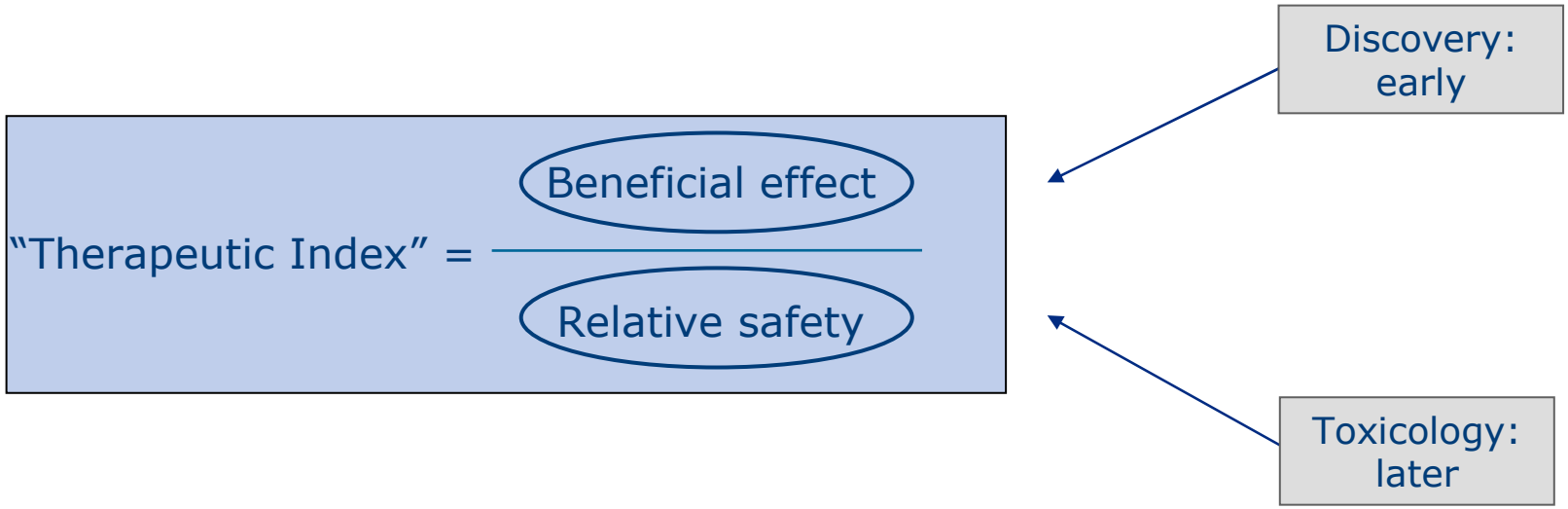
Hans Gmuender Scientific Consultant

Application of Microarrays to Toxicity Studies

The Challenge for Toxicogenomics

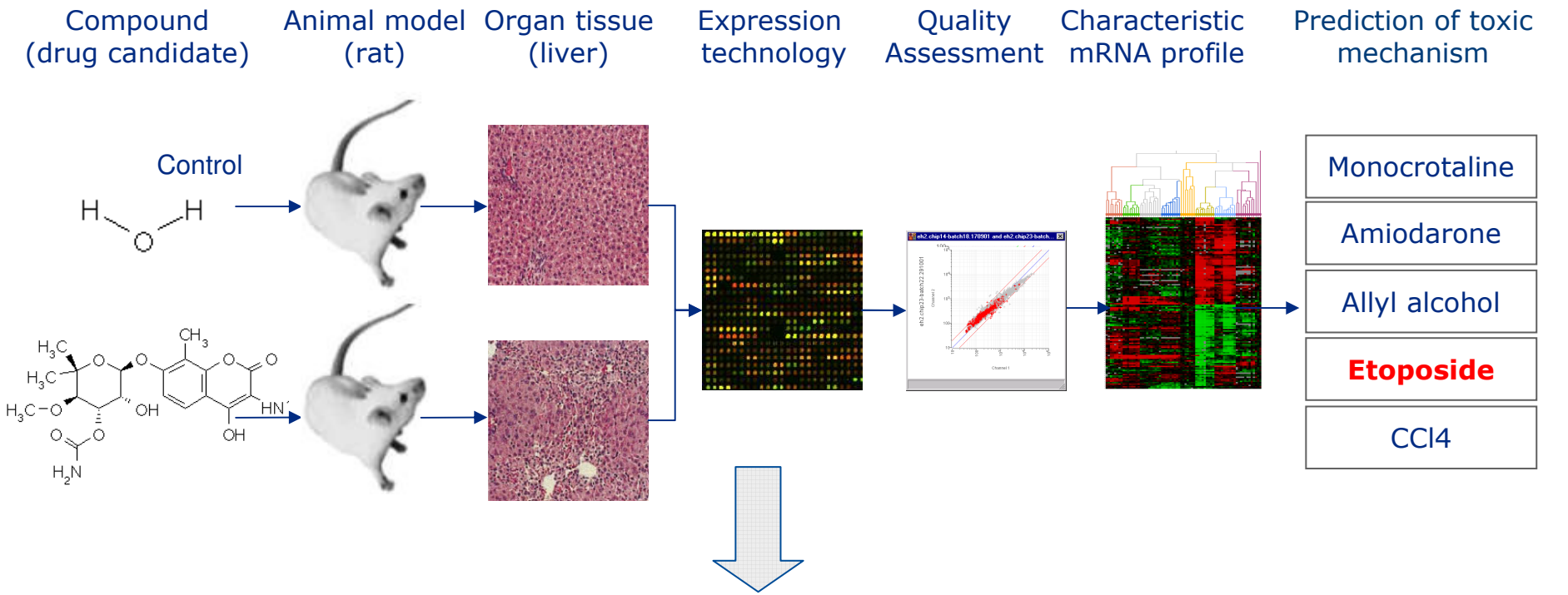


Ideally, safety and efficacy of a new drug are determined simultaneously

Enables qualified decisions for the likelihood of success early in the discovery process, before initiating costly development programs

Toxicogenomics

Early prediction of the risks associated with a given drug candidate by gene expression analysis

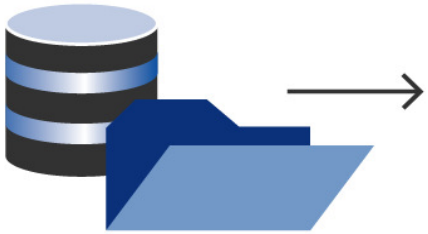


Data quality assessment → Identify toxicity marker genes → Establish toxicology reference database → Automated classification of new compounds → MOA elucidation

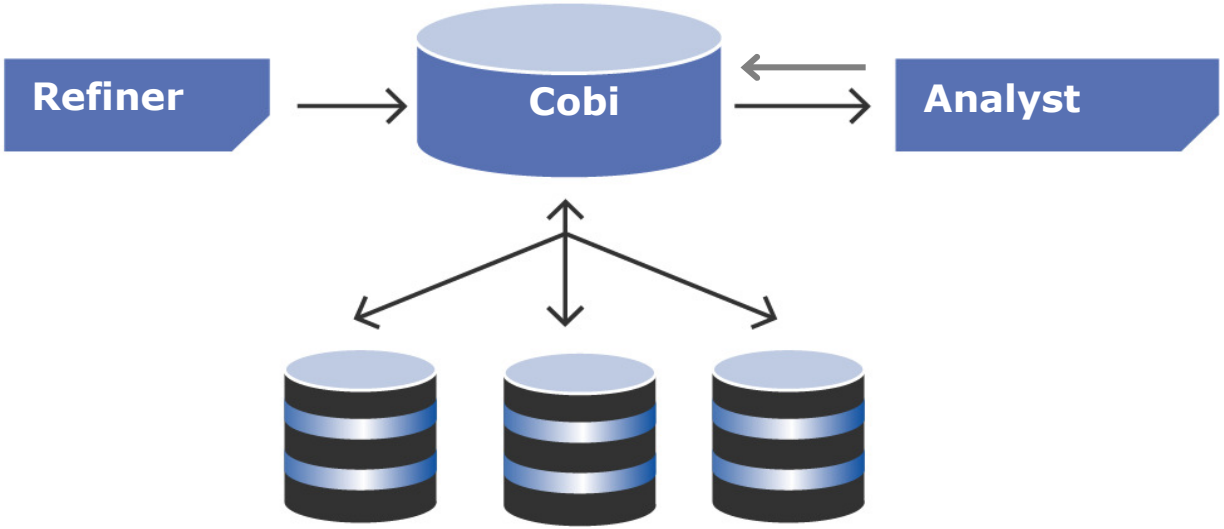
Genedata

Genedata Toxicogenomics Analysis Workflow

LIMS Image Analysis SW

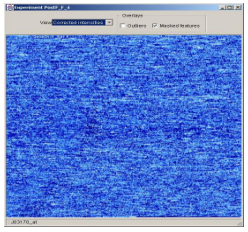


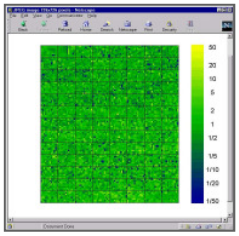
Raw Data

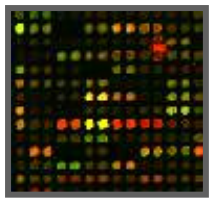


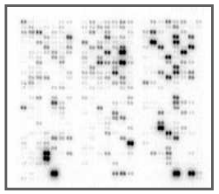
Proprietary & Public Databases

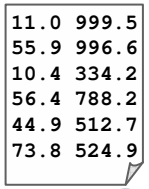
Synthesized Oligos	Spotted Oligos	Spotted cDNA	Filters	Generic (RT- PCR)







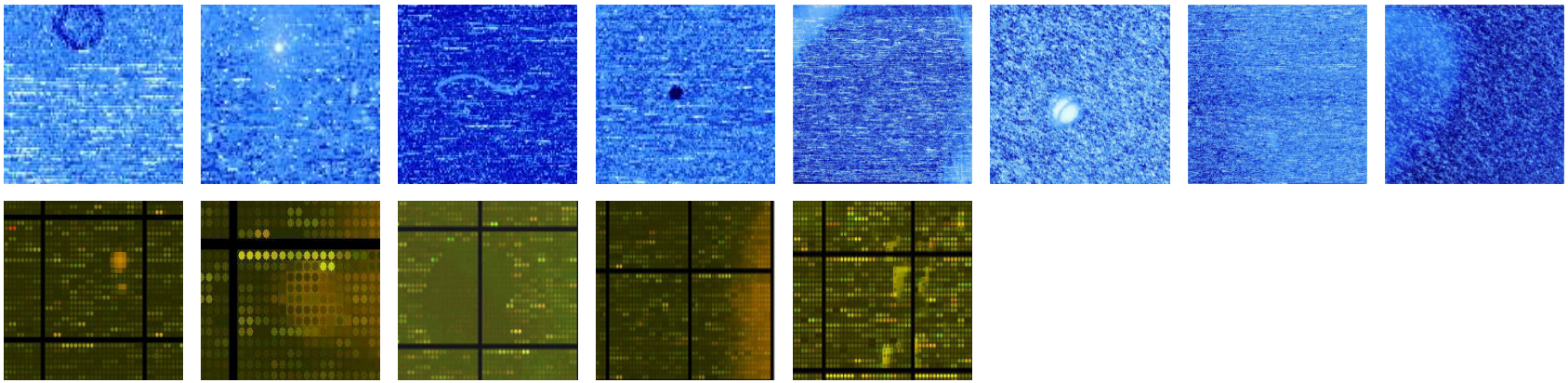




Genedata

Data Quality Assessment

These issues occur in few experiments only – but they do occur!



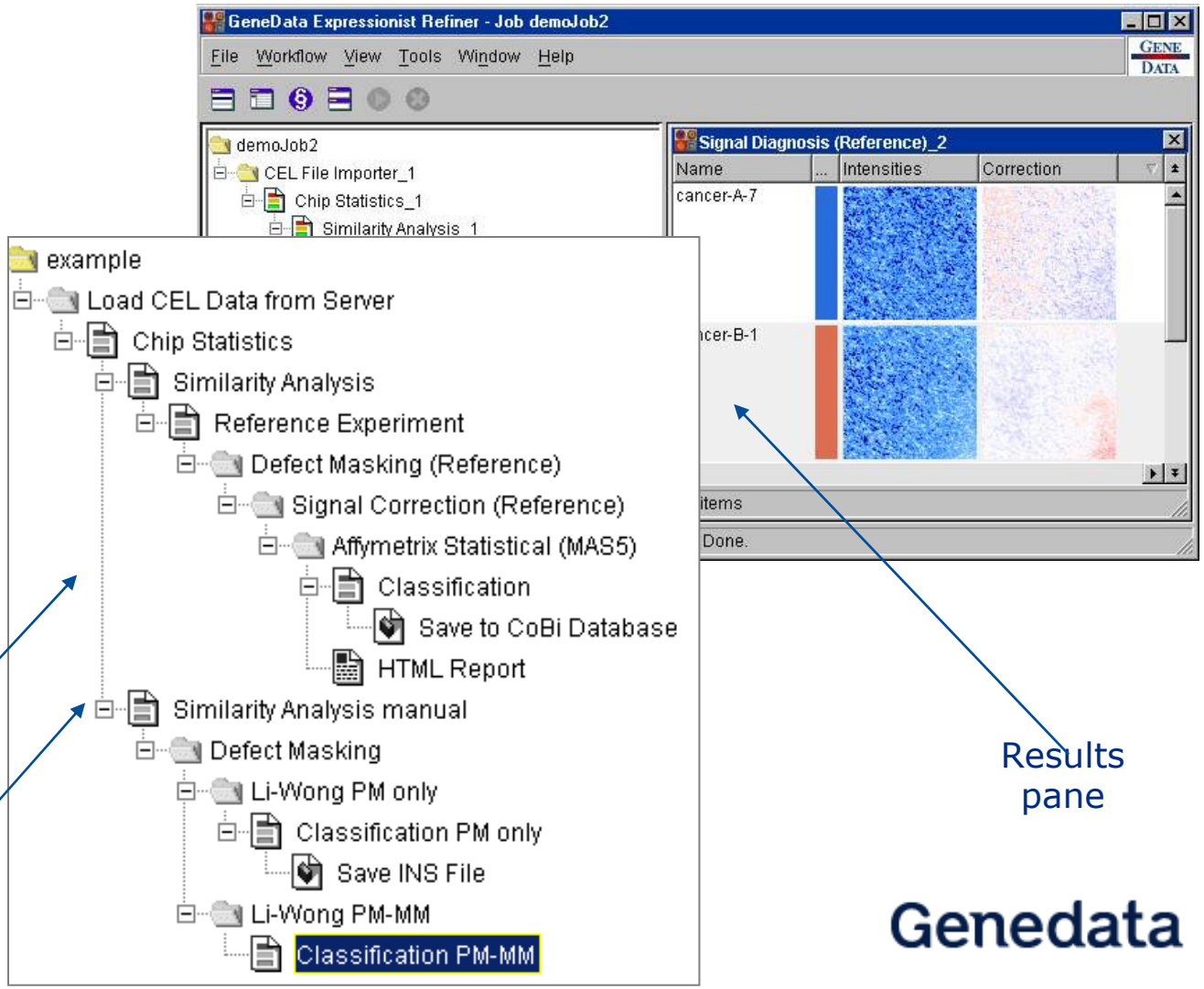
Genedata

Refiner Workflows

Workflows are assembled from a choice of actions

Available for both relative and absolute data

- Import
- Diagnostic
- Correction
- Visualization
- Reporting
- Condensing
- Export



Workflow tree

Individual activity

Results pane

Genedata

Refiner Diagnosis

One channel technology

Similarity analysis

Hierarchical clustering of experiments by their dissimilarity

Reference experiment

Computes an average experiment from groups of experiments (‘virtual chip’)

Defective area

Detection of defective areas (too-bright or too-dark spots)

Gradient diagnosis

Determines the intensity gradient across the surface

Signal diagnosis (reference)

Computes the distortion, i.e. signal dependent systematics

Two channel technology

Distortion

Identifies the deviation of the data points from the straight line

Imbalance

Adjusts the relative brightness of red and green channel

Contrast

Identifies the percentage of features whose S/N ratio is below 2

Defective area

Estimates the percentage of potentially defective area on the array from the background information

Refiner Corrections

One channel technology

Defect masking

- Interactive masking of defects identified through visual inspection or
- Automatic masking of identified defects based on a comparison against a reference experiment

Gradient correction (individual)

- Corrects a gradient present in the average intensity of the low expressed genes

Signal correction

- Corrects non-linear signal responses with respect to the reference

Two channel technology

Normalization

- Applied to balance the sample and reference channels

Background correction

- Applied to eliminate the additive effect of the background signal

Lowess correction

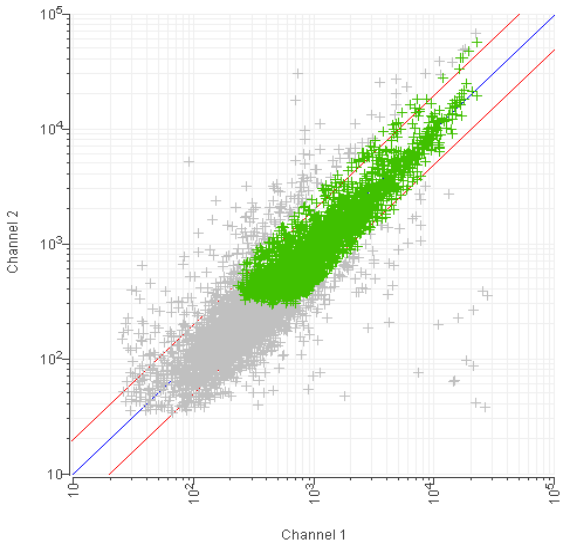
- A locally weighted correction applied to improve signal dependent distortion

Masking

- Automatically flags defective features and excludes them from processing

Normalization, LOWESS Correction

Normalization

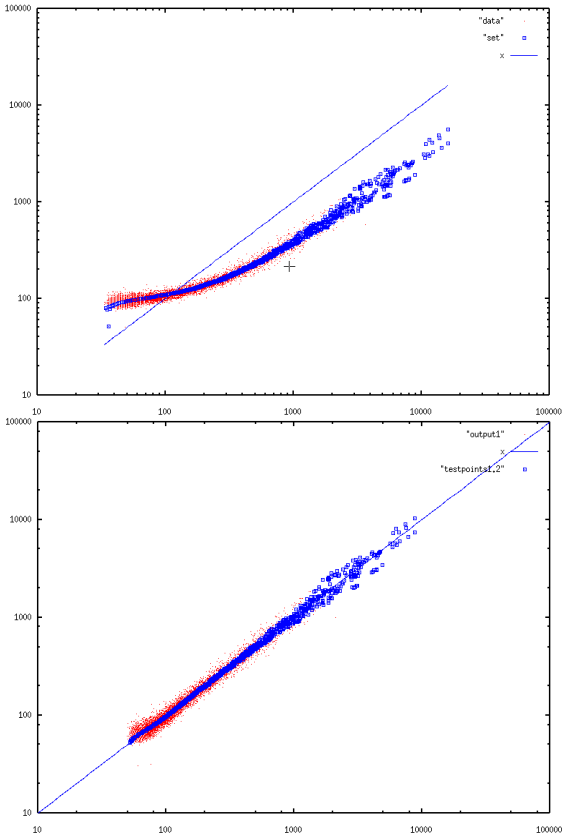


S/N > 2

S/N < 2

LOWESS correction

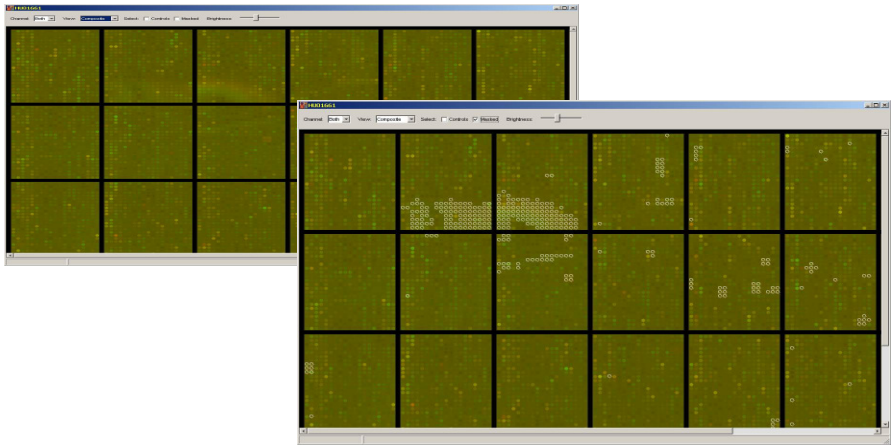
locally weighted least mean squares scatter plot smoother is a statistical algorithm for finding non-linear fit to large data sets

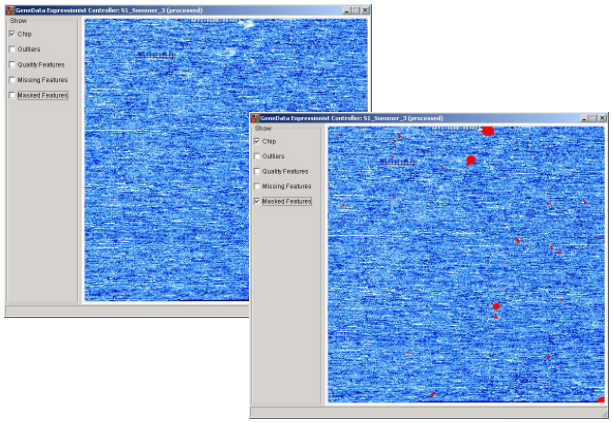


Genedata

Defect Detection and Masking

This activity masks the dark and (optionally) saturated defects to exclude them from further processing. The algorithms identify local clusters of extremely low and extremely high signals as being defective and masks them.

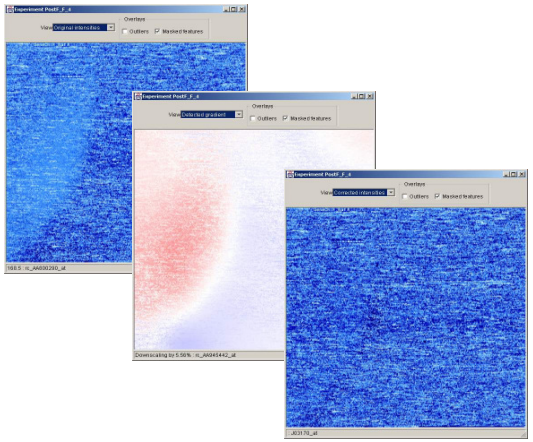


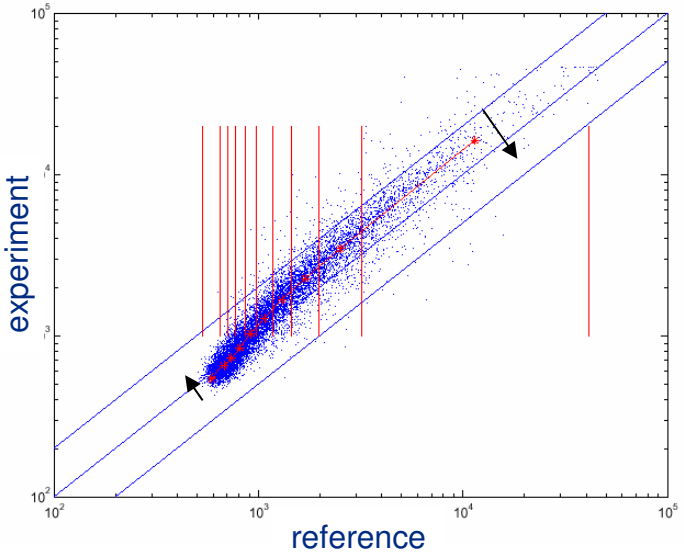


Compares the current experiment against the corresponding 'reference', identifies systematic deviations as defects and immediately masks them. Both dark and bright defects are detected and masked.

Genedata

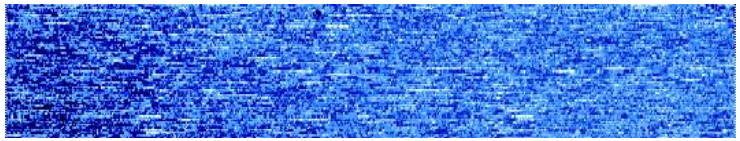
Signal & Gradient Correction





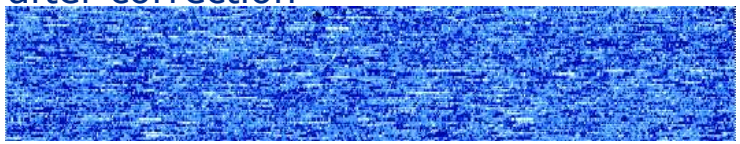
Signal Correction

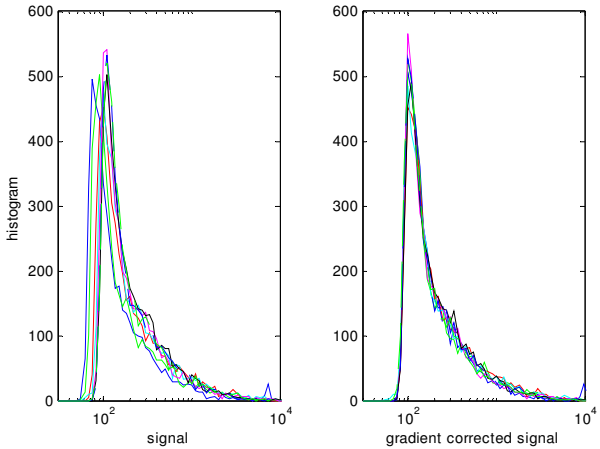
before correction



measured gradient

after correction





Gradient Correction

Genedata

Diagnostics / The “Traffic Light” System

Good

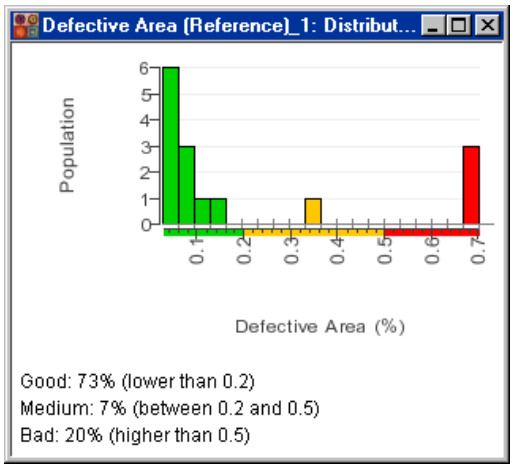
- No significant quality problems found

Medium

- Defects found
- Manual inspection is suggested

Bad

- Chips show serious defects
- Should only be included for further processing after manual inspection



Histogram of diagnostic values

Diagnostics _1

Experiment	Classification	Distortion	Imbalance	Contrast	Defective Area
HU01655		0.134	0.547	41.4	6.91
HU01661		0.125	0.254	70.7	3.02
HU01663		0.117	0.363	36.5	0.959
HU01665		0.103	0.421	36.9	4.06
HU01667		0.091	0.308	32.2	1.93
HU01669		0.141	0.0881	73.4	2.82
HU01688		0.116	0.0555	50.8	2.34
HU01689		0.113	0.066		
HU01690		0.16	0.09		
HU01698		0.152	0.12		
HU01699		0.182	0.06		
HU01700		0.193	1.1	51.6	4.41
HU01701		0.207	0.125	64.8	3.16
HU01702		0.144	0.0552	59.8	5.26
HU01704		0.052	0.457	68	1.17
HU01705		0.418	0.391	55.1	2.99
HU01706		0.111	0.766	81.4	2.5
HU01707		0.0798	0.674	57.9	3.29
HU01708		0.159	0.00106	71.5	5.99
HU01709		0.0752	0.00333	31.7	1.93

Before Correction

Diagnostics _3

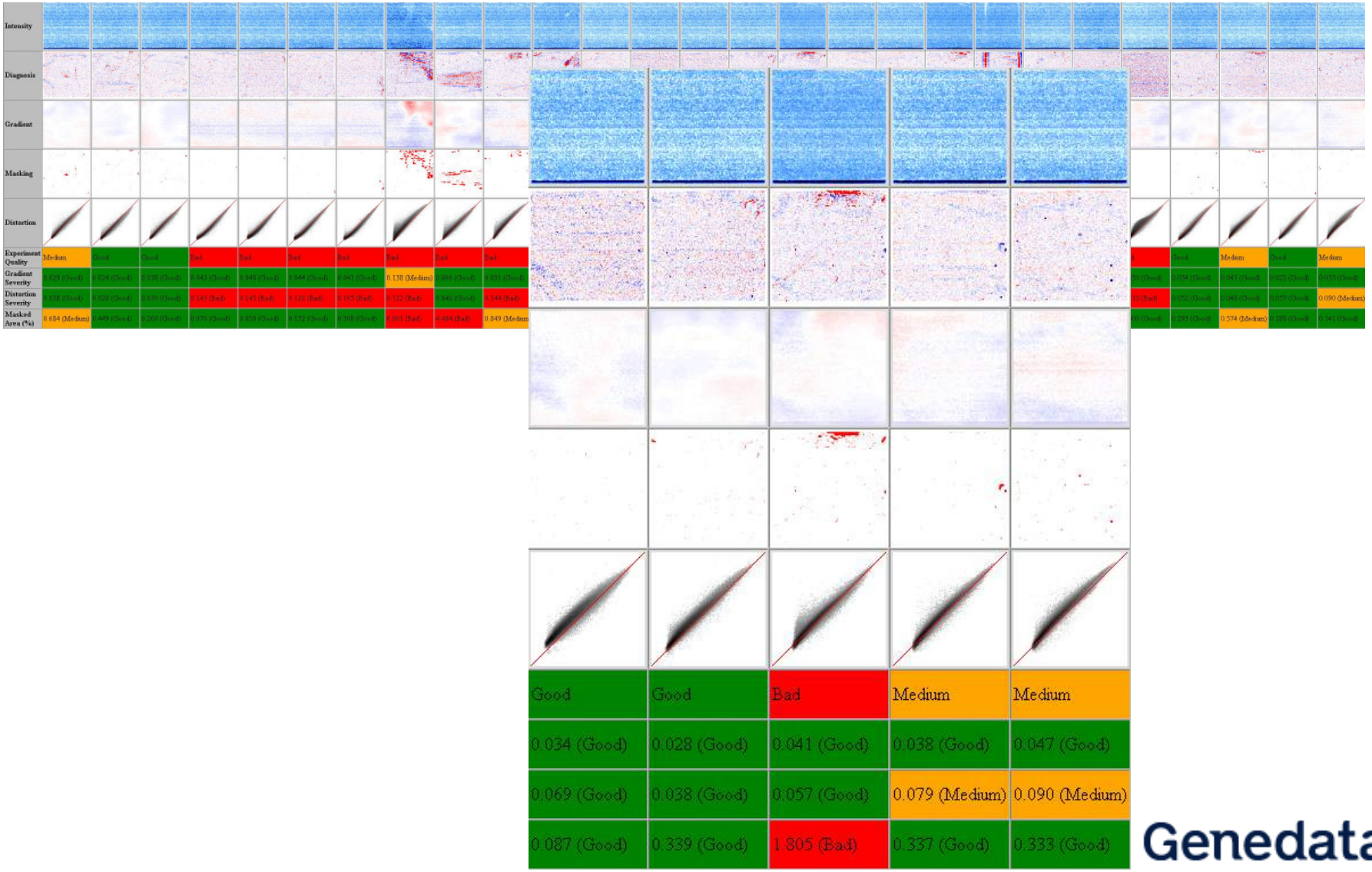
Experiment	Classification	Distortion	Imbalance	Contrast	Defective Area
HU01655		0.125	0.0425	41.4	6.91
HU01661		0.183	0.148	70.7	3.02
HU01663		0.168	0.107	36.5	0.959
HU01665		0.157	0.047	36.9	4.06
HU01667		0.123	0.0685	32.2	1.93
HU01669		0.127	0.0318	73.4	2.82
HU01688		0.0855	0.00133	50.8	2.34
HU01689		0.0689	0.01		
HU01690		0.0975	0.02		
HU01698		0.179	0.01		
HU01699		0.173	0.0765	56.8	2.93
HU01700		0.157	0.0577	51.6	4.41
HU01701		0.175	0.0707	64.8	3.16
HU01702		0.126	0.031	59.8	5.26
HU01704		0.135	0.0311	68	1.17
HU01705		0.0578	0.0489	55.1	2.99
HU01706		0.0688	0.0313	81.4	2.5
HU01707		0.0595	0.029	57.9	3.29
HU01708		0.164	0.0405	71.5	5.99
HU01709		0.106	0.0175	31.7	1.93

After Correction

Genedata

Quality Reports

Quality control reports are archived and directly accessible



Genedata

Database

Genedata

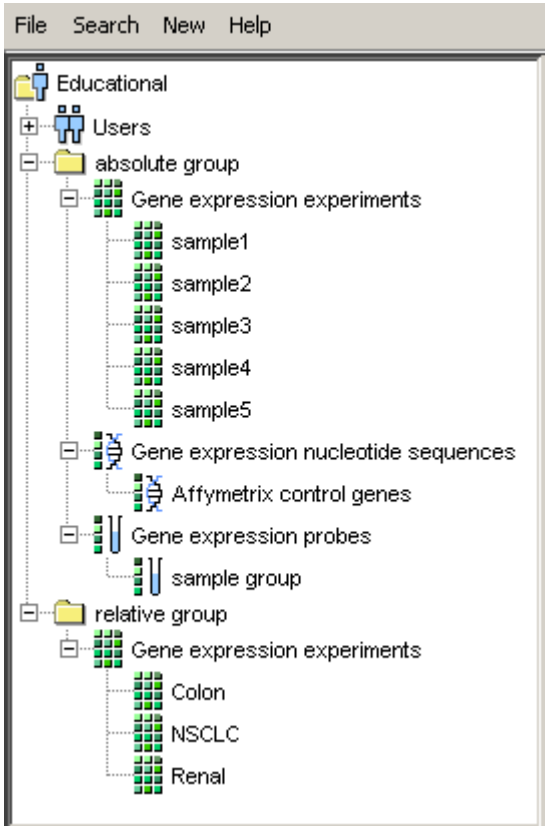
Searchable Database

Database stores the names of materials, e.g. probes and chips, used in microarray hybridizations

Database also stores data that results from applying various algorithms to the hybridizations

The data types under which you can search for data include:

- Experiment
- Hybridization
- Probe
- Chip
- Nucleotide sequence (genes)
- Groups
- Label
- Report

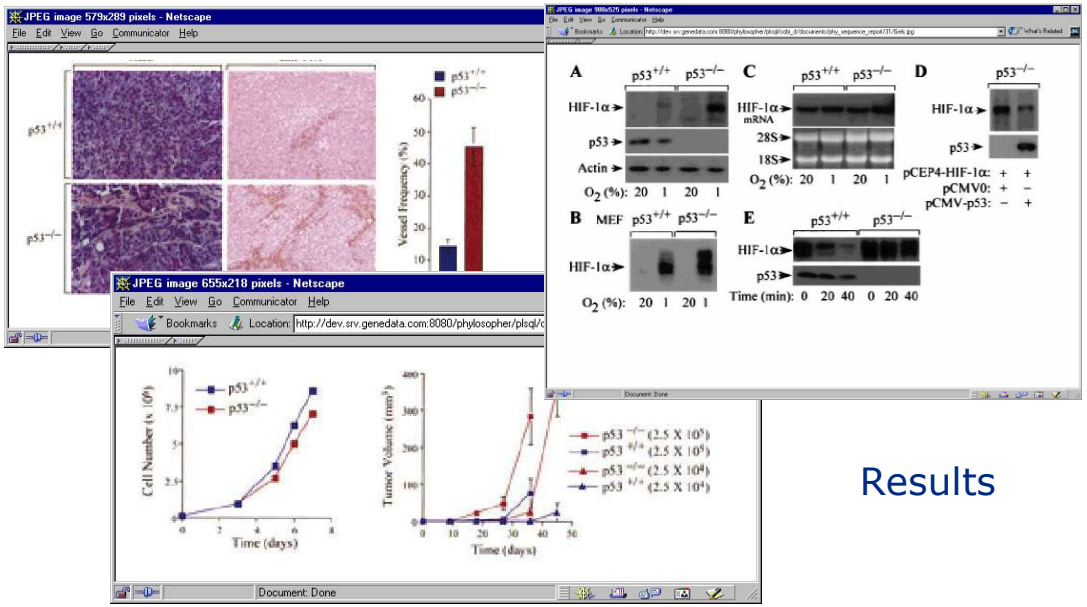


Genedata

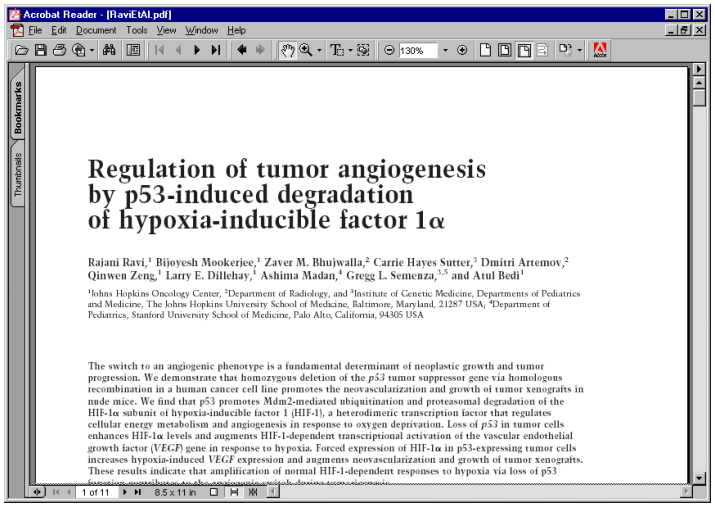
Annotation / Experimental Results / Literature

Annotation columns:

- Sequence description
- GenBank Acc.
- UniGene ID
- UniGene Title
- Gene Symbol
- Sequence Type
- Map Location
- LocusLink
- GO_Biological_Process
- GO_Molecular_Function
- GO_Cellular_Component



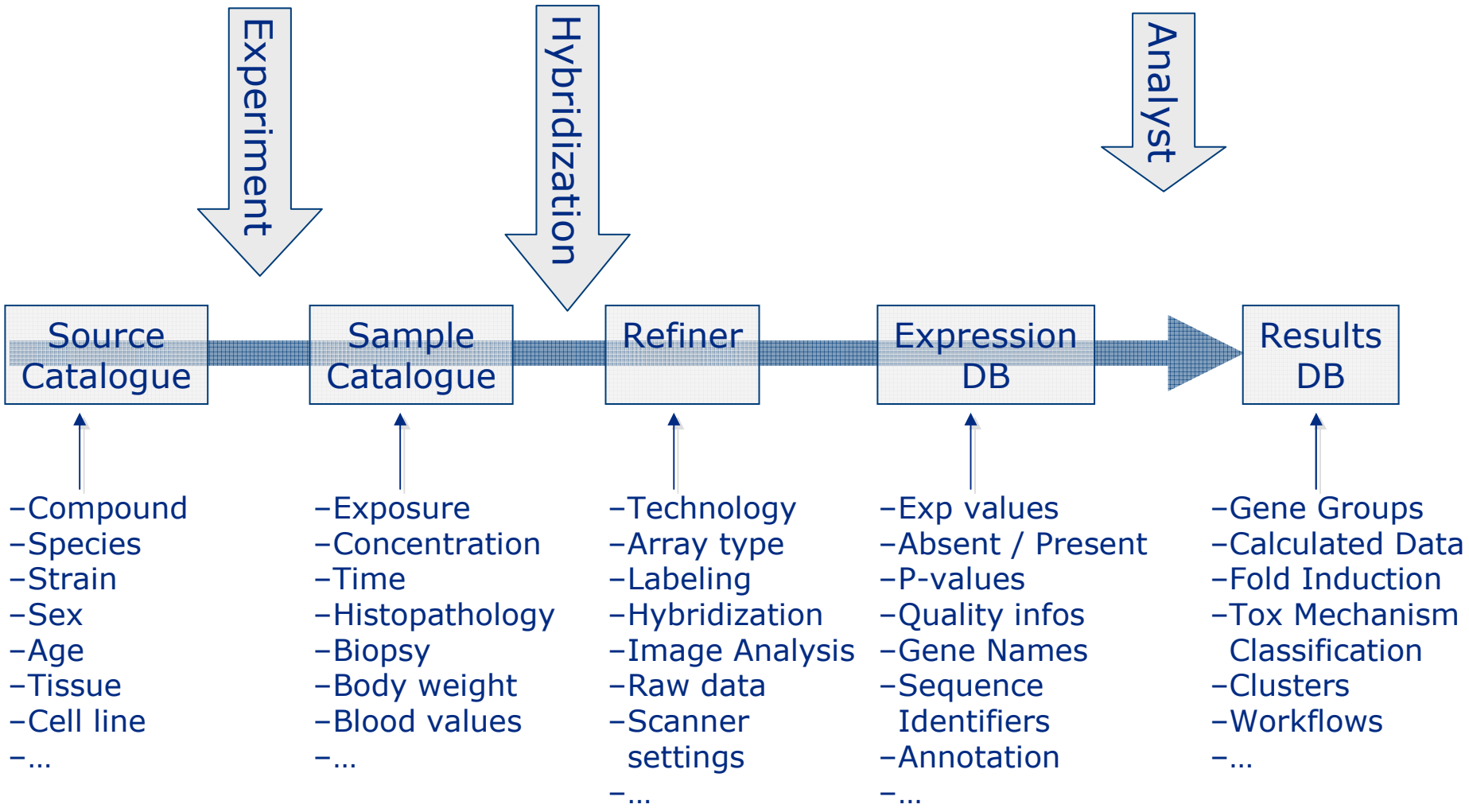
Results



Scientific Literature
Protocols

Genedata

Customization



Genedata

Significant Genes
Reference Compendium
Classification of New Compounds

Genedata

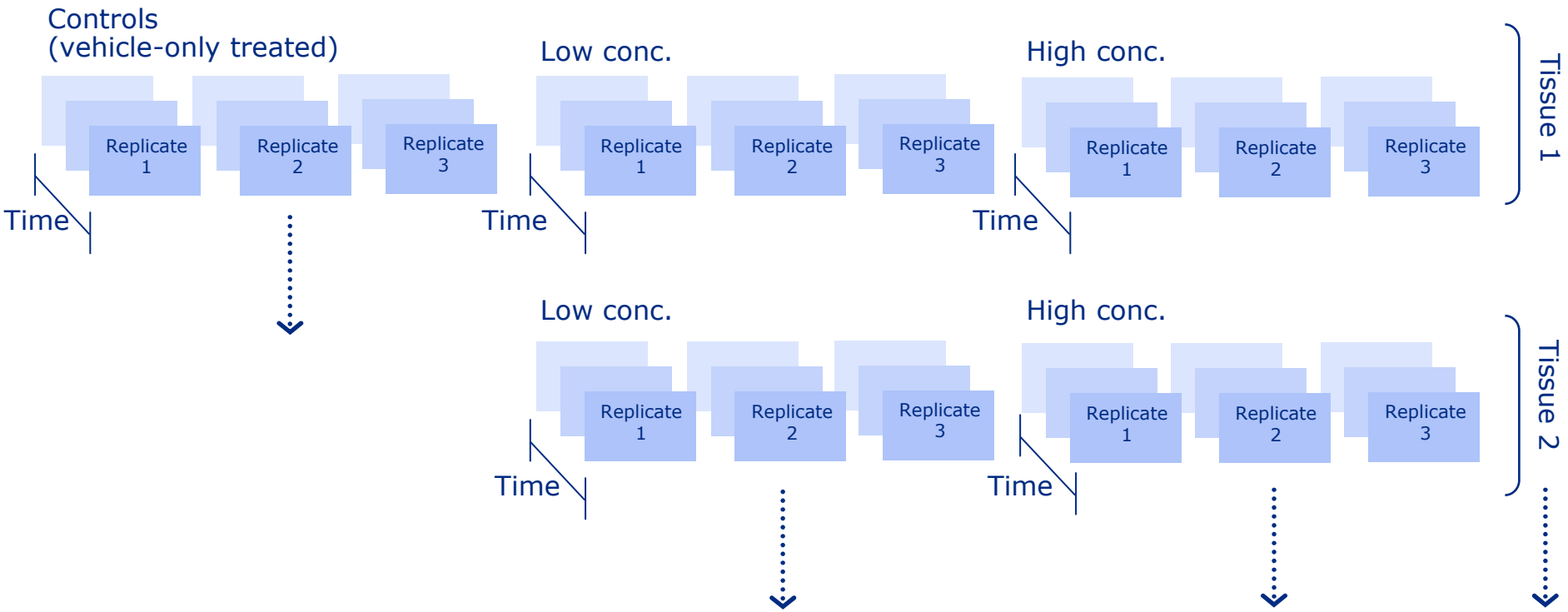
Analysis of Data

Find compound specific effects on gene expression

Find optimal gene set which discriminates the most between compound classes

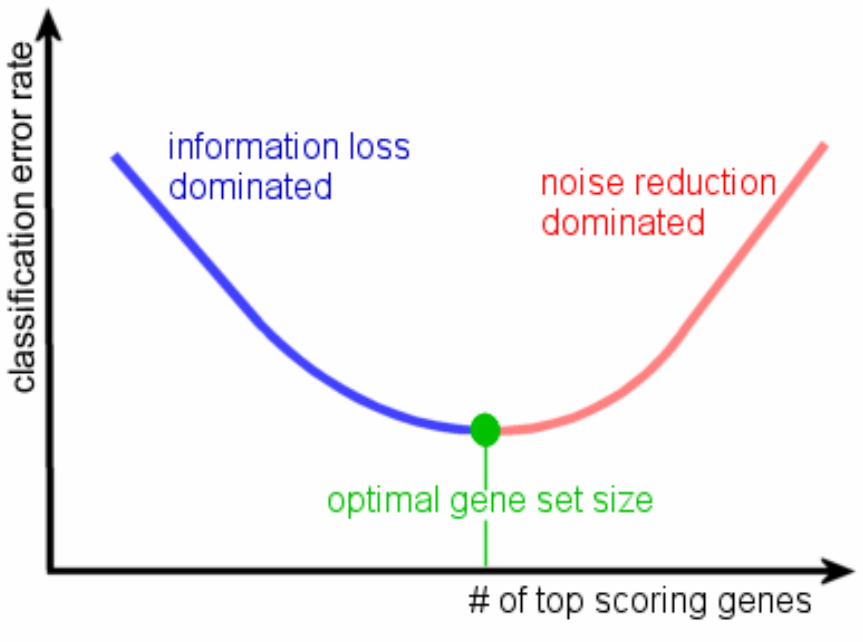
Verify classes by cross validation

Classify new compounds into reference compendium



Genedata

Identification of Significant Genes



Sketch of a typical error rate curve when removing genes

Gene ranking methods

- Support Vector Machine Weight
- Recursive Feature Elimination
- Sparse Linear Ranking
- Supervised Gene Shaving
- ANOVA
- Kruskal Wallis

Significant Genes: Statistical Identification

Computational identification of statistically significant genes

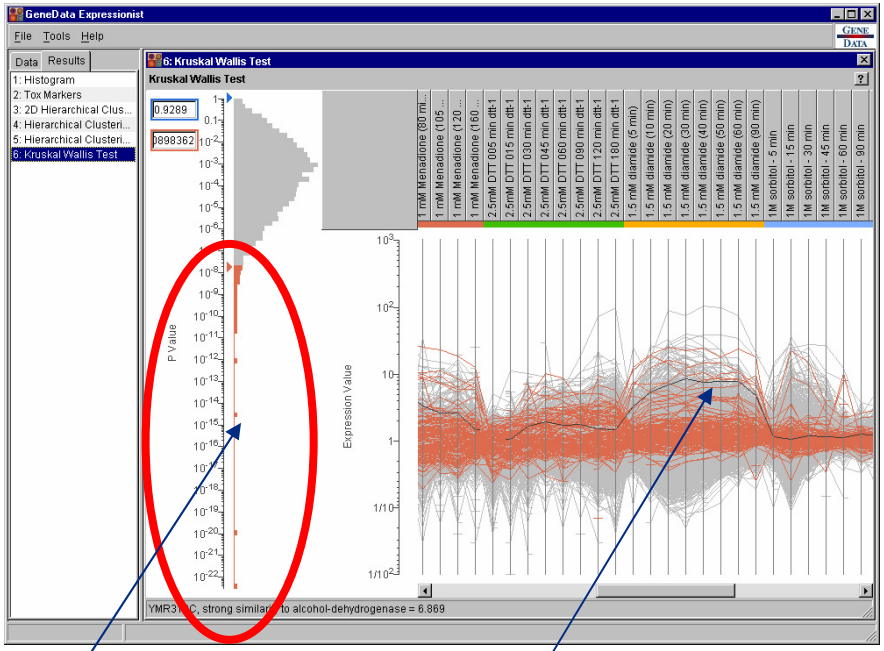
Unsupervised learning

- ANOVA (analysis of variance)
- Kruskal Wallis ranking test

Supervised learning

- SVM (support vector machine)
- Recursive Feature Elimination
- Sparse Linear Discriminant Analysis
- Supervised gene shaving

ANOVA



Gene-specific p-value reflecting the discriminative power as a marker gene

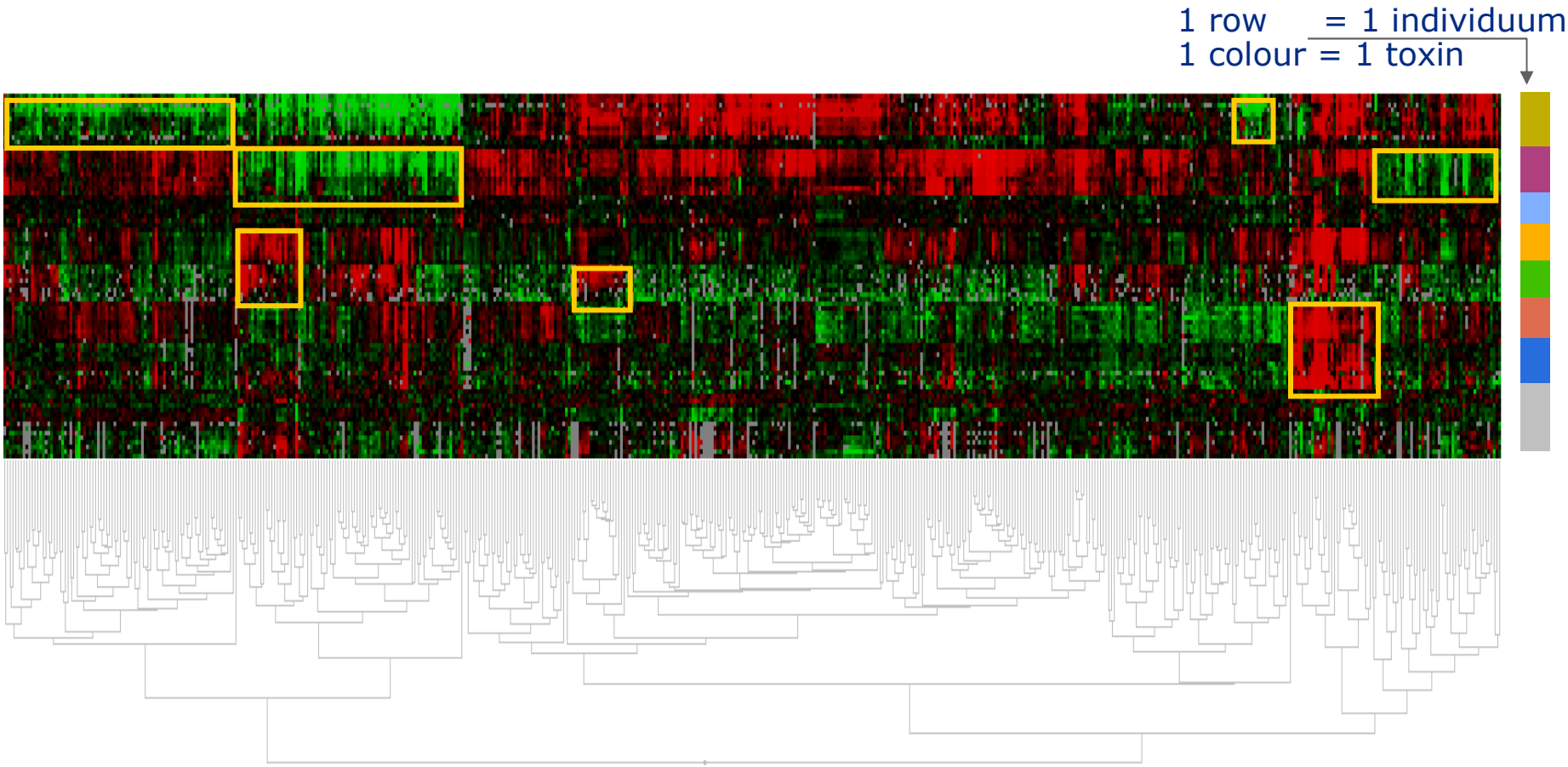
Subtle mRNA changes characteristic for a certain toxic mechanism

Genedata

Hepatoxin Induced mRNA Responses

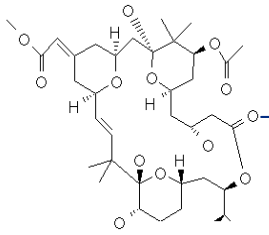
Hepatic gene expression for a selection of 8 hepatoxins (relative expression fold-factors, i.e. treated rats vs. vehicle-treated rats)

Characteristic block structure patterns indicate distinct toxicological events with a similar toxic mode of action



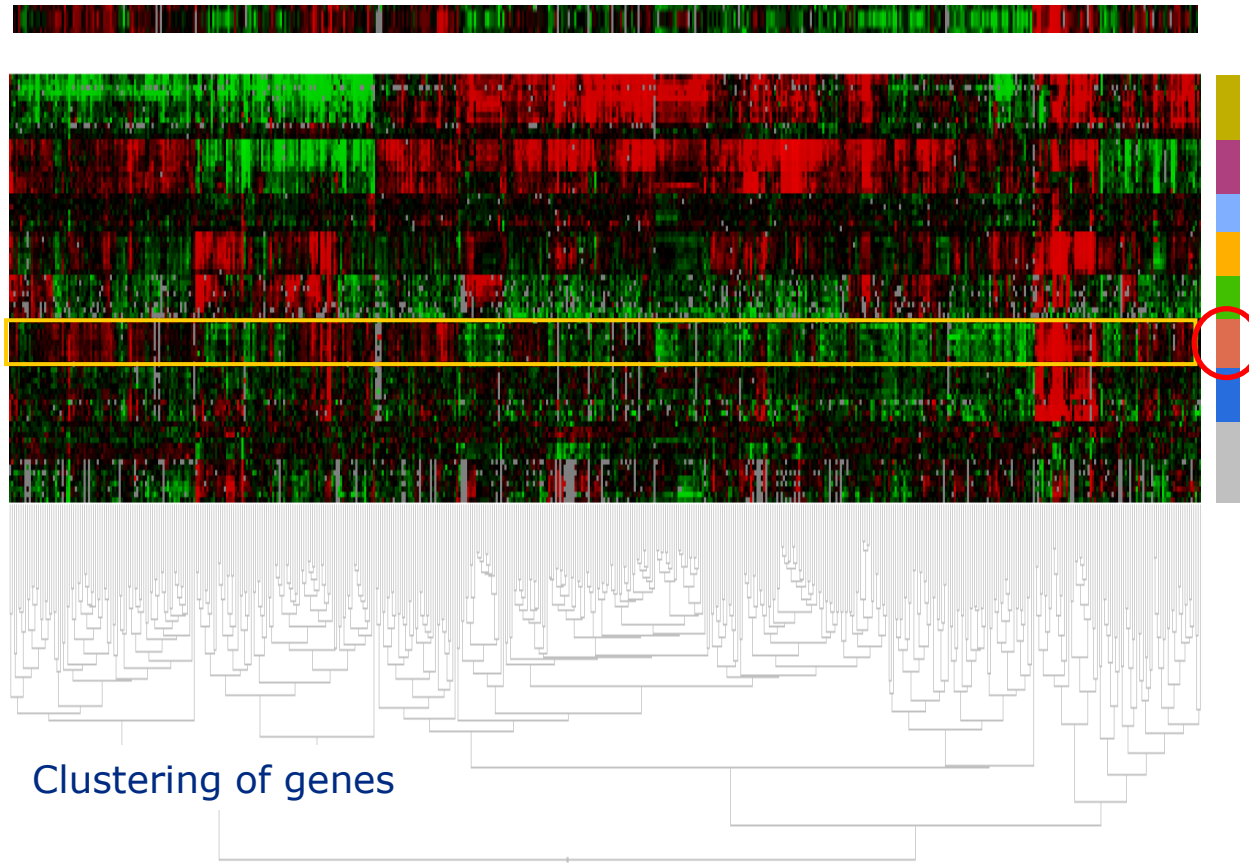
Genedata

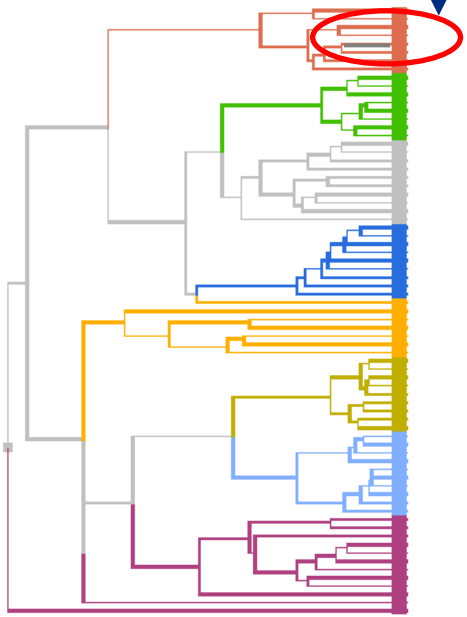
Predicted Toxicity



Drug candidate

Expression profile



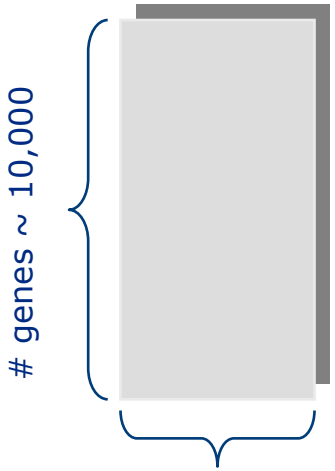


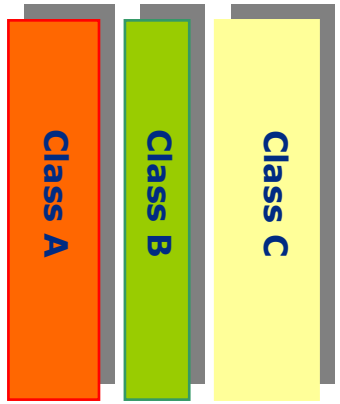
Clustering of experiments

Clustering of genes

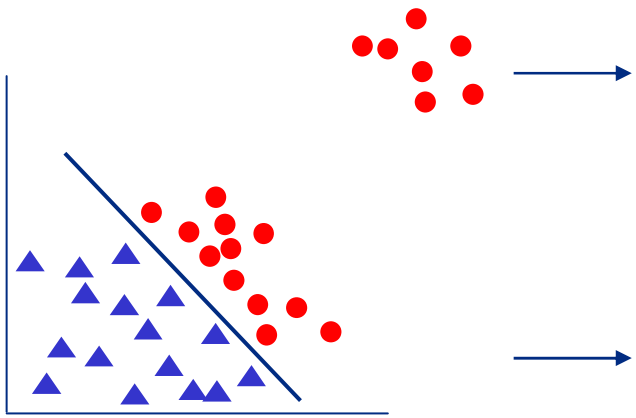
Genedata

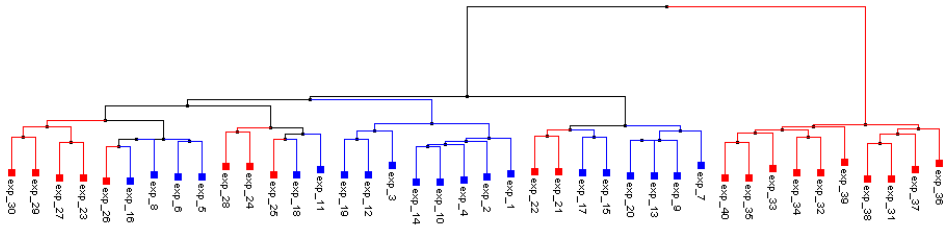
Supervised Learning





Previously determined groups

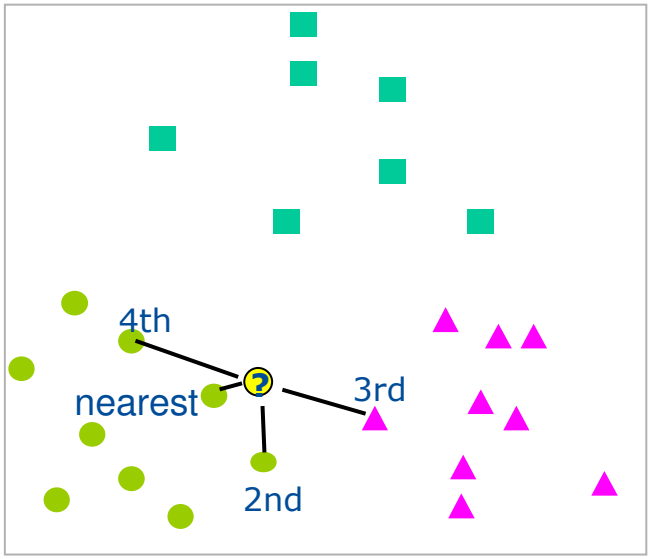




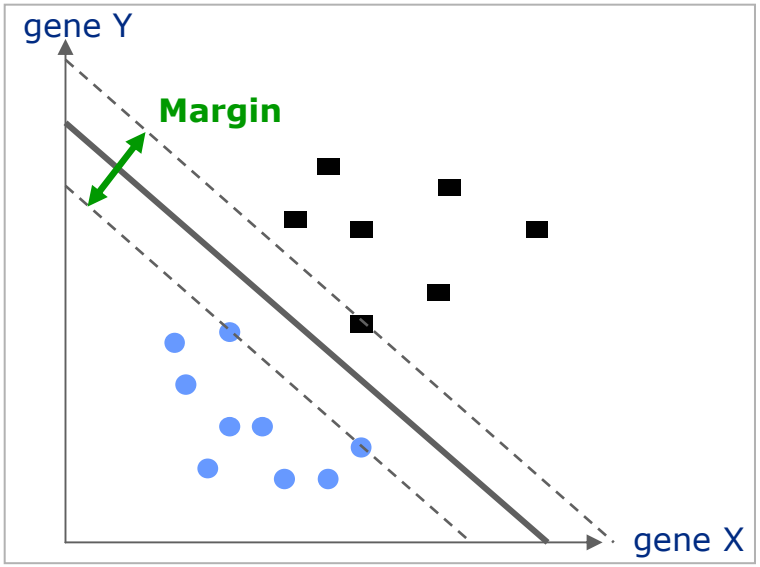
Genedata

Classifiers

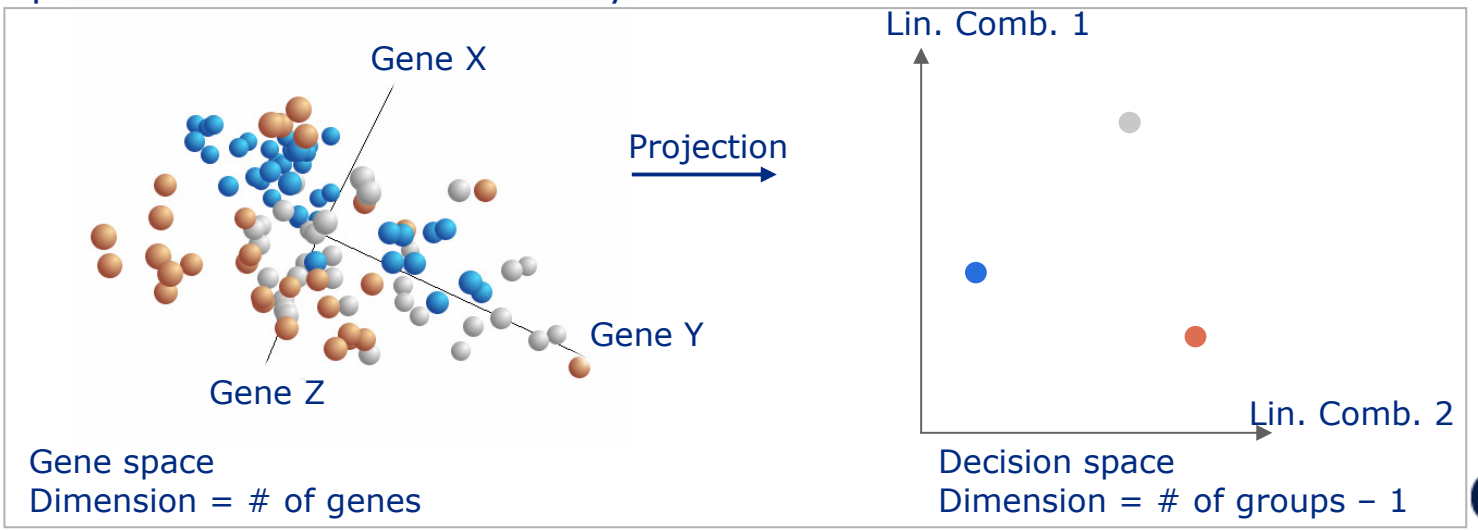
K-Nearest Neighbors (KNN)



Support Vector Machine

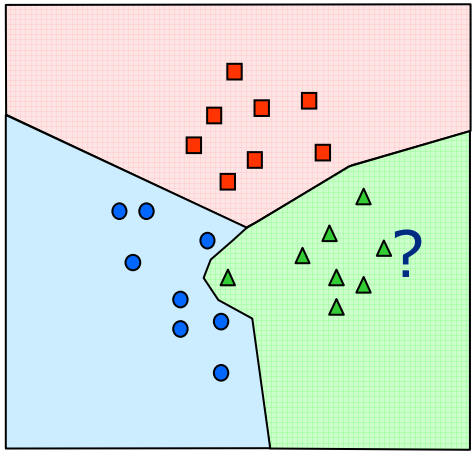


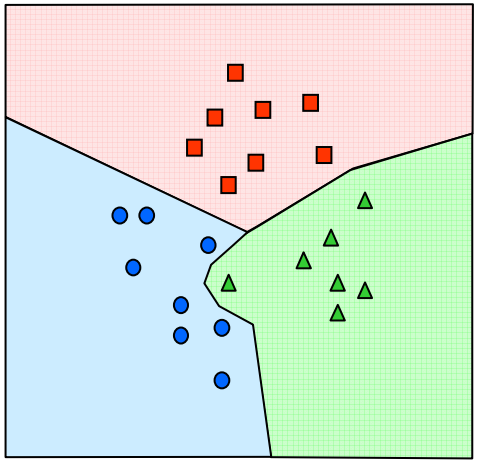
Sparse Linear Discriminant Analysis

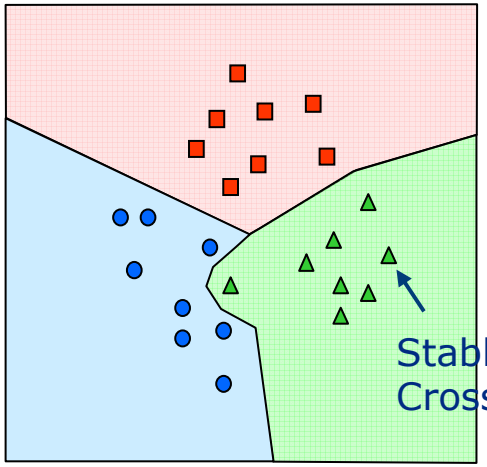


Genedata

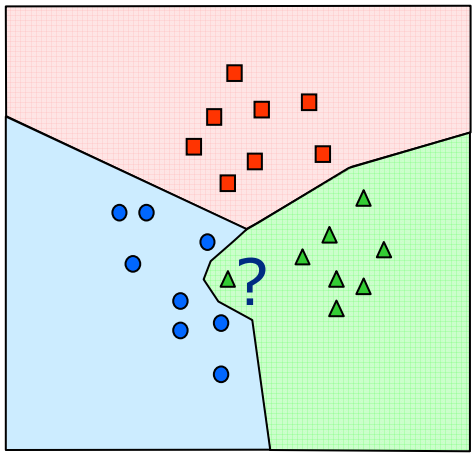
Cross Validation

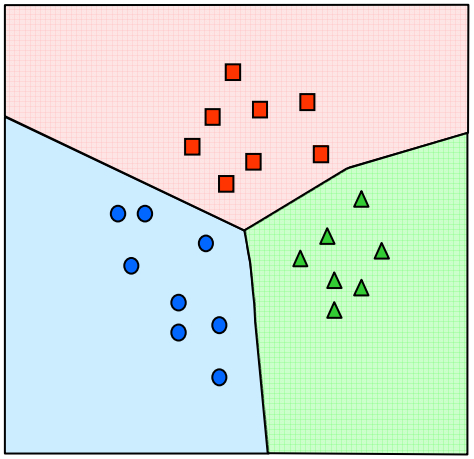


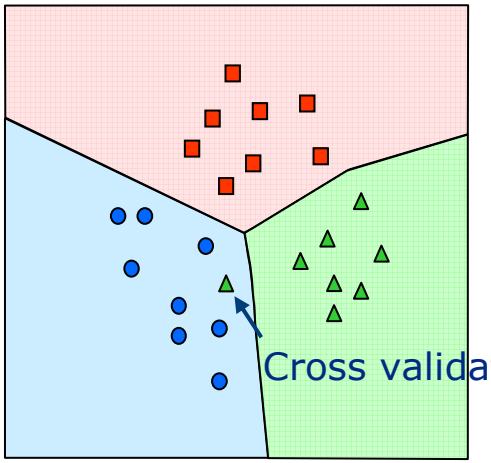




Stable under
Cross validation



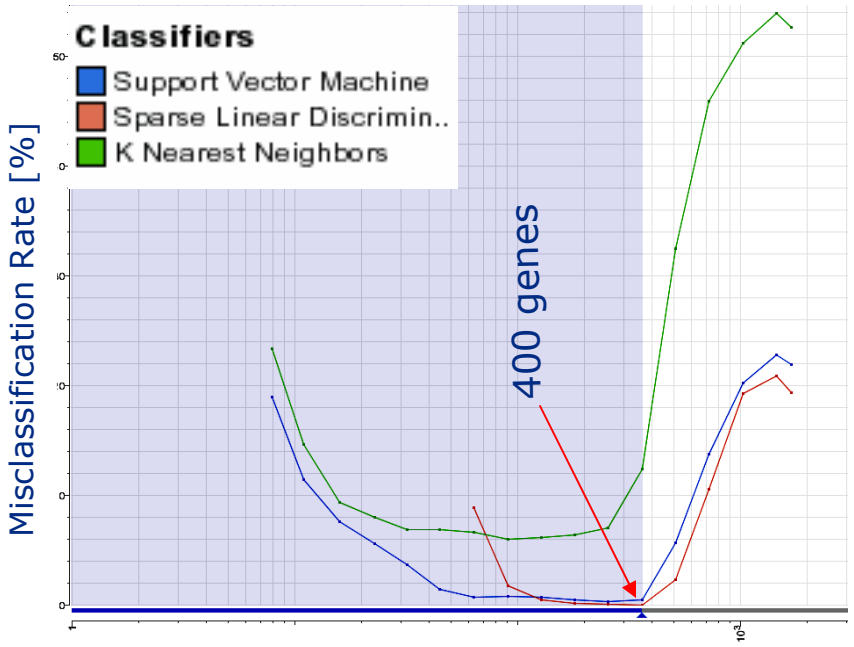




Cross validation error

Genedata

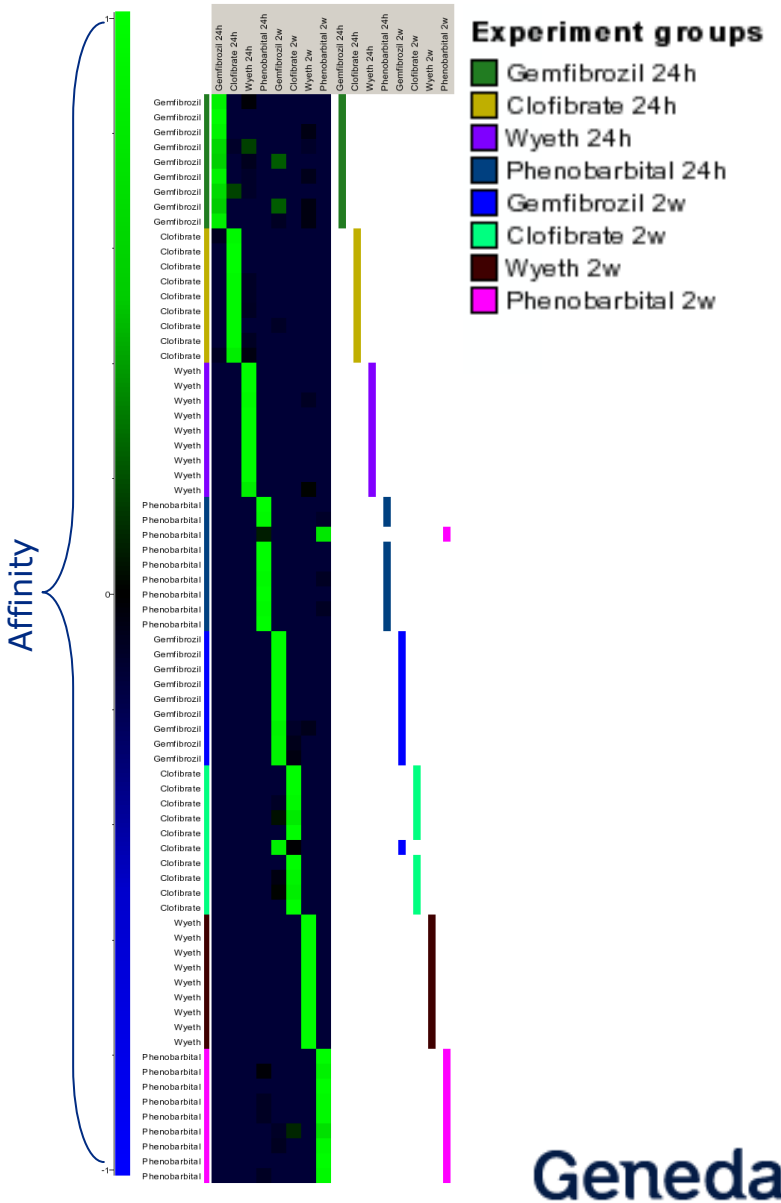
Cross Validation With Optimal Gene Set



Ranking Method: ANOVA

Genes: ~ 360 (ANOVA)
Classifier: K Nearest Neighbors
Distance: Positive Correlation, K = 1
Average cross validation error rate = 7 %

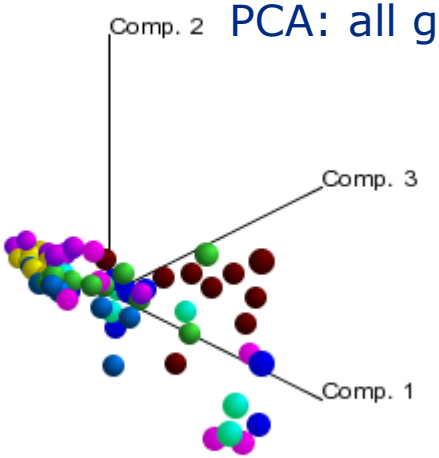
In contrast to:
Genes: all (~ 1700)
Classifier: Support Vector Machines
Average cross validation error rate = 19.89 %



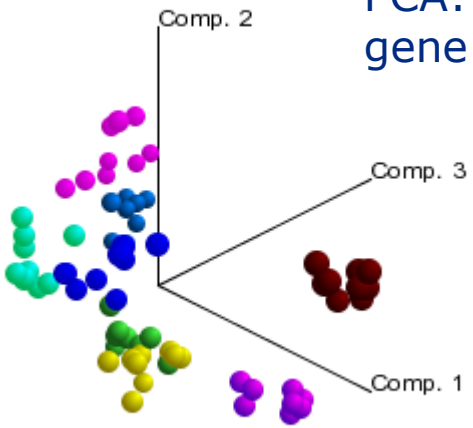
Genedata

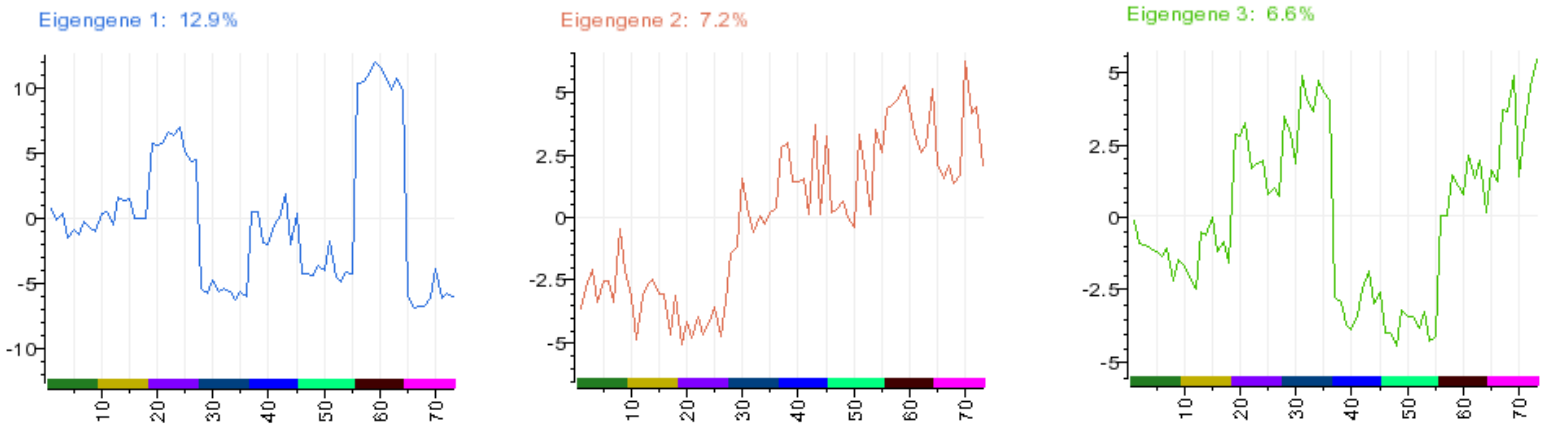
Principal Components Analysis (PCA)

PCA: all genes



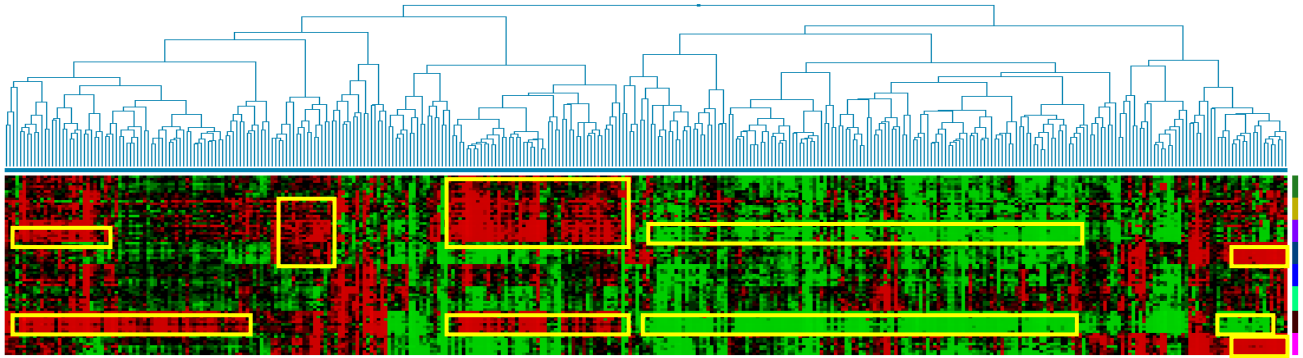
PCA: 360 genes from gene ranking with ANOVA





Genedata

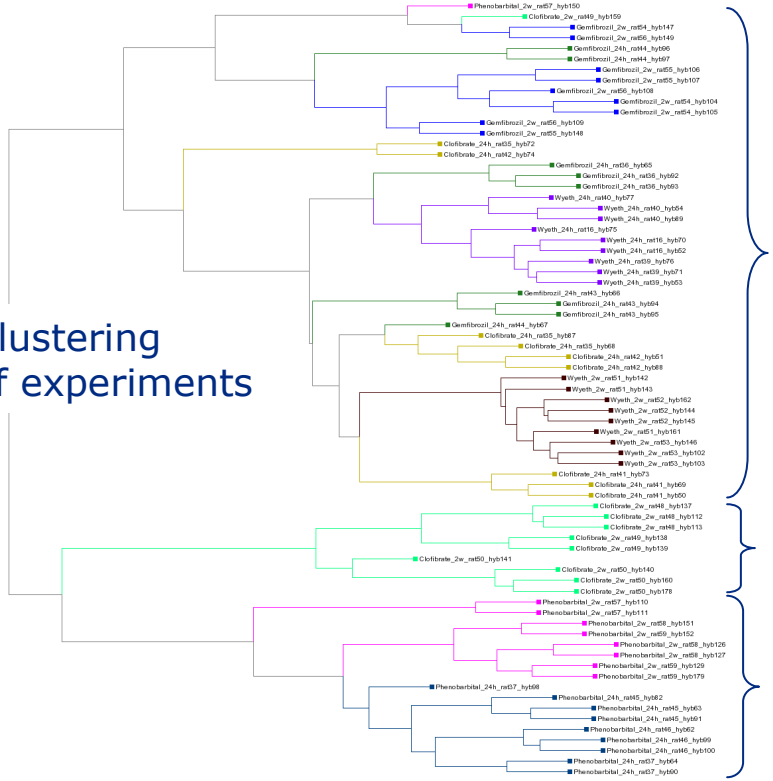
Hierarchical Clustering



Clustering of genes

Experiment groups

- Gemfibrozil 24h
- Clofibrate 24h
- Wyeth 24h
- Phenobarbital 24h
- Gemfibrozil 2w
- Clofibrate 2w
- Wyeth 2w
- Phenobarbital 2w



Clustering of experiments

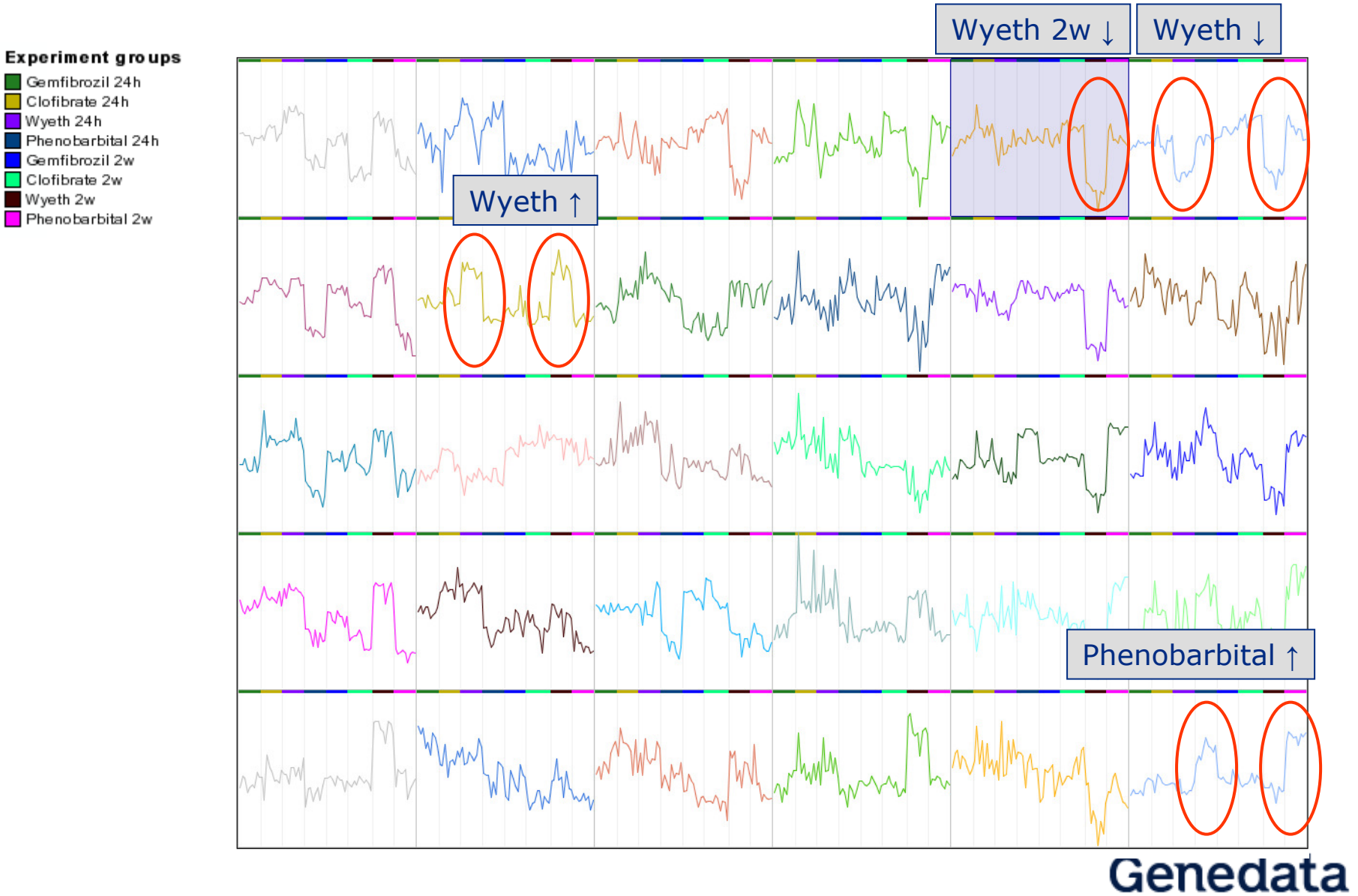
Peroxisome proliferators

Clofibrate 2w

Phenobarbital 24h and 2w

Genedata

Self Organizing Map (SOM)

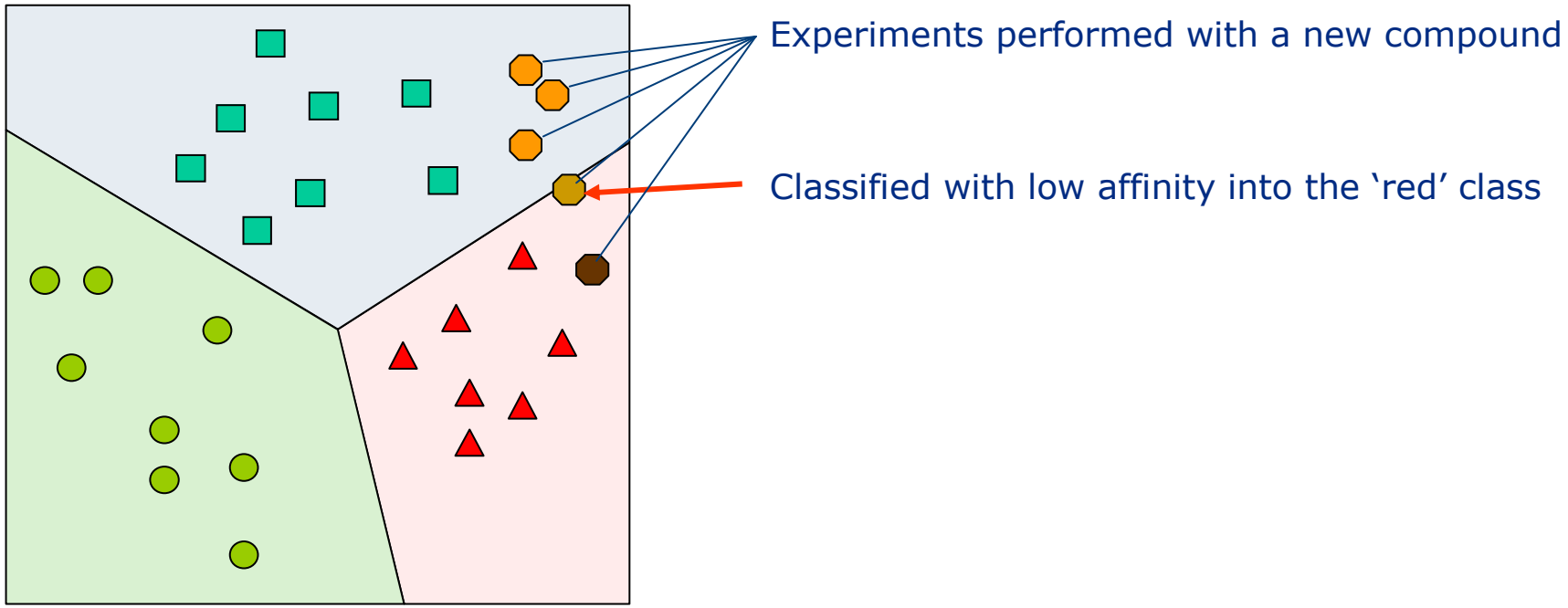


Genedata

Classification of New Experiments

The classifier first takes the reference compendium experiments and defines characteristic regions in the data space

The classifier then positions the new compound experiments in the data space

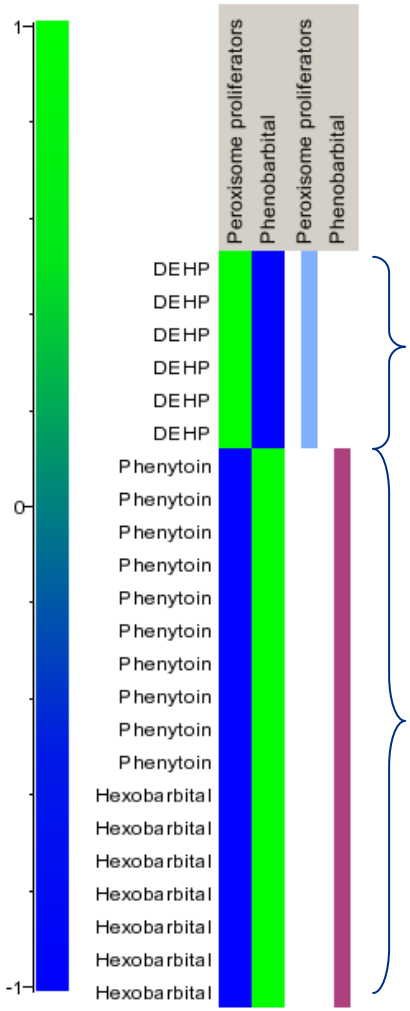


Genedata

Classification of 'Unknown' Compounds

Matrix containing measures of "affinity" for each experiment to each of the groups in the reference compendium

Experiment is assigned to the group with highest affinity



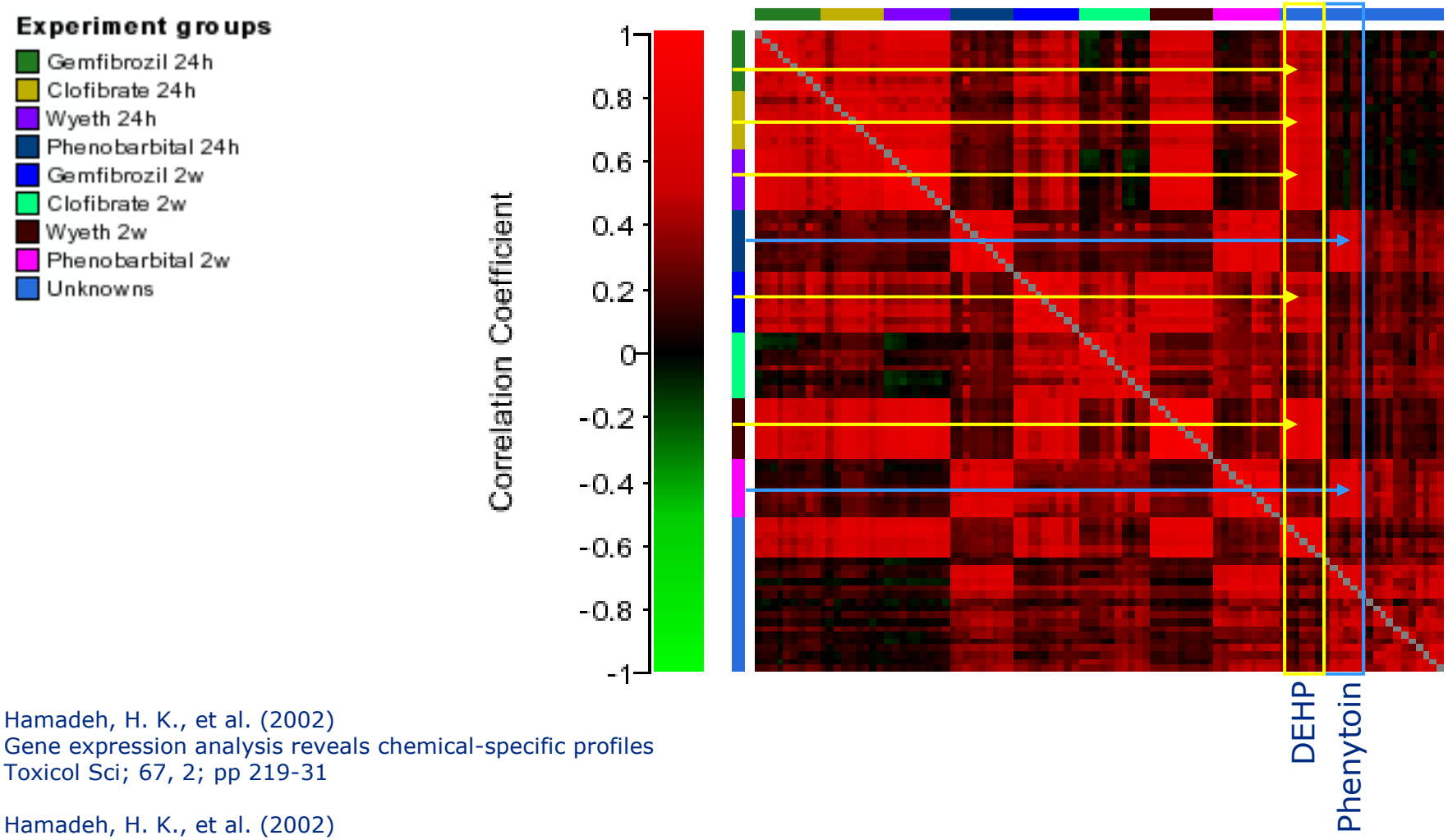
Classifier: K Nearest Neighbors
Distance: Positive Correlation
K: 1

DEHP =
peroxisome proliferator

Phenytoin and Hexobarbital =
Phenobarbital

Genedata

Experiment Correlation



Hamadeh, H. K., et al. (2002)
Gene expression analysis reveals chemical-specific profiles
Toxicol Sci; 67, 2; pp 219-31

Hamadeh, H. K., et al. (2002)
Prediction of compound signature using high density gene expression profiling
Toxicol Sci; 67, 2; pp 232-40

Genedata

Promoter Analysis of Expressed Genes

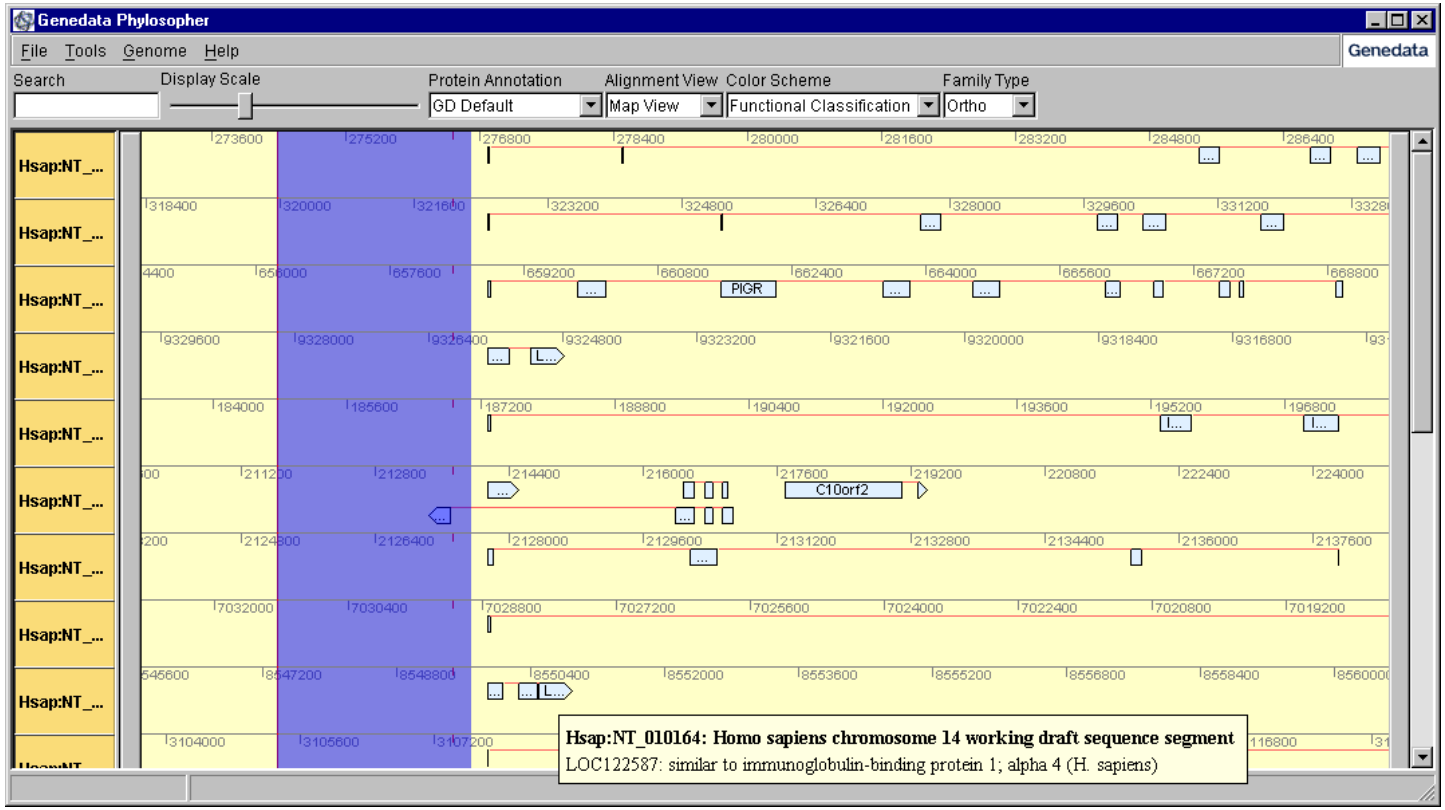
Given a set of co-expressed genes, is there a characteristic and modular structure of transcription factor binding sites?

If yes, can one identify additional putatively co-regulated genes?

Genedata

Identification of Upstream Regions

Mapping genes onto genome, align and identify upstream regions

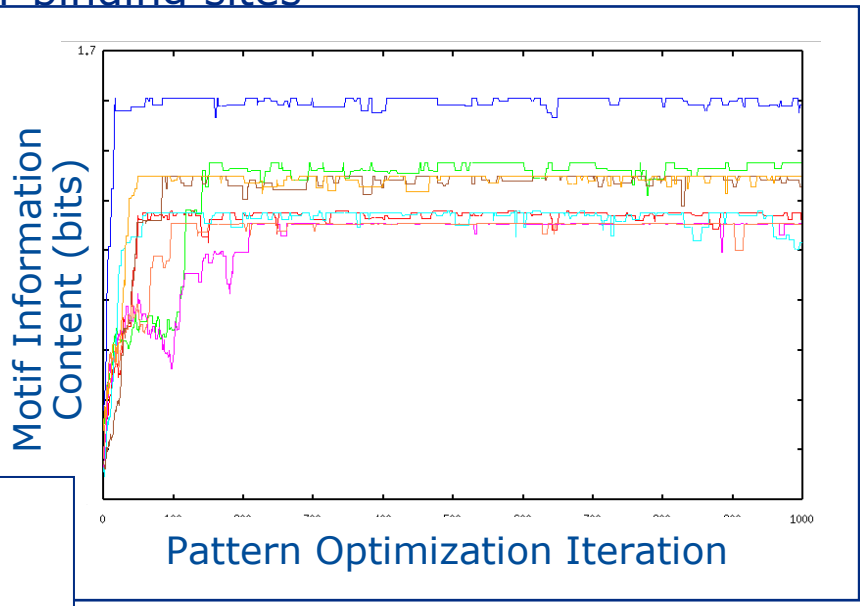


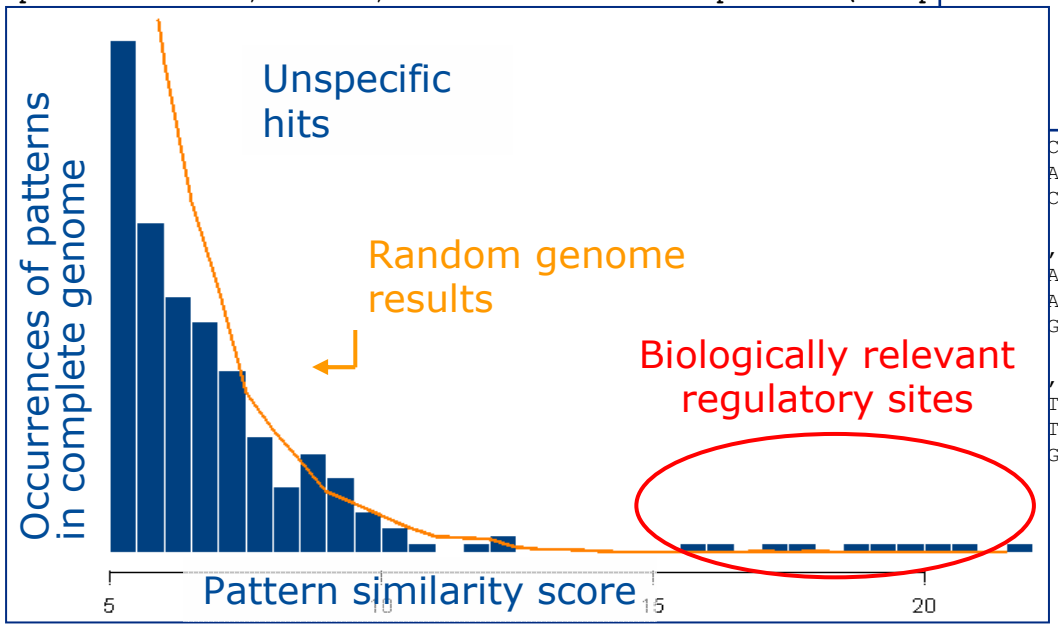
Genedata

Identification of Specific Regulatory Motifs

Identification of motifs for transcription factor binding sites

>purEK start=552323, orient=−, from start=552823 to stop=552323 (500 bp up
ATGGCGACACGCTGTGCACCGATGACGCGGGTTATCAGGCTTTTCGCGCCAAGGTCCACAAACCCTGGCTGCAC
GAAGCCAACAGCAGTAAATCGCTGGCGATCATGGACGTTAACCAAAACGCGGTGGTCAGTGCGATGGAAAAACA
ACCTGCTTTTCGCGTGGTACTG**GGTGCCTGGCATACGGAAGGTTC**AATGGTGAAAGTCACGGCGGATGACGTTC
TGCTCTCTTTCCGTGCTATTCTCTGTGCCCTCTAAAGCCGAGAGTTGTGCACCACAGGAGTTTTAAGACGCATC
>purB start=2595638, orient=−, from start=2596138 to stop=2595638 (500 bp
GTGGTTTGGCAACGTCTGCCAGCGGCGCTGGAAAAAGTGGGCATGAAAGTGACCGACAGCACCCGTTCGCAGGC
CCTGGCATCCGGTGACTATAAACTGCAGGTTGGCGATTTAGATAACCGCAGCAGCCTACAGTTCATCGATCCGA
AAAAATACAGGGCTGGAATCATCCGGCCCTTTTTTCTGATATGATACGCAAACGTGTGCGTCTGCAGGAAAACC
CCGACAAAACATCCGGCACACCAGACAGCAAAAGATTTTAAAACGTTAATTCACACCCAGGAGTGATAAAGATC
>purT start=2619217, orient=+, from start=2618717 to stop=2619217 (500 bp
CCGTTCCAGCCTTCGATAGTTACTTTTTCCGTTTCGAGGTCGGCGGTCGCTTC**GTAAGTCAGCAGGCTACCCAC**
GACGAGTGGGTGTTTGACTTCCACGATCTTCATACTCTTTCTCCTTTGAGGGGCAGCCACAAAAAAAATCGACC
AGATAAAGTTATTTTATATTCAGATGGTTATGAAAGAAGATTATTCCATCCGAAAACTAACCTTTACCCTGGCA
GTTTGCTTTCCCTGTTAGAATTGCGCCGAATTTTATTTTTCTACCGCAAGTAACGCGTGGGGACCCAAGCAGTC
>purF start=2693563, orient=−, from start=2694063 to stop=2693563 (500 bp





CAGTCGCCAGGGCTTAAACGCCGAATTAACCCGCTACACACTCAGCTTGATGGTGCTTGAG
ATTTACAGTCCGAAACGCTGATGAGCGCGATGGCTGCTATCTATGTTGATGTGATTAGCCC
CTGGCAGGCATTCGCGCCGCCGTGCTCTGGCACCAGGTCGGCGGCGGACGTCTGCAACTGA

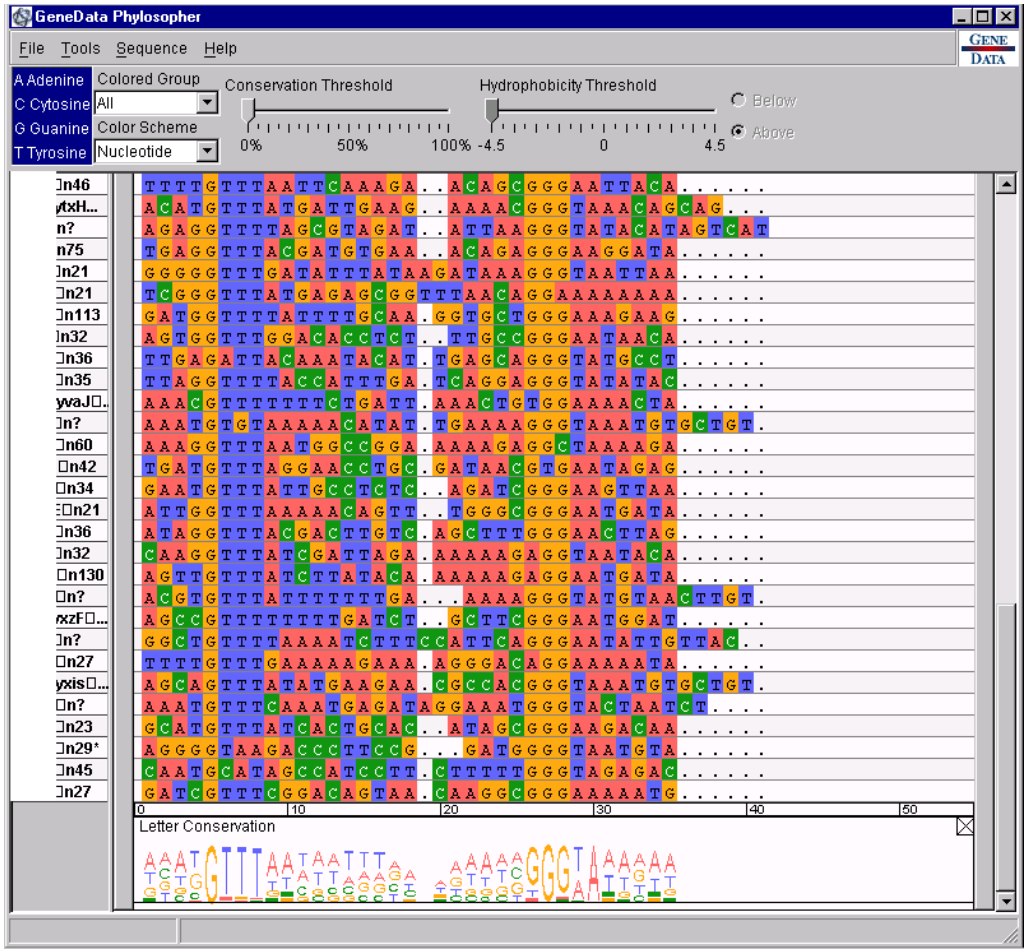
, 0 downstream)
ATATTGGCGCGAACTAAAACTTGTTGCCCATGCTGTGGGCGGAAGGTCACCCGGCGGTTGC
ATTAGAAATTTCGCCGCTGATCCAAACCTGTCCCATCTCATGCTCAAGCAGCAGACGAACC
GTTATTCAGTCGATAGTAACCCGCCCTTCGGGGATAGCAAGCATTTTTTGCAAAAAGGGGT

, 0 downstream)
TCTTTCAGAGTCAACCCTGAATTTCAGGATTTTTCTCTTCAACCGAACCGGCTGTTTGTGT
TGACTGACTGCTGCATTCCCCAGCAAAAGCCCGCTTTATACCTTTTTACGCACAGAGTTAT
GCGAGCGTTGCGCAAACGTTTTCGTTACAATGCGGGCGAAAAATAAGGATGCCCCGTTAGG

Genedata

Module Identification

Automatically identified eukaryotic promoter structures that are organized in a modular fashion

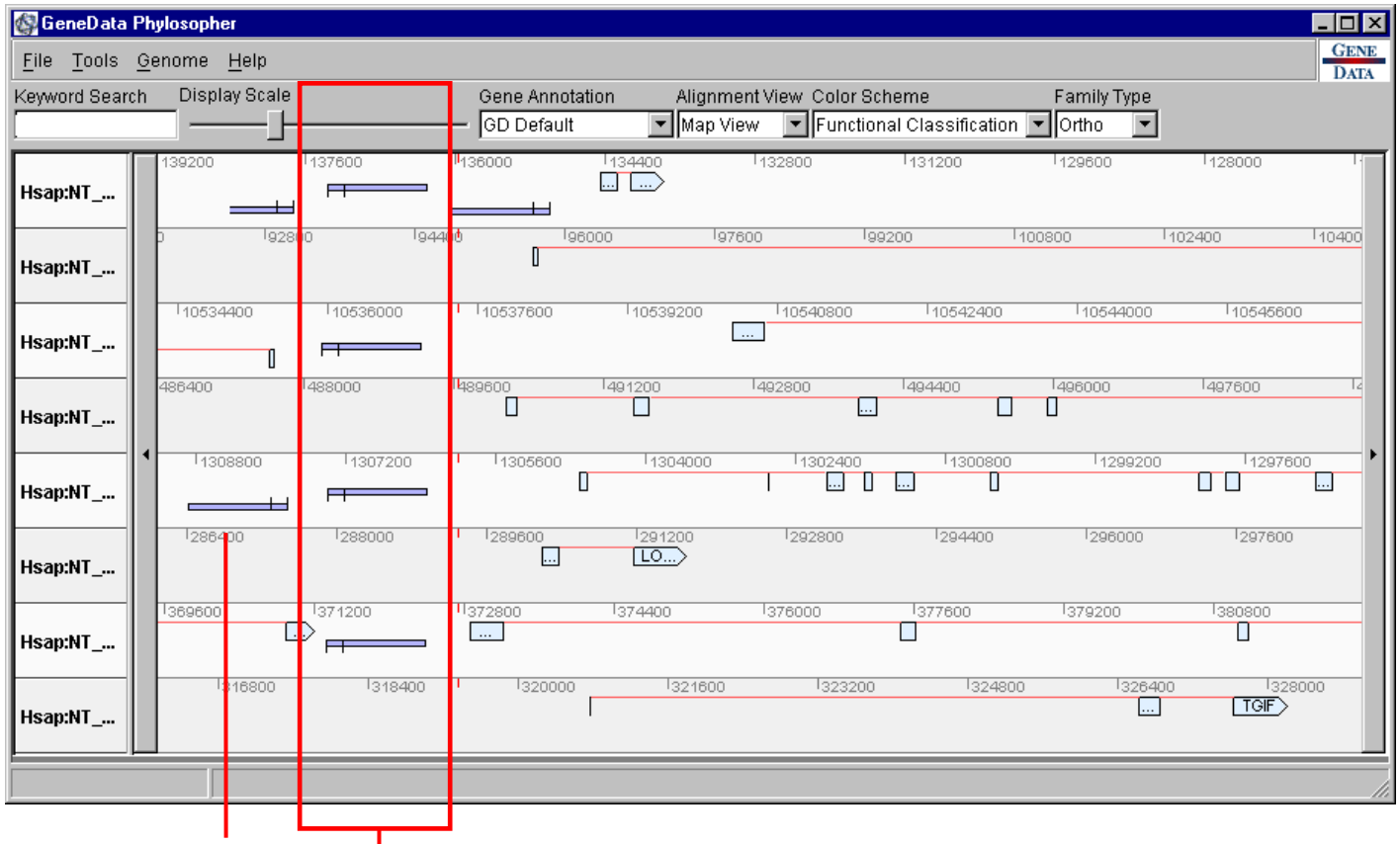


Two short transcription factor binding sites separated by a variable spacer → **GTTTAA–N(12–15)–GGGTA**

Characterization of Regulatory Networks

Searching whole genome for other genes with similar promoter structure

Global view on large-scale structure of promoters and enhancers



Reversed elements (opposite strand)

Alignment of promoter structures located upstream of a set of co-regulated genes

Genedata

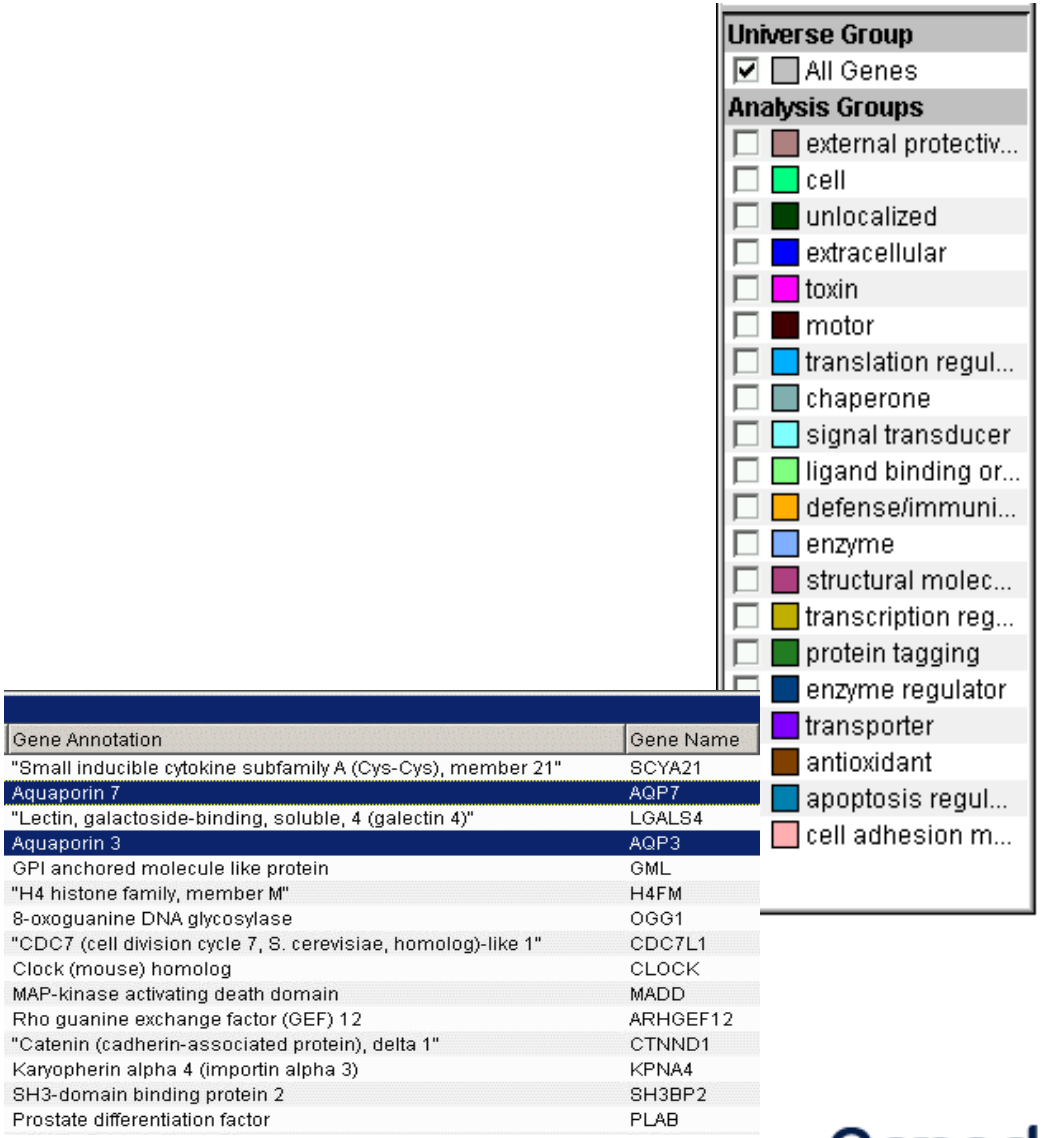
Mode of Action Analysis

Genedata

Annotation

Annotation columns:

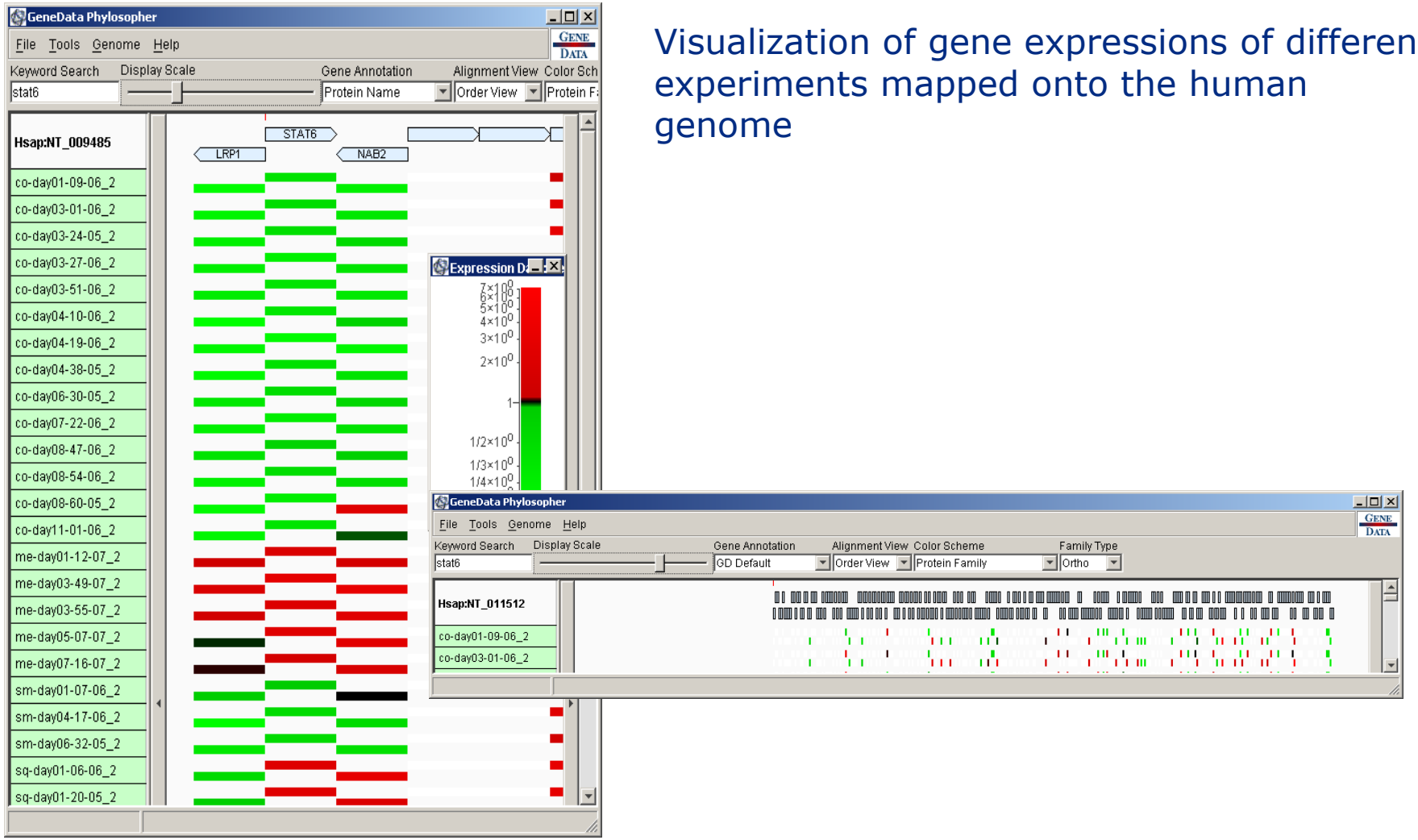
- Sequence description
- GenBank acc.
- UniGene ID
- UniGene Title
- Gene Symbol
- Sequence Type
- Map Location
- LocusLink
- GO_Biological_Process
- GO_Molecular_Function
- GO_Cellular_Component



Gene Annotation	Gene Name
"Small inducible cytokine subfamily A (Cys-Cys), member 21"	SCYA21
Aquaporin 7	AQP7
"Lectin, galactoside-binding, soluble, 4 (galectin 4)"	LGALS4
Aquaporin 3	AQP3
GPI anchored molecule like protein	GML
"H4 histone family, member M"	H4FM
8-oxoguanine DNA glycosylase	OGG1
"CDC7 (cell division cycle 7, S. cerevisiae, homolog)-like 1"	CDC7L1
Clock (mouse) homolog	CLOCK
MAP-kinase activating death domain	MADD
Rho guanine exchange factor (GEF) 12	ARHGEF12
"Catenin (cadherin-associated protein), delta 1"	CTNND1
Karyopherin alpha 4 (importin alpha 3)	KPNA4
SH3-domain binding protein 2	SH3BP2
Prostate differentiation factor	PLAB

Genedata

Mapping of Gene Expression Data Onto Genome

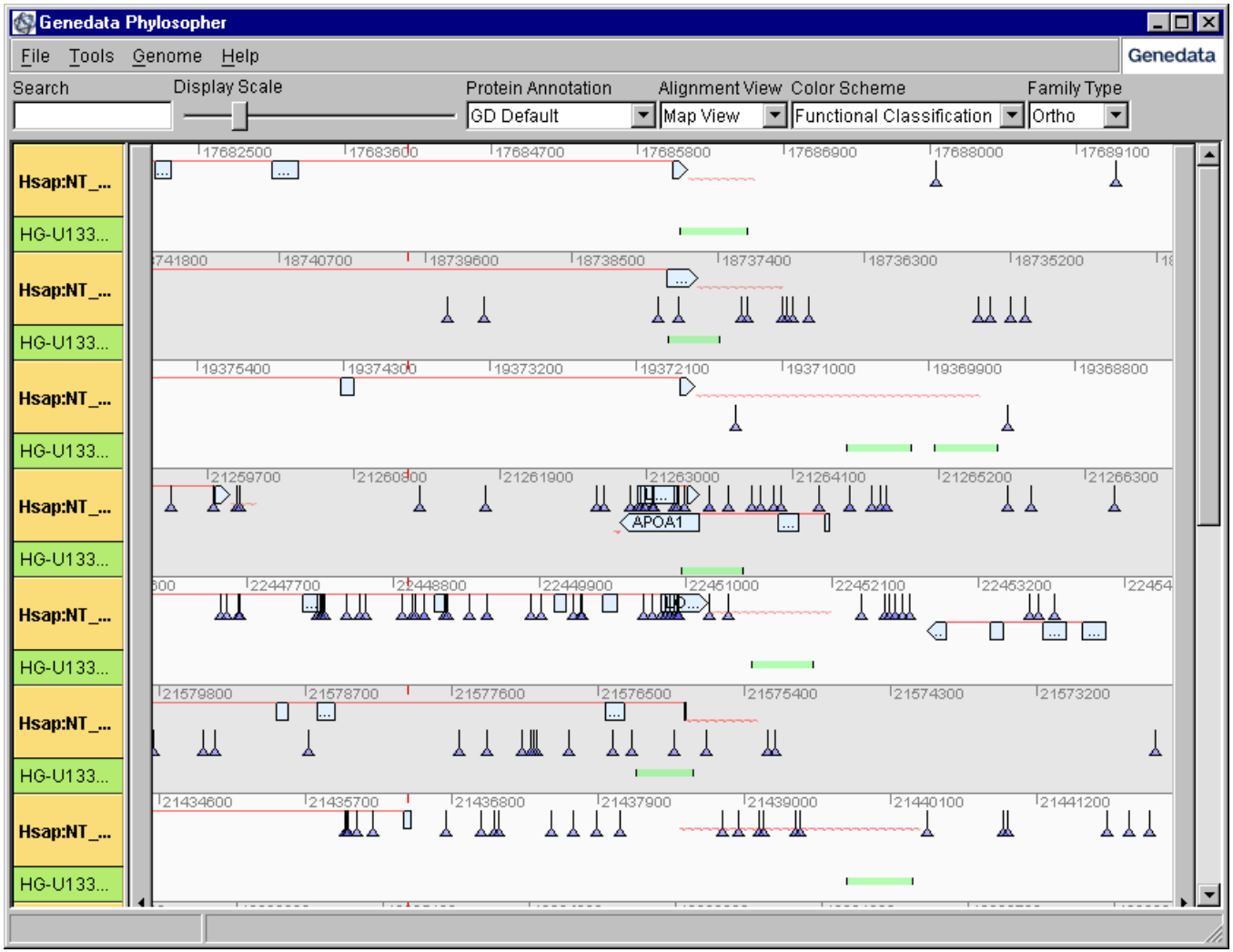


Visualization of gene expressions of different experiments mapped onto the human genome

Genedata

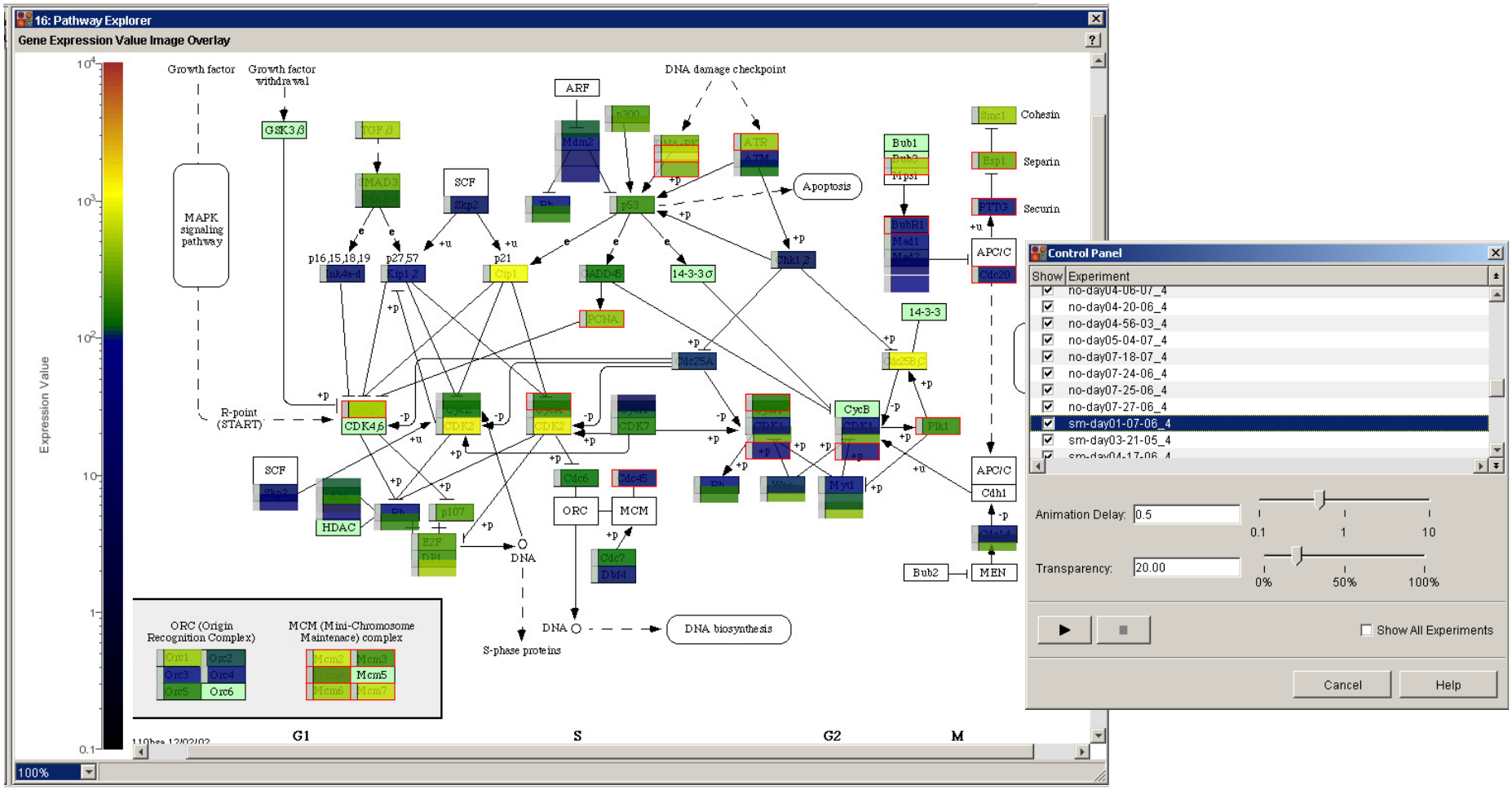
Binding Specificity Affected by SNPs

Investigation of SNPs that may affect the oligonucleotide binding specificity, resulting in reduced hybridization signals



Genedata

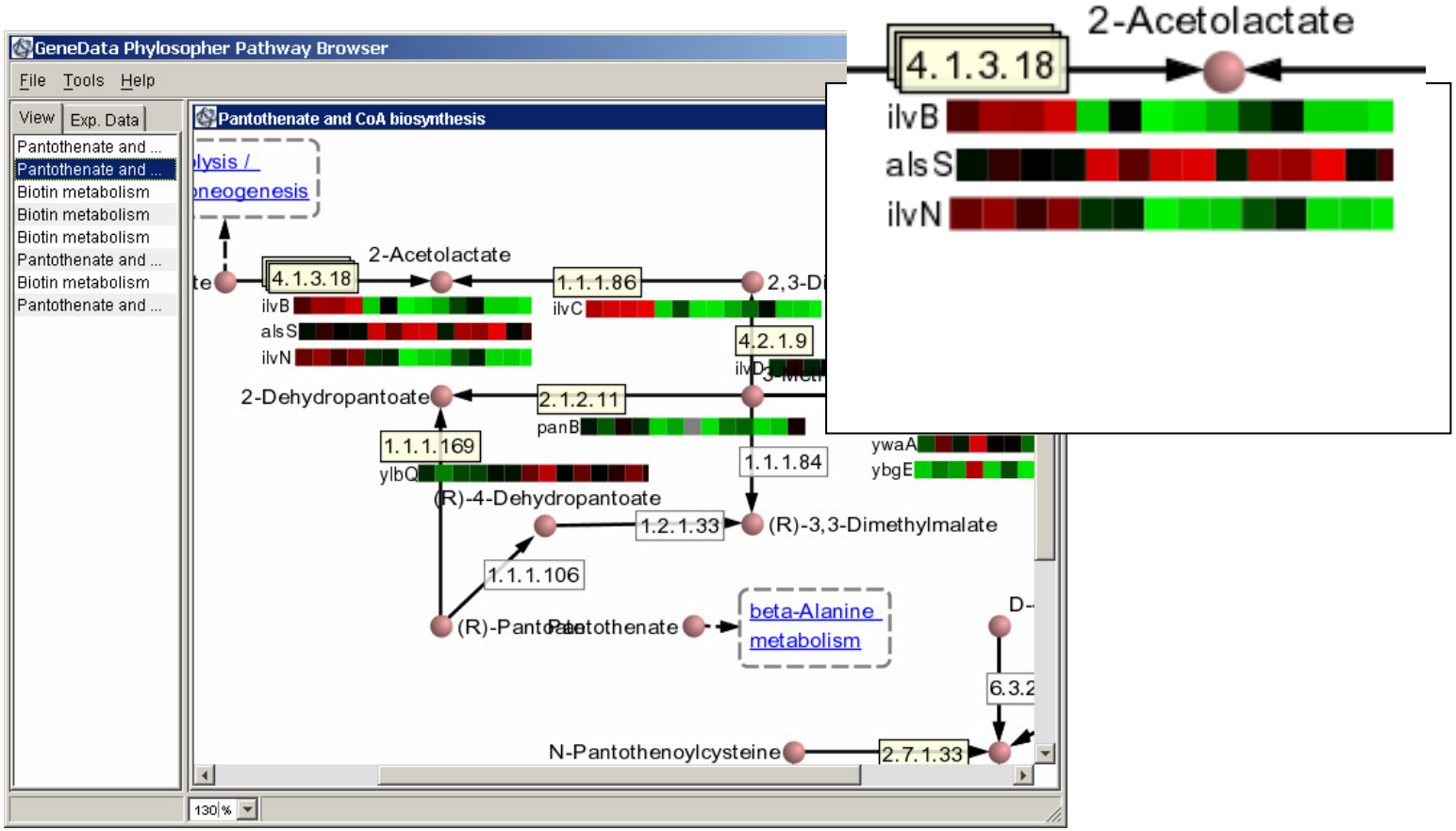
Pathway Exploration



Genedata

Time Series Expression Profiling

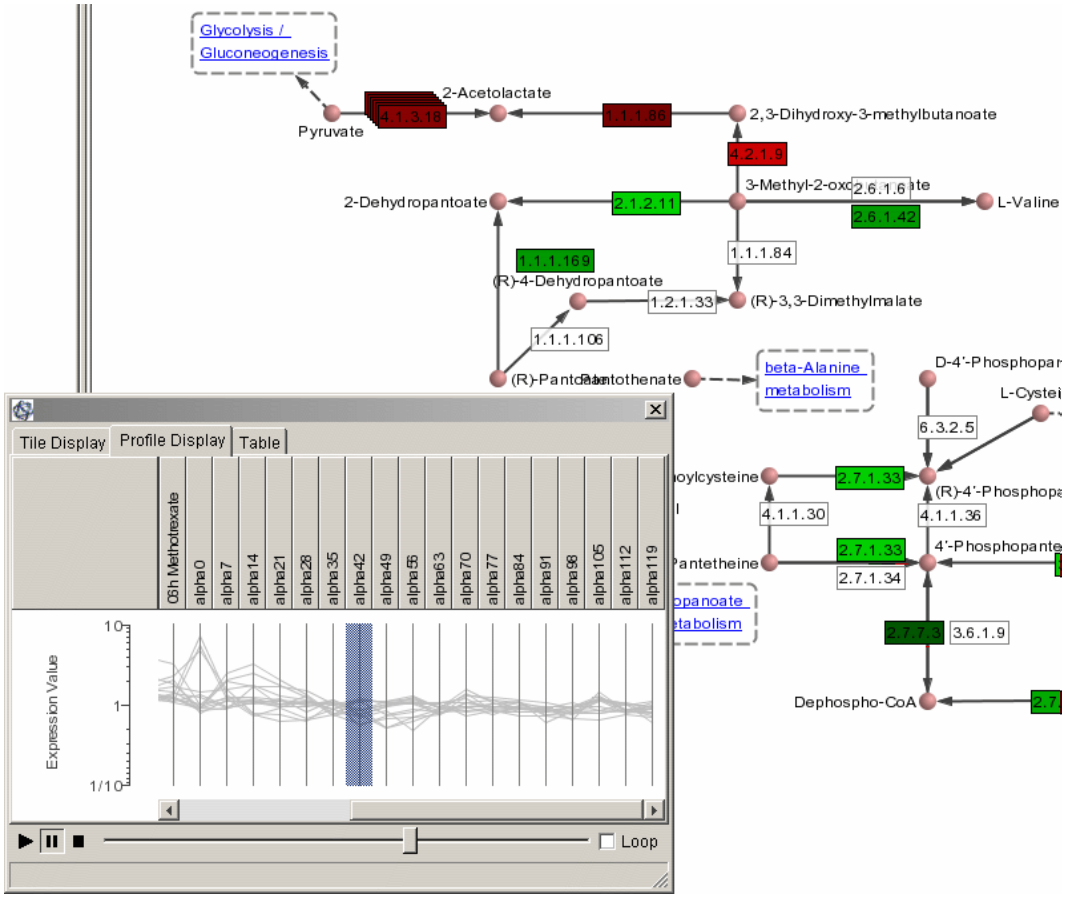
Mapping of gene expression profiling experiments onto known metabolic pathways (e.g. time series)



Genedata

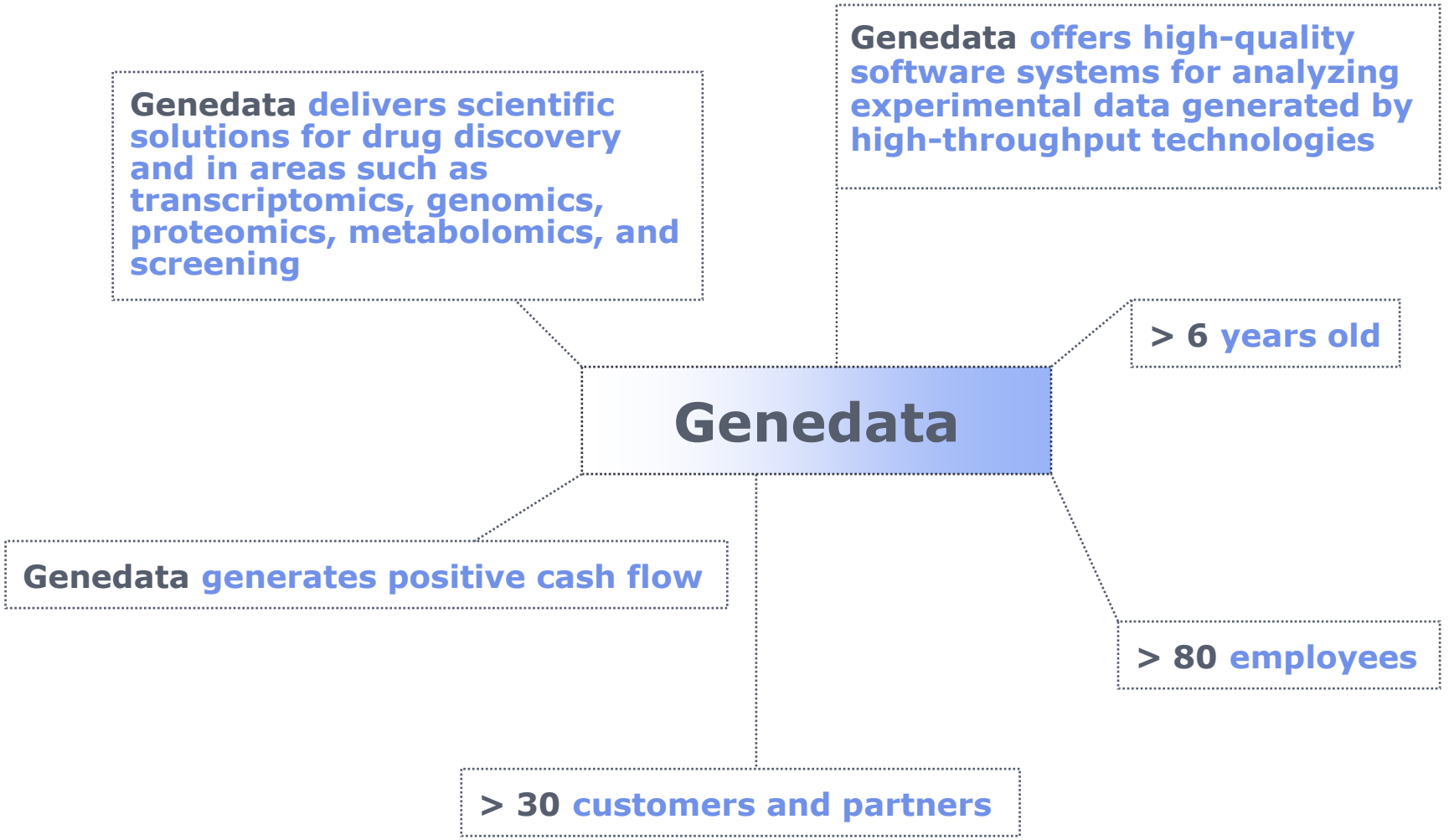
Analysis of Expression Experiment Series

Large sets of expression experiments can be analyzed by automatic cycling through the datasets



Genedata

Company Facts



Business Partners

Pharmaceutical Companies

e.g. Altana Pharma, AstraZeneca, Aventis Pharma,
Bayer Healthcare, Berlex, Novartis Pharma, Roche Pharma, Schering,
Wyeth Pharmaceuticals …

Agrochemical, Biotechnology, and other Life Science Related Companies

e.g. 454 Corporation, Arrow Therapeutics, Bayer CropScience, Bayer Diagnostics, Degussa, diaDexus, MWG-Biotech, Masterfoods, Roche Vitamins, Syngenta …

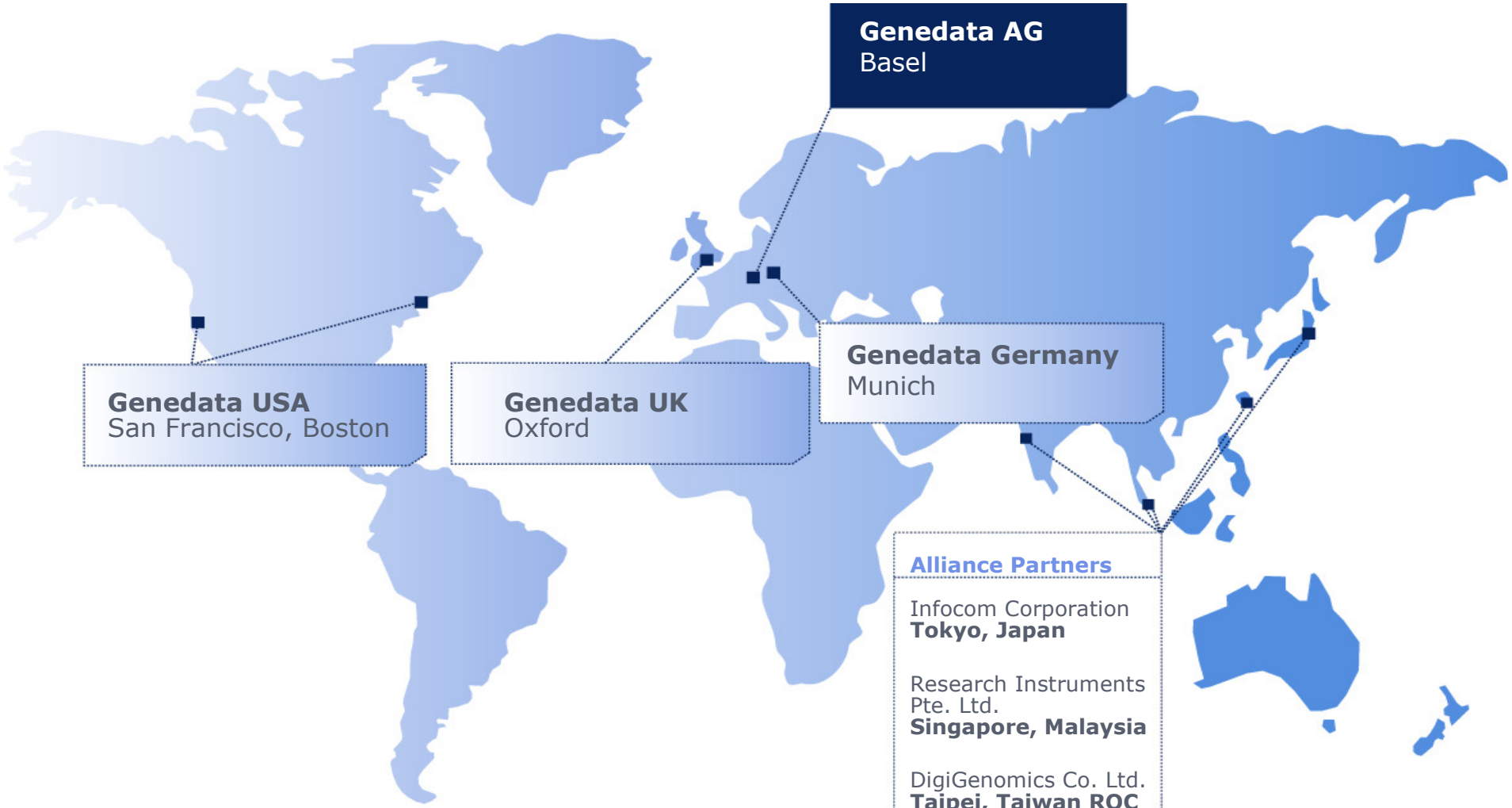
Academic Institutions

e.g. German Cancer Research Centre, Massachusetts Institute of Technology, National Institute of Technology & Evaluation Japan, National Cancer Center Singapore and Japan, PathoGenoMik Network Germany, Stanford University, University of Aarhus, University of California, Berkeley and Irvine, University of Goettingen, University of Minnesota, University of Muenster, University of Wuerzburg …

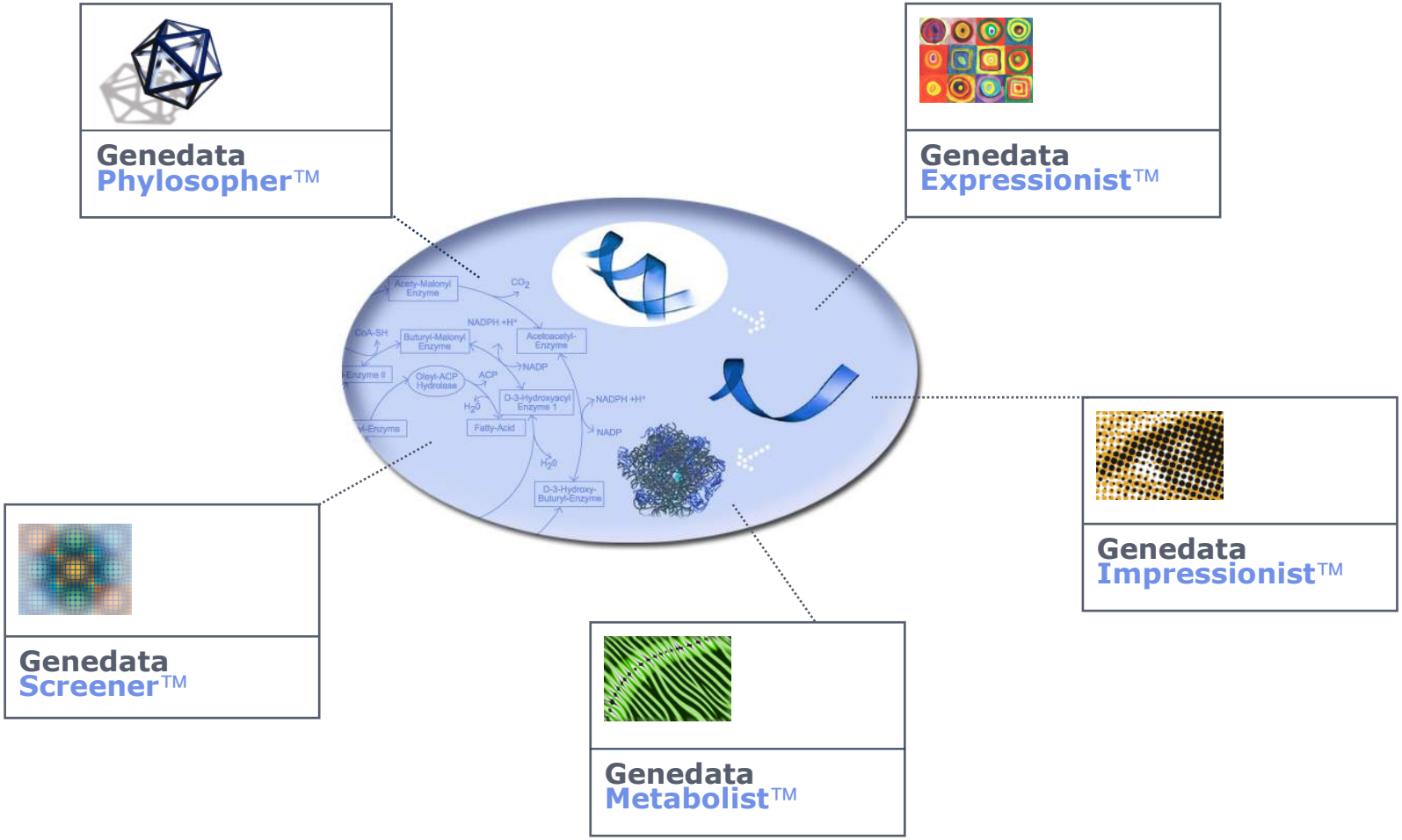
Technology Alliances

e.g. Affymetrix, Hewlett-Packard, IBM, Oracle, Silicon Graphics, Sun Microsystems …

Organization



Integrated Software Systems



www.genedata.com