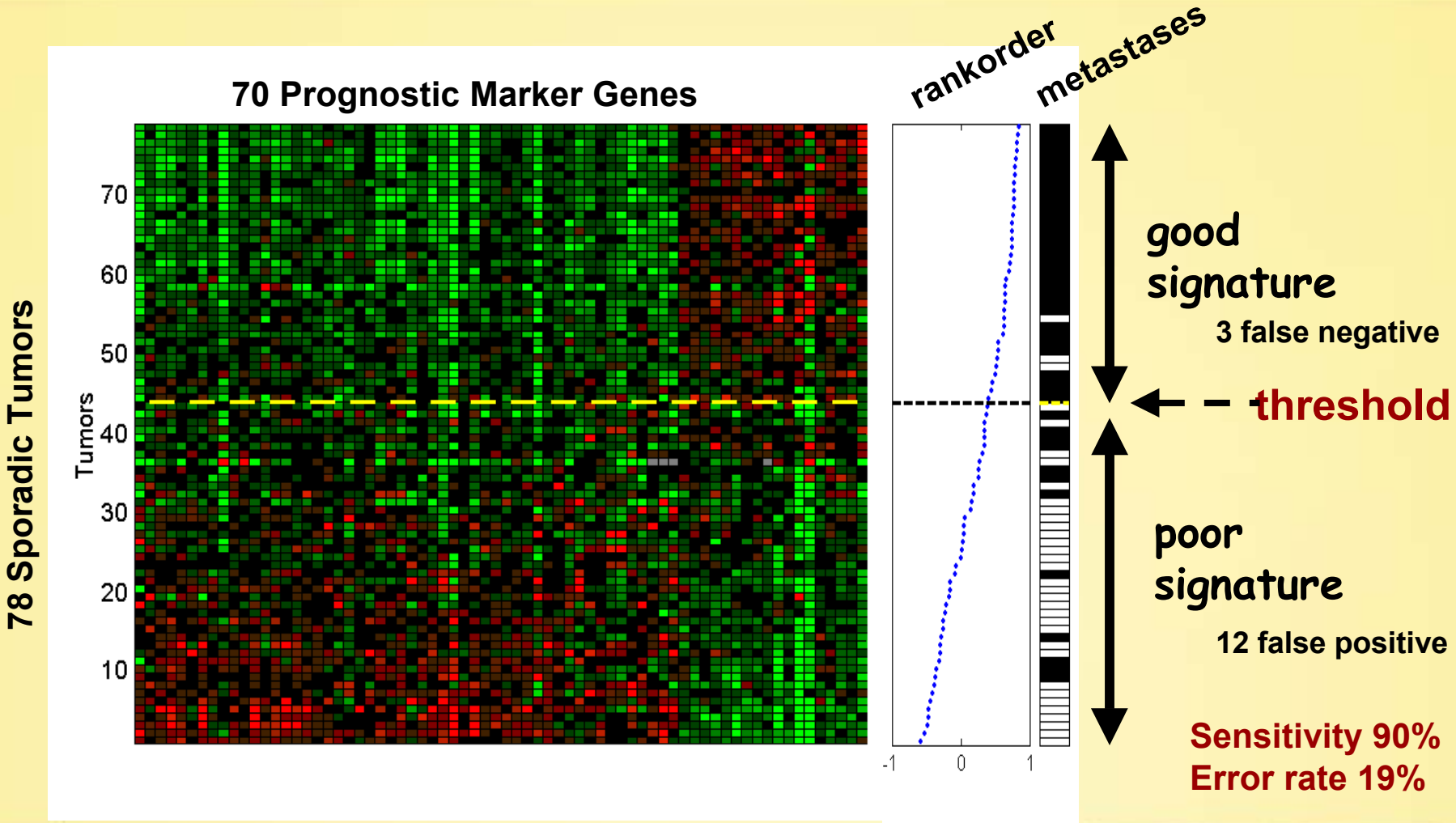
Red curve:
Histogram of correlation coefficients of genes with prognostic category (metastasis group versus no-metastasis group)

Blue curve:
Histogram of correlation coefficients of genes with prognosis from Monte Carlo analysis, where the metastasis designations have been randomized

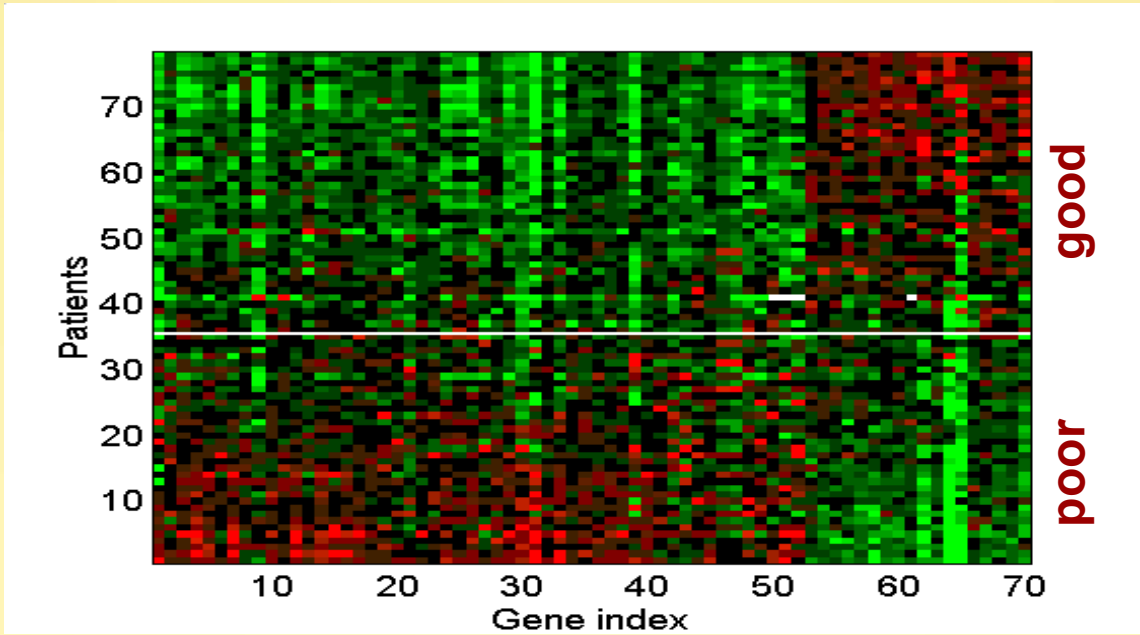
Optimize the Number of Reporter Genes Using Leave-One-Out Cross Validations



Supervised Classification Prognosis: Leave-One-Out Method for Defining Optimal Classifier

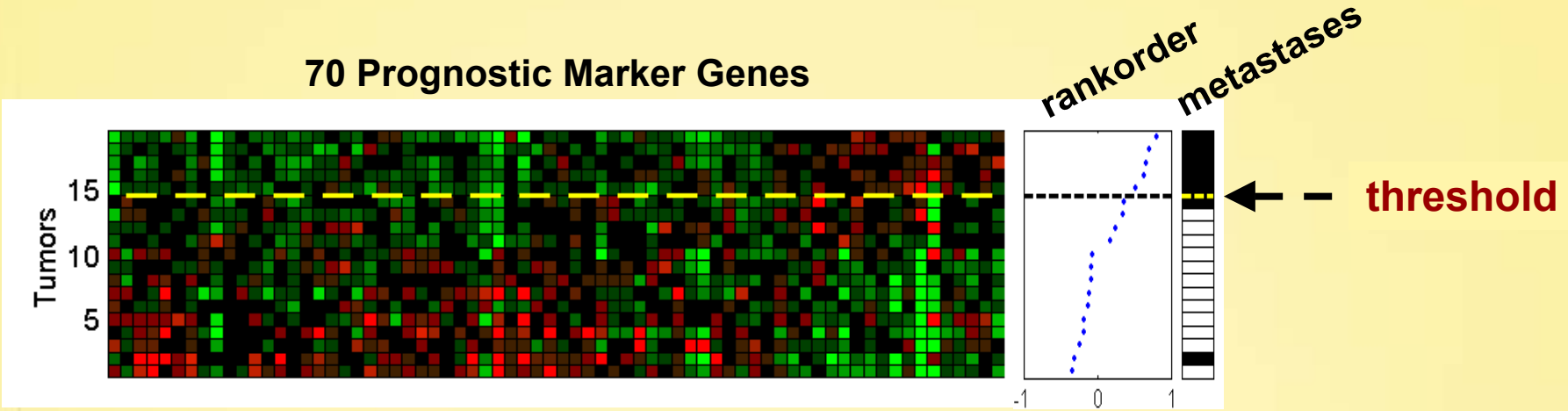


Functional Classes of Prognosis Reporter Genes



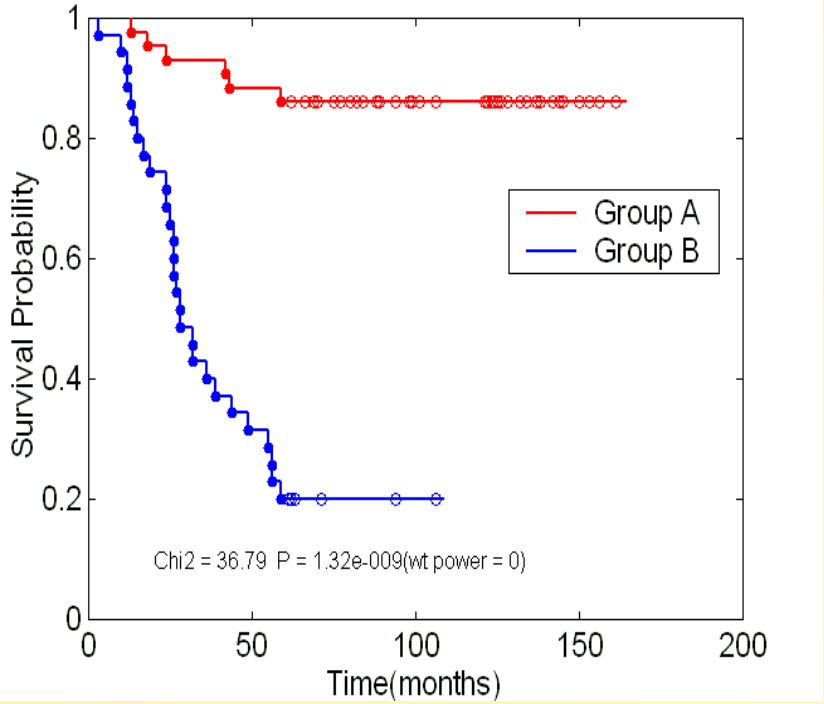
- **Cell cycle:** Cyclins, DNA replication, Checkpoint proteins
- **Invasion, metastasis:** (Metallo)proteinases
- **Angiogenesis:** VEGF Receptor
- **Signal transduction:** IGFBPs, Kinases

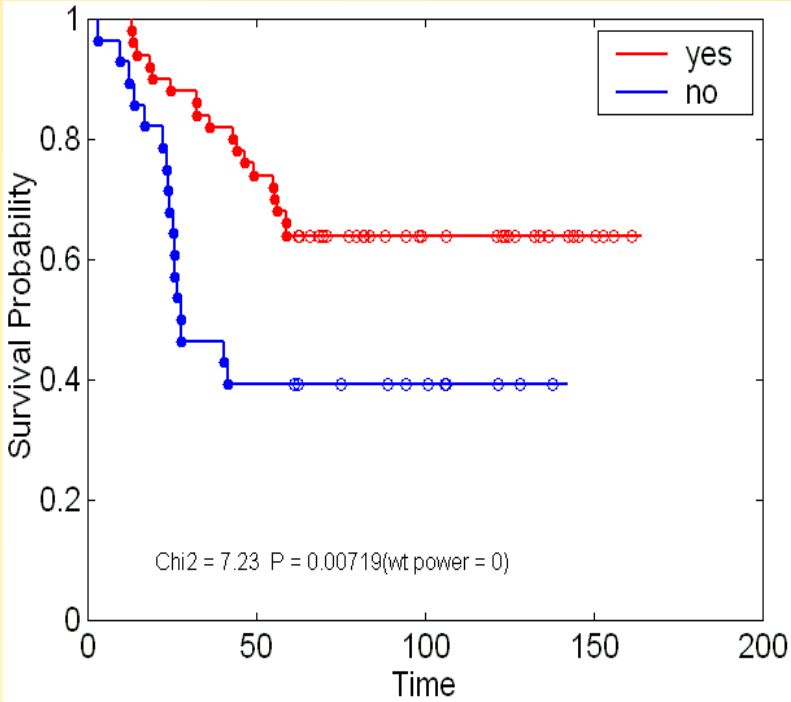
Classification of Set of 19 Tumors Using Prognosis Reporter Genes



only 2/19 tumors were misclassified:
90% accuracy, Fisher's exact p=0.0018

Microarray Reporters Predict Prognosis Better than Clinical Markers





Patients grouped by microarray-defined prognosis reporters

Patients grouped by Estrogen Receptor status

Profiling in Clinical Practice: Selection for Adjuvant Chemotherapy

Selection criteria:	Total patient group (n=78)	Metastatic disease < 5 years (n=34)	Disease free at >5 years (n=44)
St. Gallen consensus	64/78 (82%)	33/34 (97%)	31/44 (70%)
NIH-consensus	72/78 (92%)	32/34 (94%)	40/44 (91%)
Prognosis profile	43/78 (55%)	31/34 (91%)	12-18/44 (27%-41%)

Breast cancer patients eligible for adjuvant systemic therapy

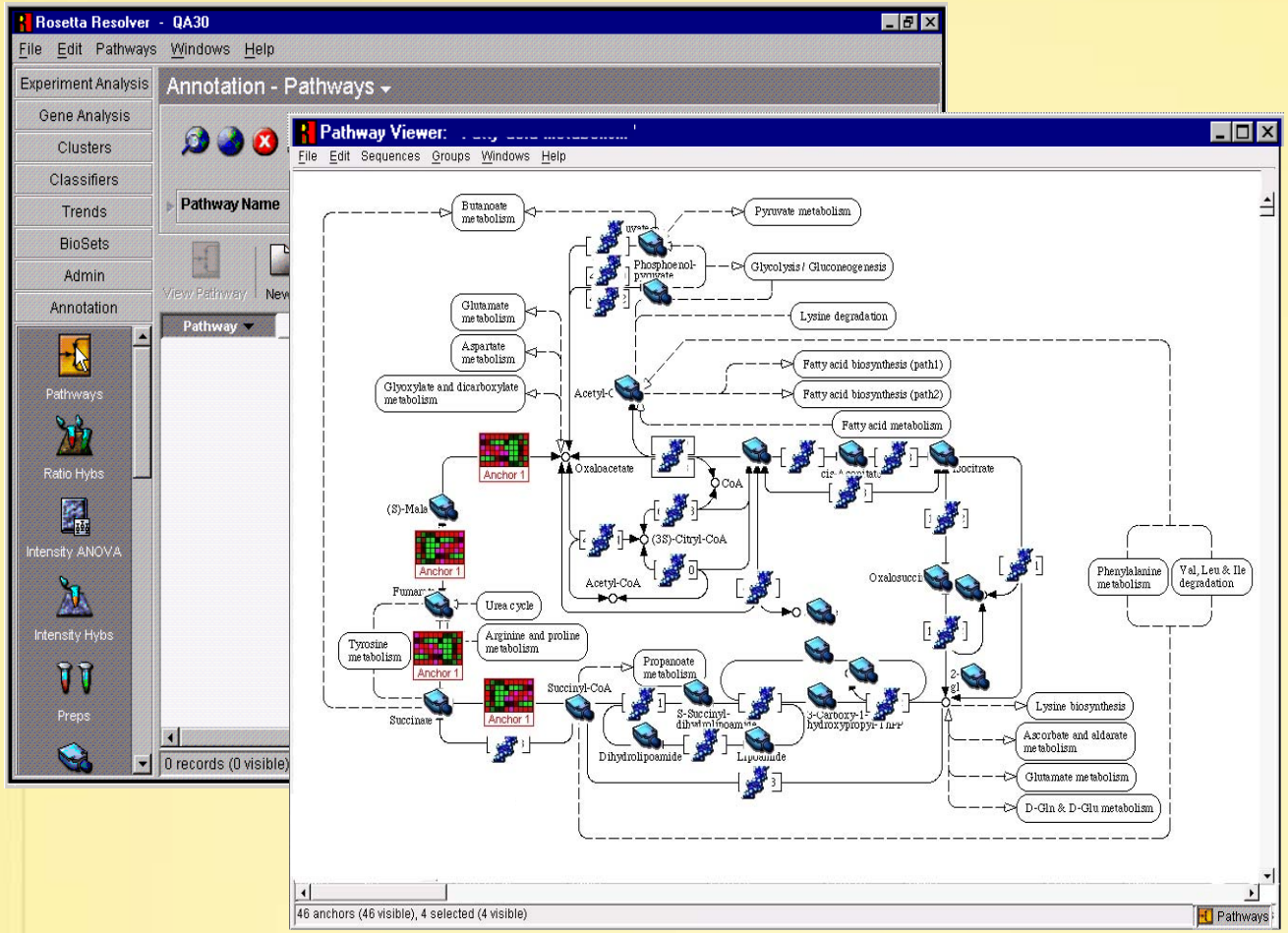
Profiling in Clinical Practice: Conclusions

Even small tumors are already programmed to metastatic phenotype.

Expression profiling:

- » selects patients that *should* receive adjuvant therapy similarly to conventional criteria.
- » can significantly *reduce* the number of patients who would receive adjuvant therapy unnecessarily.

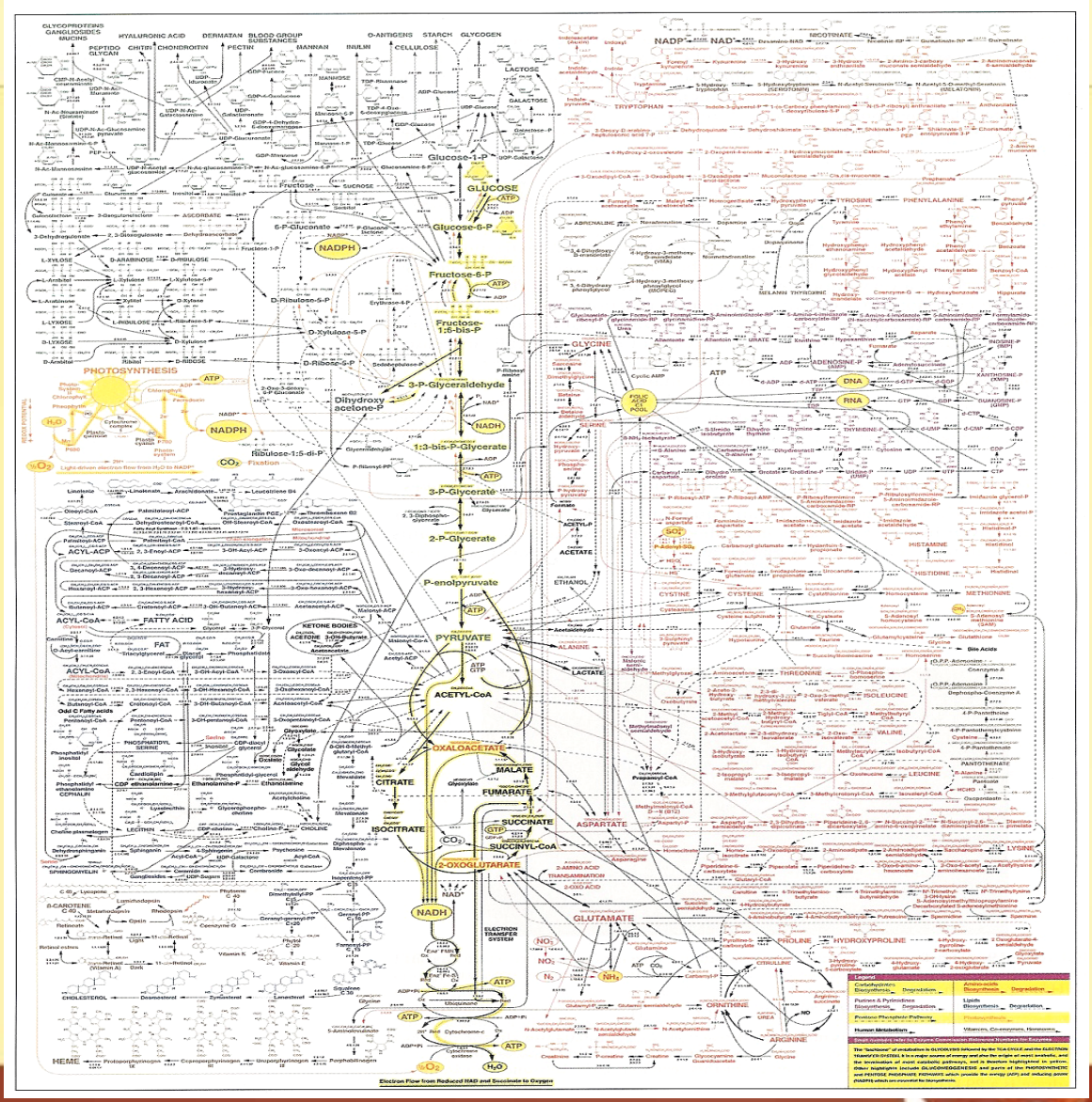
A Final Remark: Never Forget Bringing Data back to Biology



Visualize gene expression data in the context of your favorite pathways.

Download publicly available pathway maps or create your own.

The Future of Computational Biology: Can We Find It on This Map?



Never underestimate the complexity of a simple biological system.

Example:
A metabolic pathway map

Conclusions

New data acquisition technologies, such as DNA microarrays, have made molecular profiling possible. They also have created strong demands for better computational tools and systems.

To support the industrialization in molecular biology, enterprise solutions, such as the Rosetta Resolver system, have been engineered to meet the strong demand.

With these solutions, biologists can answer many challenging questions during molecular profiling. Many data-driven computational methods help biologists gain new knowledge in pharmaceutical and other bio-technology research areas.