In silico toxicity evaluation and mode of action prediction based on reference compendia

February 2006

Dr. Hans Gmuender
Scientific Consultant

Toxicogenomics

- Ideally, safety and efficacy of a new drug is determined simultaneously, enabling qualified decisions for the likelihood of success early in the discovery process

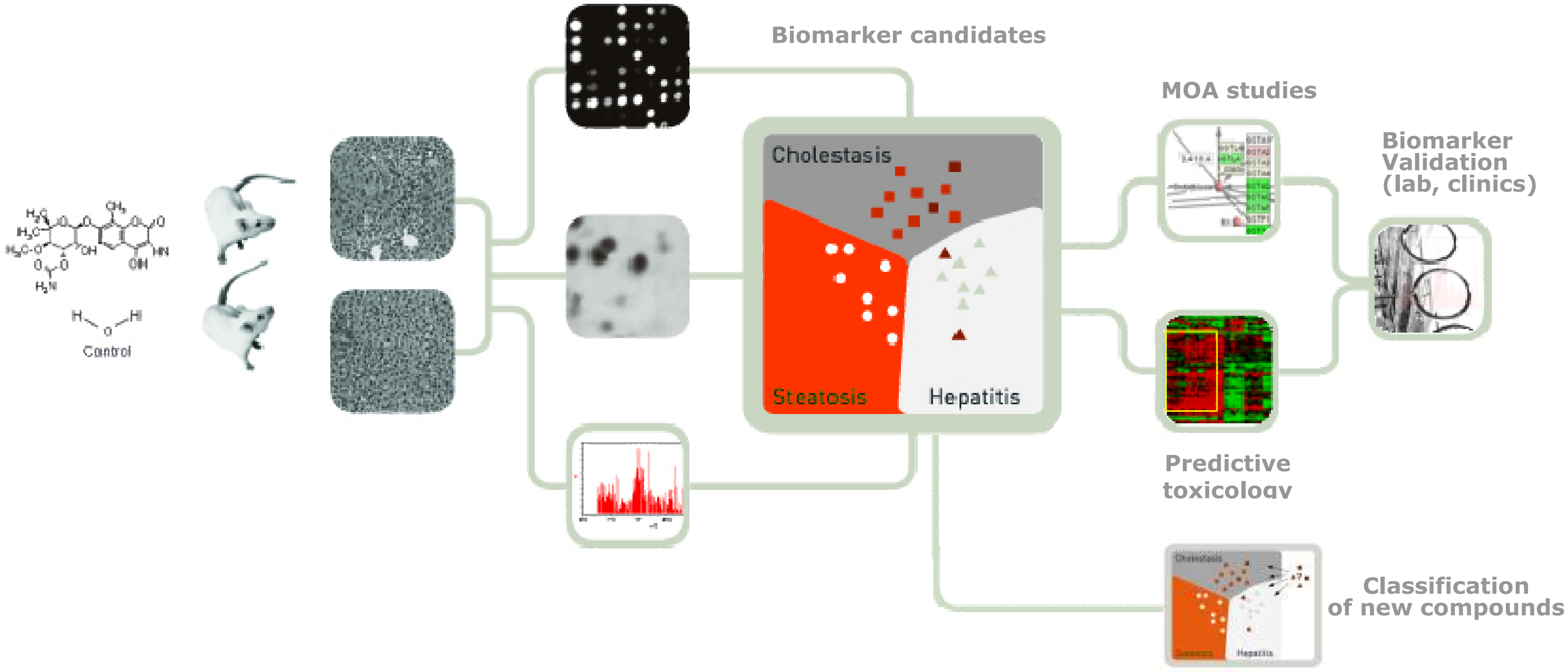
- Toxicogenomics is the study of how genomes respond to environmental stressors or toxicants

- Toxicogenomics combines classical toxicology and the technologies of -omics and bioinformatics to identify and characterize mechanisms of action of known and suspected toxicants

- Goals of toxicogenomics:
  - Prediction of long-term toxicity
  - Understanding of toxic effects (MOA or MOT)
  - Identification of cross-species biomarkers
  - Prediction of MOA of new compounds

Research process

Lab Work → T/P/M Omics → Computational Analysis → Lab Work



Biomarker candidates

MOA studies

Biomarker Validation (lab, clinics)

Predictive toxicology

Classification of new compounds

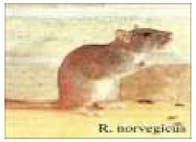
Predictive Tox Database

Compounds

- Compound
- Compound class
- Concentration
- Treatment time
- Dosing route
- Dosing frequency
- Vehicle
- Endpoints
- …

Animals

- Species
- Strain
- Sex
- Age
- Weight
- Observations
- …

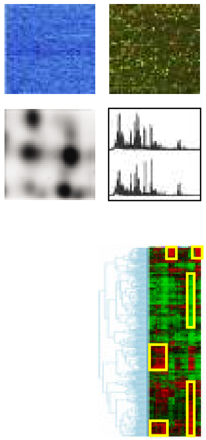


Treated animals

- Tissue
- Histopathology
- Clinical endpoints
- Serum chemistry
- Urine chemistry
- Hematology
- …

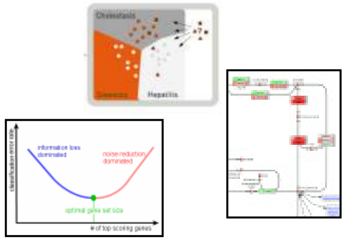
-omics data

- 1 channel data
- 2 channel data
- 2D gel
- LC/MS, GC/MS
- NMR
- Raw data
- Processed data
- Expression values
- Classification
- …



MOA, Tox Prediction

- MOA
- Biomarker candidates
- Tox Mechanism
- Tox Prediction
- …



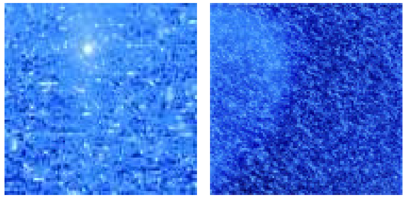
Data quality control and data normalization

- Toxicogenomics is crucially dependent on high quality expression data

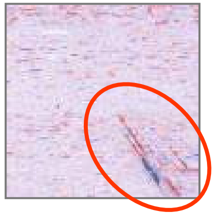
- Data quality control has to ensure:
  - Data quality assurance over large experimental series
  - High throughput analysis with standardized data processing
  - Diagnosis of technically conditioned effects
  - Enabling of consortial work and the submission of toxicogenomics data

Refiner Transcriptomics

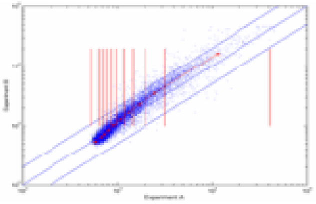
- Detection and correction of defects on microarrays



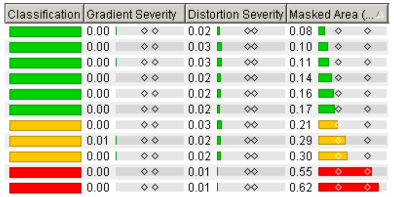
- Automated data quality control:
  - Loads uncondensed raw data
  - Detects and masks defective regions



  - Detects and corrects gradients and distortions



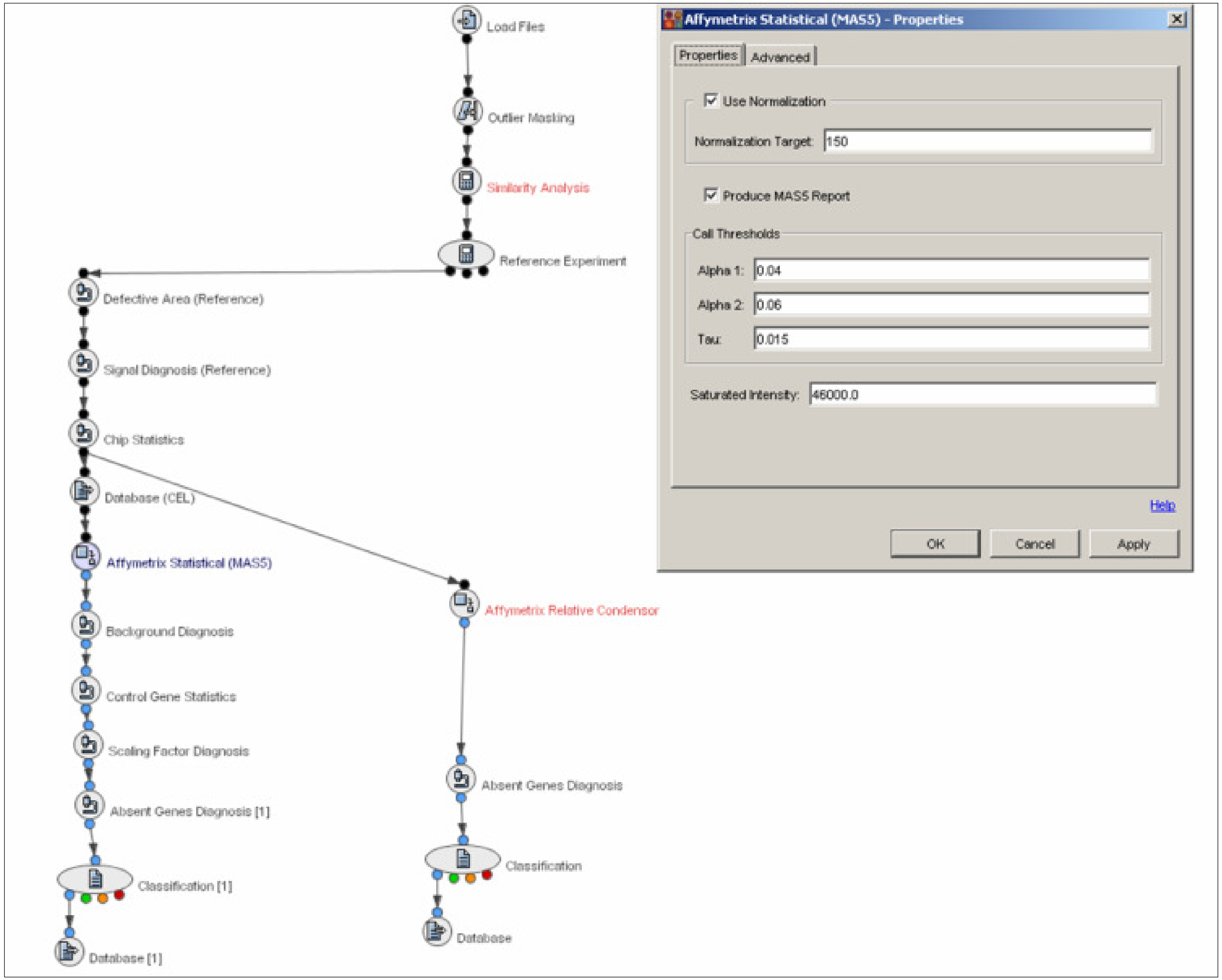
  - Condenses the data (MAS5, Li-Wong, RMA, GC-RMA)
  - Generates a quality classification for each chip



Classification	Gradient Severity	Distortion Severity	Masked Area (...
🟩	0.00	0.02	0.08
🟩	0.00	0.03	0.10
🟩	0.00	0.03	0.11
🟩	0.00	0.02	0.14
🟩	0.00	0.02	0.16
🟩	0.00	0.02	0.17
🟨	0.00	0.03	0.21
🟨	0.01	0.02	0.29
🟨	0.00	0.02	0.30
🟥	0.00	0.01	0.55
🟥	0.00	0.01	0.62

  - Saves condensed data into database

Workflow for data quality assessment for one-channel data



☑ Use Normalization

Normalization Target: 150

☑ Produce MAS5 Report

Call Thresholds

Alpha 1: 0.04

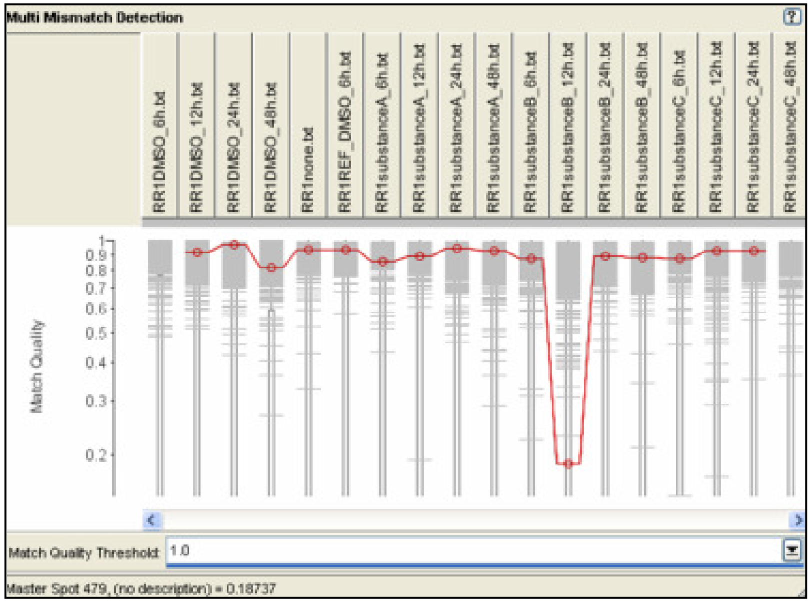
Alpha 2: 0.06

Tau: 0.015

Saturated Intensity: 46000.0

Refiner Proteomics

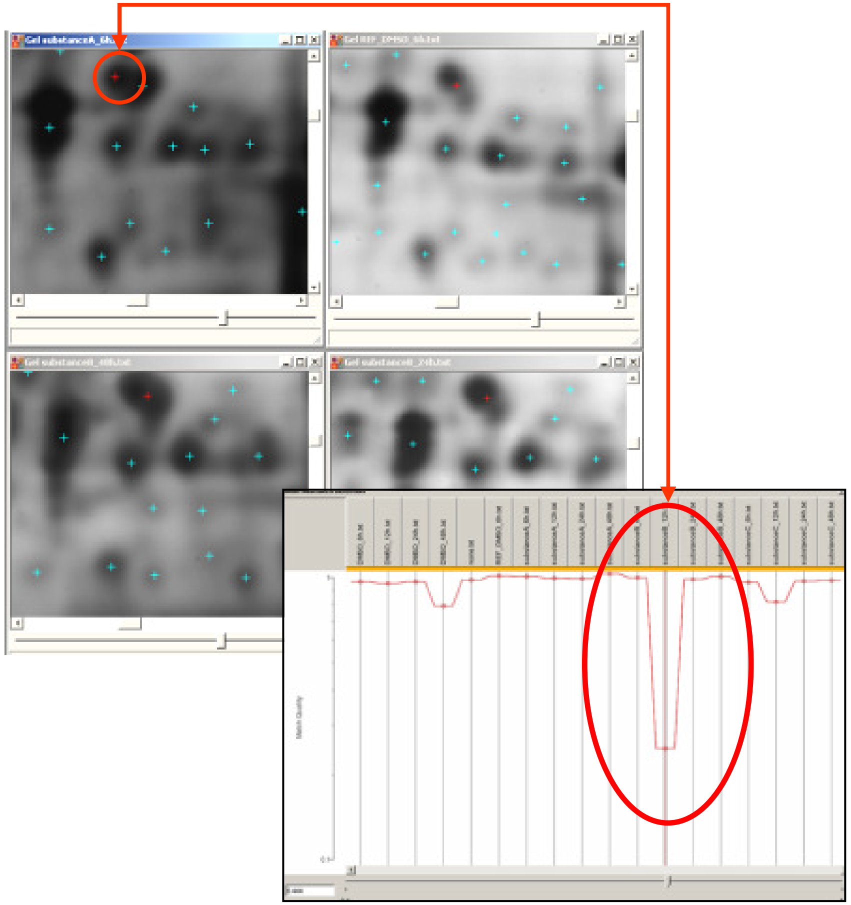
Compares location of spots over complete gel data set



Match Quality Threshold: 1.0

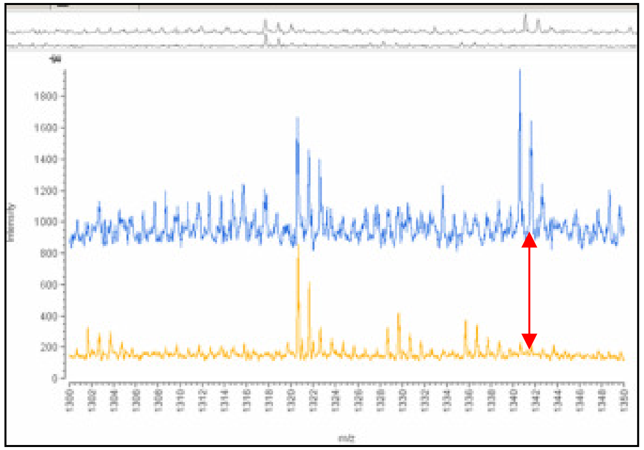
Master Spot 479, (no description) = 0.18737

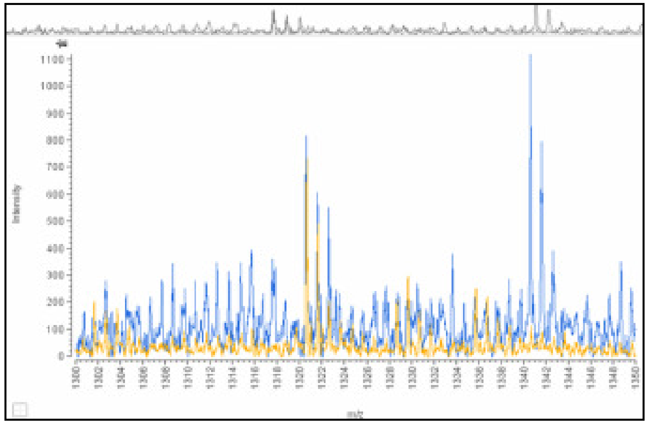
Automated mismatch detection based on calculation of standardized match scores



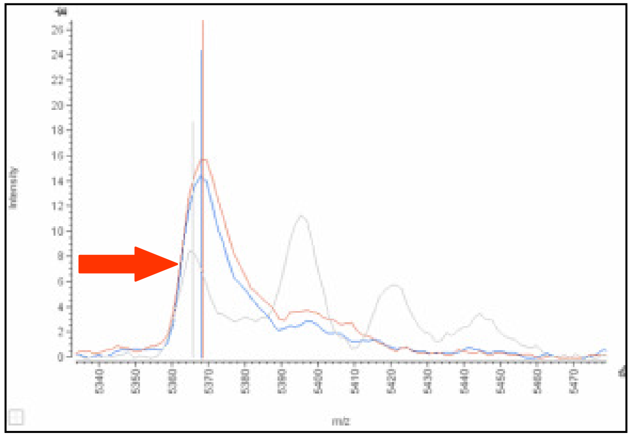
Refiner Metabolomics

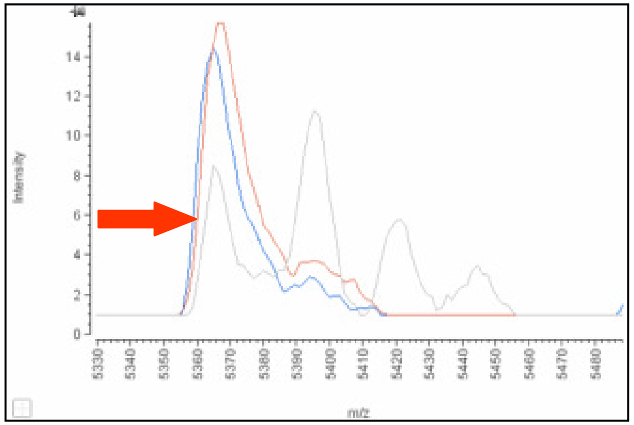
Baseline subtraction increases the comparability of spectra





m/z alignment prevents false positives in biomarker detection

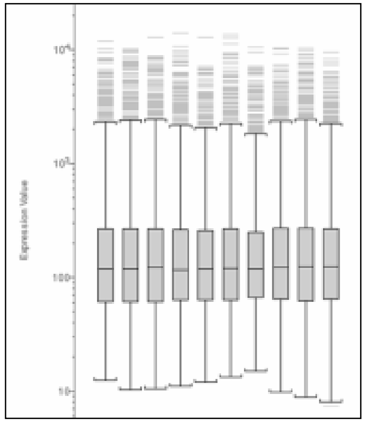


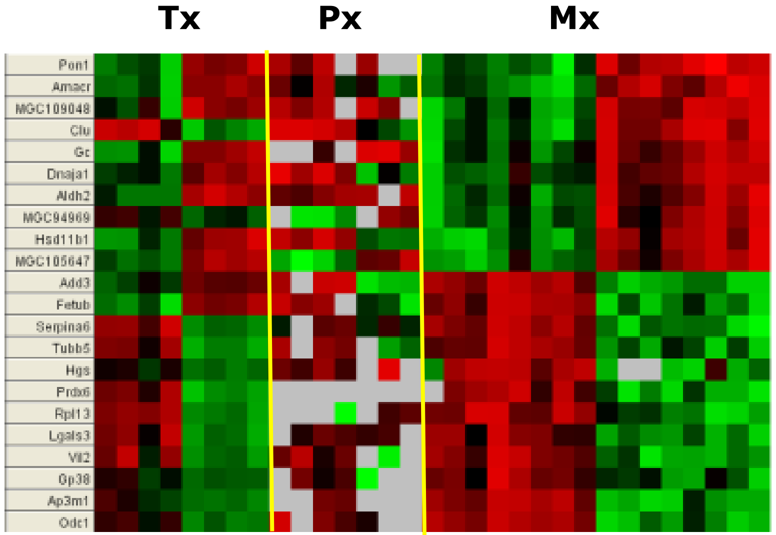


Mapping and normalization

- Integration and simultaneous analysis of:
  - Different Affymetrix chips (e.g. HG-U95 and HG-U133)
  - Chips from different providers (e.g. Affy and Agilent)
  - Chips covering different species (e.g. Mouse and human)
  - Different technologies (transcripts, metabolites, proteins)

- Normalization:
  - Arithmetic Mean
  - Logarithmic Mean
  - Median
  - Pointwise Division
  - LOWESS
  - Half Z-Norm.
  - Z-Norm.

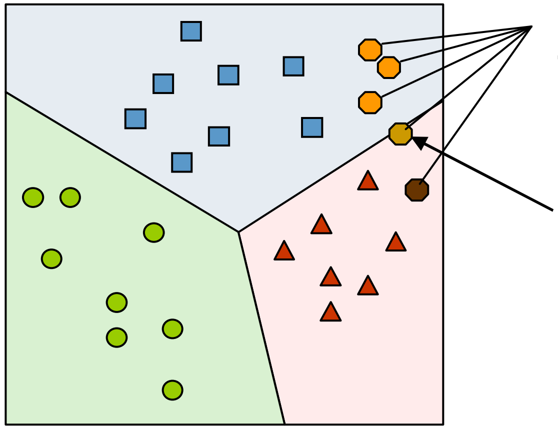




Mapping of data into a gene symbol space

Reference compendium for toxicity prediction

- Expression profiles of known, well-described compounds applied under diverse conditions frame a reference compendium
- The idea of a reference compendium is to predict the 'toxicity' of a new compound (with unknown toxicity) by assigning it to the Tox class of the compounds in the reference compendium with the 'closest' expression profile



Experiments performed with a new compound
3 classified into the 'blue' class
1 classified into the 'red' class

Classified with low affinity into the 'red' class

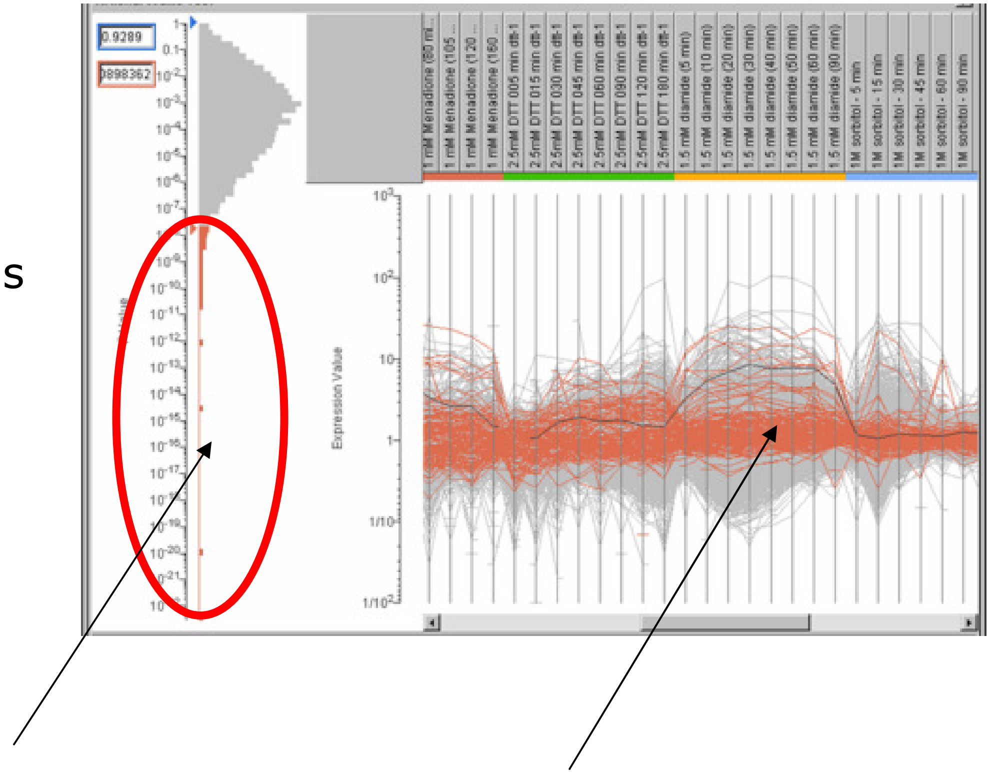
Reference compendium

Detection of significant genes by unsupervised learning algorithms

- Unsupervised computational methods can be used to arrange transcripts/proteins/metabolites in groups or clusters based solely on the similarity of their expression

- Unsupervised learning algorithms
  - ANOVA (analysis of variance)
  - N-way ANOVA
  - Kruskal Wallis ranking test
  - Principal Components Analysis
  - Hierarchical clustering
  - Self-organizing maps
  - K-means clustering

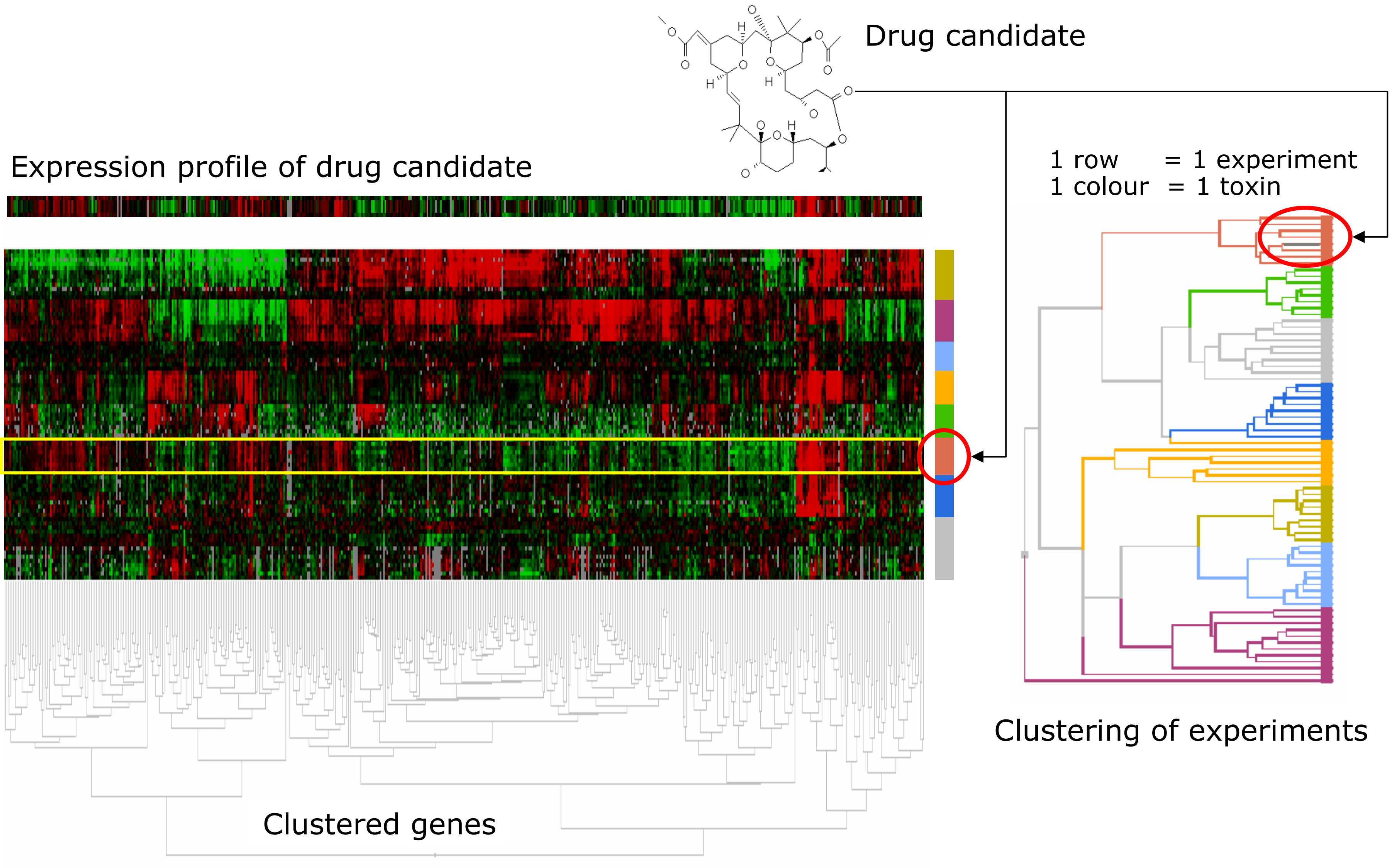
ANOVA



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

mRNA changes characteristic for a certain toxic mechanism

Prediction of toxicity of a new compound using unsupervised learning methods



Drug candidate

Expression profile of drug candidate

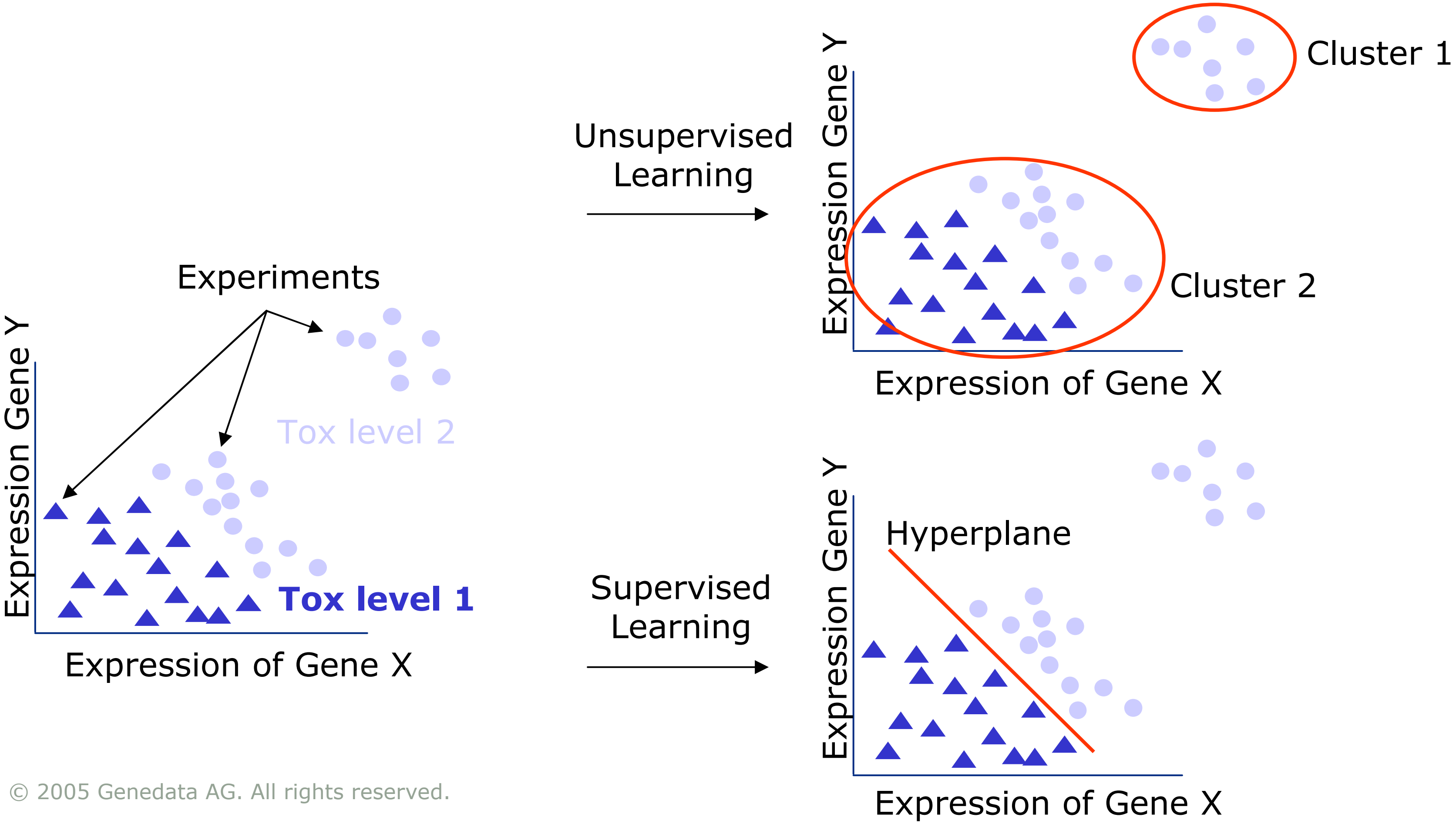
1 row = 1 experiment
1 colour = 1 toxin

Clustered genes

Clustering of experiments

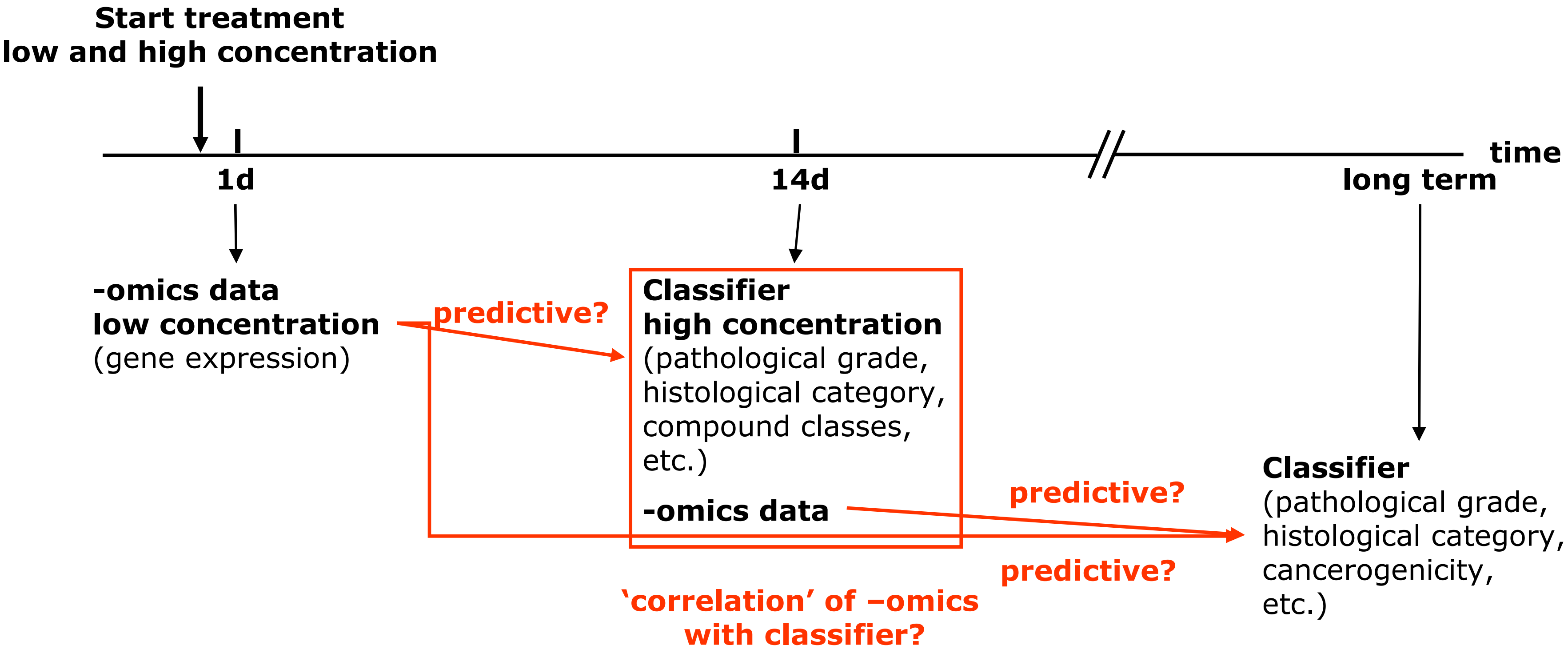
Detection of significant genes by supervised learning algorithms

- Supervised learning algorithms predict an output variable (e.g. a toxicity level) from input data (e.g. transcript, protein or metabolite expression)
- In contrast to unsupervised learning methods a priori knowledge on compounds’ 'toxicity' can be taken into account



Hypothesis-driven correlation

- Identification of relevant associations between sample phenotypes (a priori knowledge) and expression



Start treatment
low and high concentration

time
1d
14d
long term

-omics data
low concentration
(gene expression)

Classifier
high concentration
(pathological grade, histological category, compound classes, etc.)

-omics data

Classifier
(pathological grade, histological category, cancerogenicity, etc.)

predictive?

predictive?

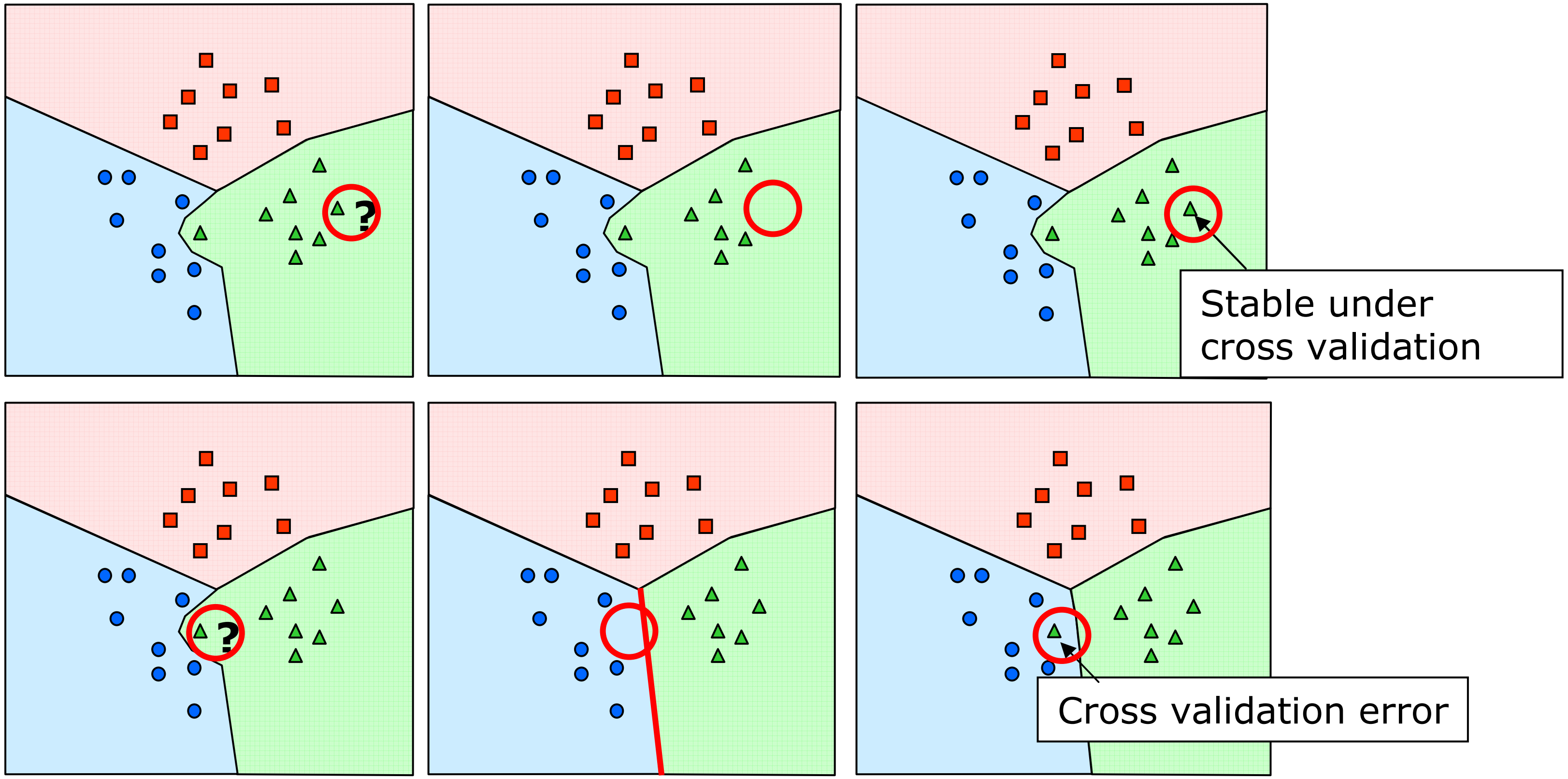
predictive?

'correlation' of –omics with classifier?

- What shall be predicted? Careful selection of hypothesis-driven classifier!

Cross validation of reference compendium

- Cross validation is a widely used method for estimating the prediction error of a reference compendium
- The goal of this intrinsic validation is to evaluate whether the reference compendium can be used for predicting the output variable of a compound based on the expression profile



Stable under cross validation

Cross validation error

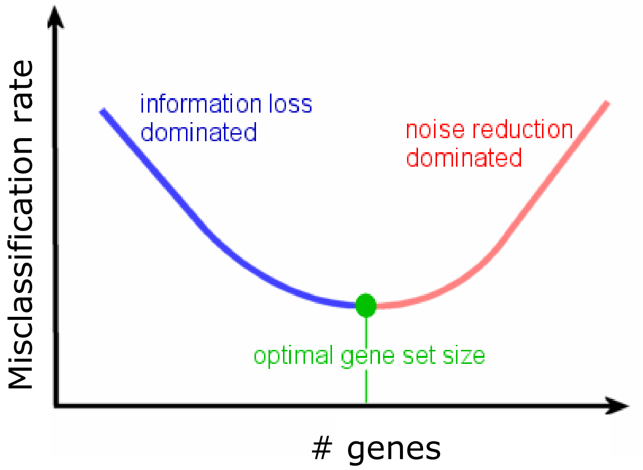
Determination of the optimal set of genes

- Besides the problem of estimating the prediction error, there also exists the issue to identify the set of genes that minimizes the prediction error and are therefore the best 'toxicity' predictors

- Genes from optimal set of genes are potential biomarkers

- Supervised learning
  - Support Vector Machine
  - Sparse Linear Discriminant Analysis
  - Fisher Linear Discriminant Analysis
  - K-Nearest Neighbours

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



Case study

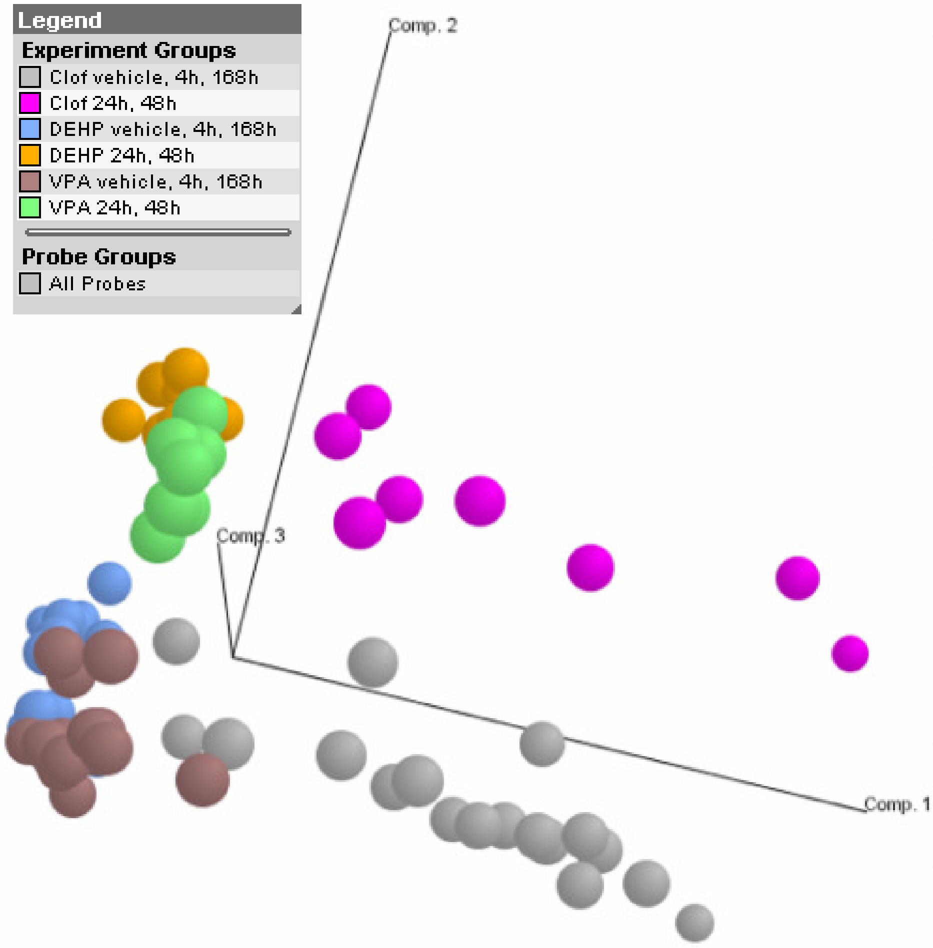
- Compounds

Clofibrate in 0.9% saline	1	g/kg
DEHP Diethylhexyl phthalate) in distilled water	20	g/kg
VPA (Valproic acid) in distilled water	2	g/kg

- Single dose (oral administration)
- Doses selected to obtain acute hepatotoxicity
- Time points: 4h, 24h, 48h, 168h (vehicles: 48h, 168h)
- 3 - 5 animals / compound and time point
- Isolation of total RNA from liver
- Hybridization to Affymetrix RG_U34A arrays
    - Each sample hybridized to a microarray
    - Samples from each compound and time point pooled and pool hybridized to microarrays

- R. A. Jolly, et al., (2005)
  Pooling samples within microarray studies: a comparative analysis of rat liver transcription response to prototypical toxicants
  Physiol Genomics, 22, 3, 346-55
  GEO Series GSE2303

Principal Components Analysis of all experiments

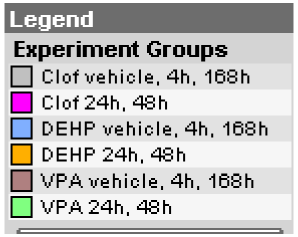


Vehicle, 4h, 168h experiments
= same colour for each compound

24h and 48h experiments
= same colour for each compound

- PCA with 3241 transcripts
  (detection p-value < 0.04, 50% valid values per group)
- Clofibrate experiments separated from DEHP and VPA experiments along axis of component 1
- 24h and 48h experiments separated from vehicle, 4h and 168h experiments along axis of component 2
- Weak response after 4h
- Recovery after 168h
- 4h and 168h experiments similar to 48h and 168h vehicle experiments
- 24h and 48h experiments not clearly separated

Clustering of all experiments

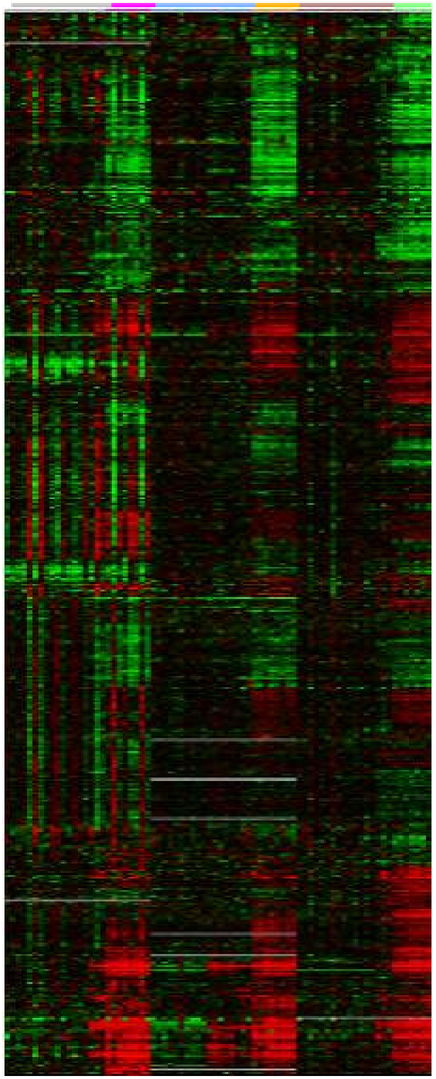


Legend

Experiment Groups

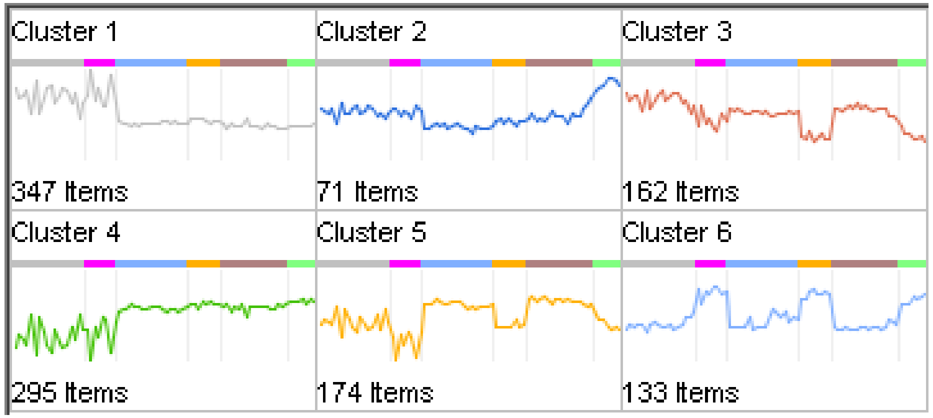
- ⬜ Clof vehicle, 4h, 168h
- 🟪 Clof 24h, 48h
- 🟦 DEHP vehicle, 4h, 168h
- 🟧 DEHP 24h, 48h
- 🟫 VPA vehicle, 4h, 168h
- 🟩 VPA 24h, 48h

Hierarchical clustering



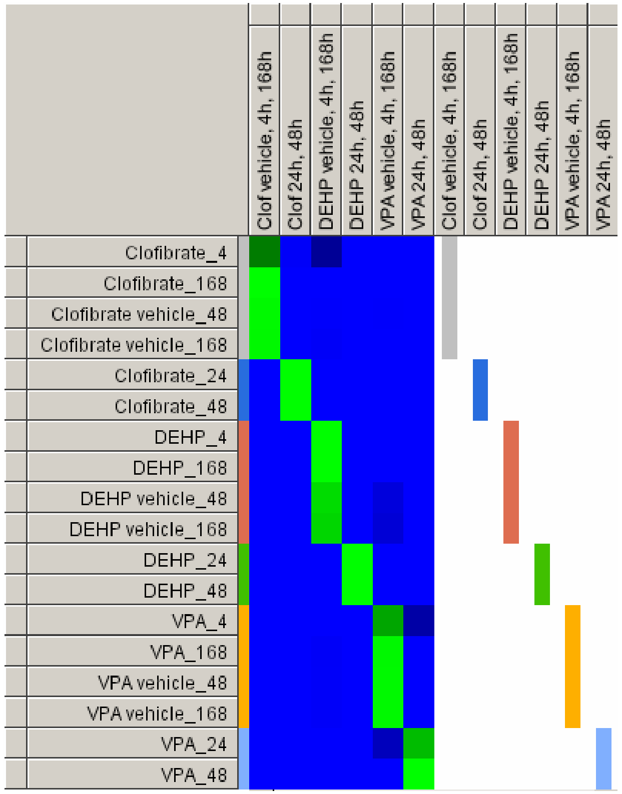
- Hierarchical clustering and K-means with 1205 transcripts after ANOVA (p-values < 0.001, fold change ≥ 2, detection p-value < 0.04, 50% valid values per group)
- Most up- or down-regulated transcripts after 24h and 48h
- Clofibrate experiments separated from DEHP and VPA experiments

K-means clustering



Which data should be selected to build a reference compendium?

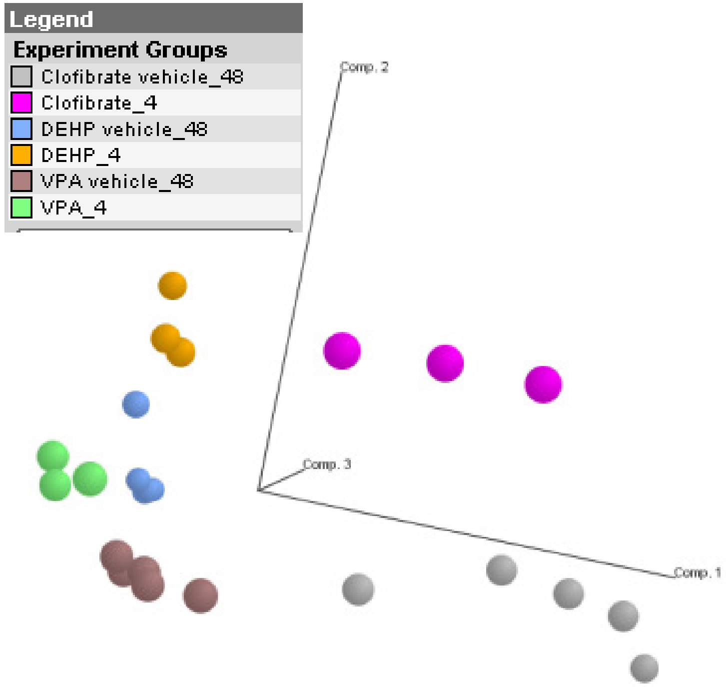
- Highly up- or down-regulated transcripts after 24h and 48h
- Experiment groups ‘no or low expression’ (4h, 168h, 48h vehicle, 168h vehicle) and ‘high expression’ (24h, 48h) could be used to build a reference compendium
- Misclassification rate ~ 0%

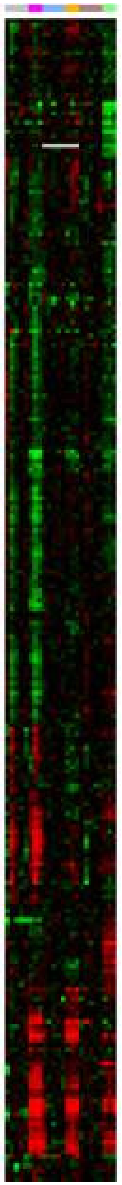


Questions:

- Is there enough information in the expression profiles of the 4h and 48h vehicle experiments to use these experiments for a reference compendium?
- Can these early time points be used to predict outcomes at later time points?

Analyses of vehicle and 4h experiments with transcripts selected from ANOVA

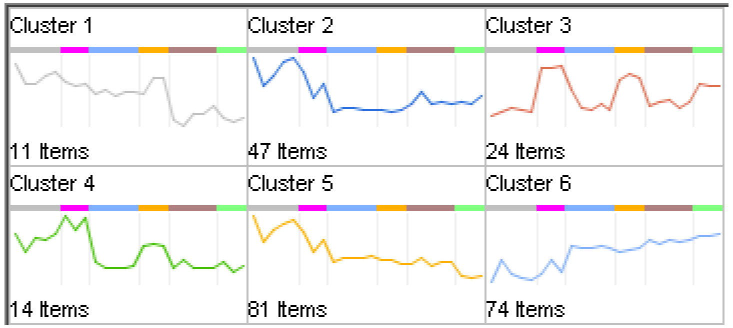




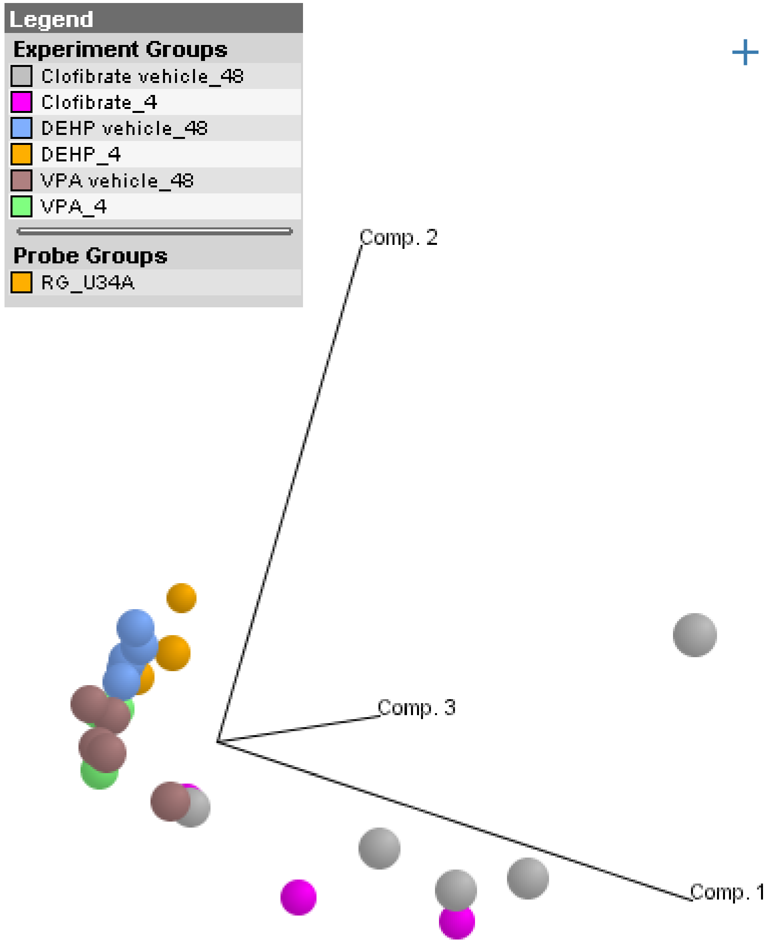
Hierarchical clustering

- PCA, K-means and hierarchical clustering with 251 transcripts after ANOVA
  (p-values < 0.001, fold change ≥ 2, detection p-value < 0.04, 50% valid values per group)
- 4h experiments separated from vehicle experiments
- Clofibrate experiments separated from DEHP and VPA experiments

K-means clustering



Principal Components Analysis of vehicle and 4h experiments

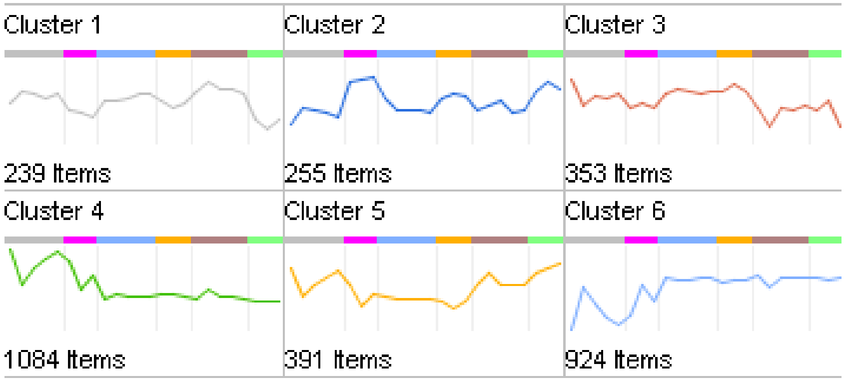


- PCA and K-means with 3241 transcripts (detection p-value < 0.04, 50% valid values per group)

  Clofibrate experiments separated from DEHP and VPA along axis of component 1

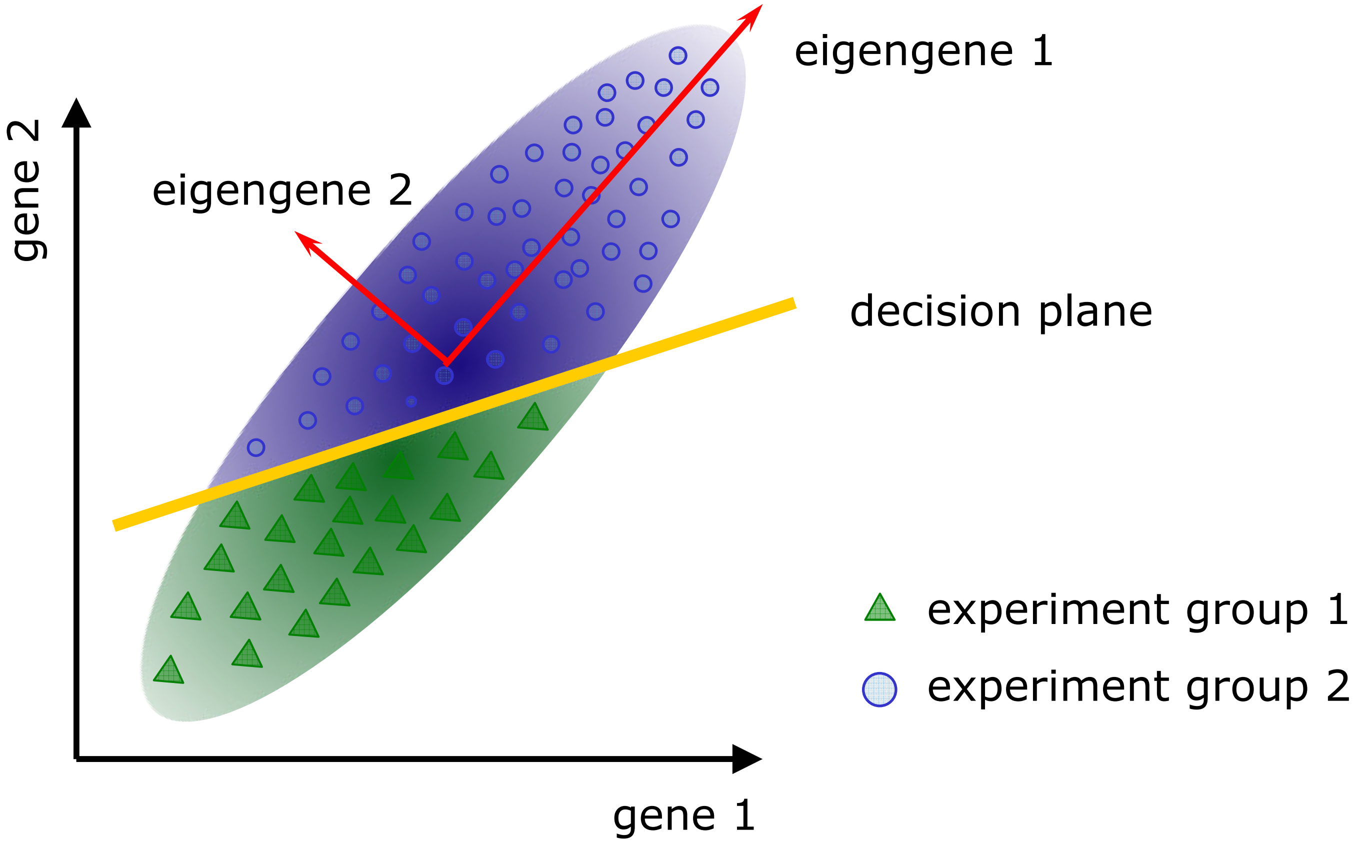
  No clear separation between vehicle and 4h experiments

K-means clustering



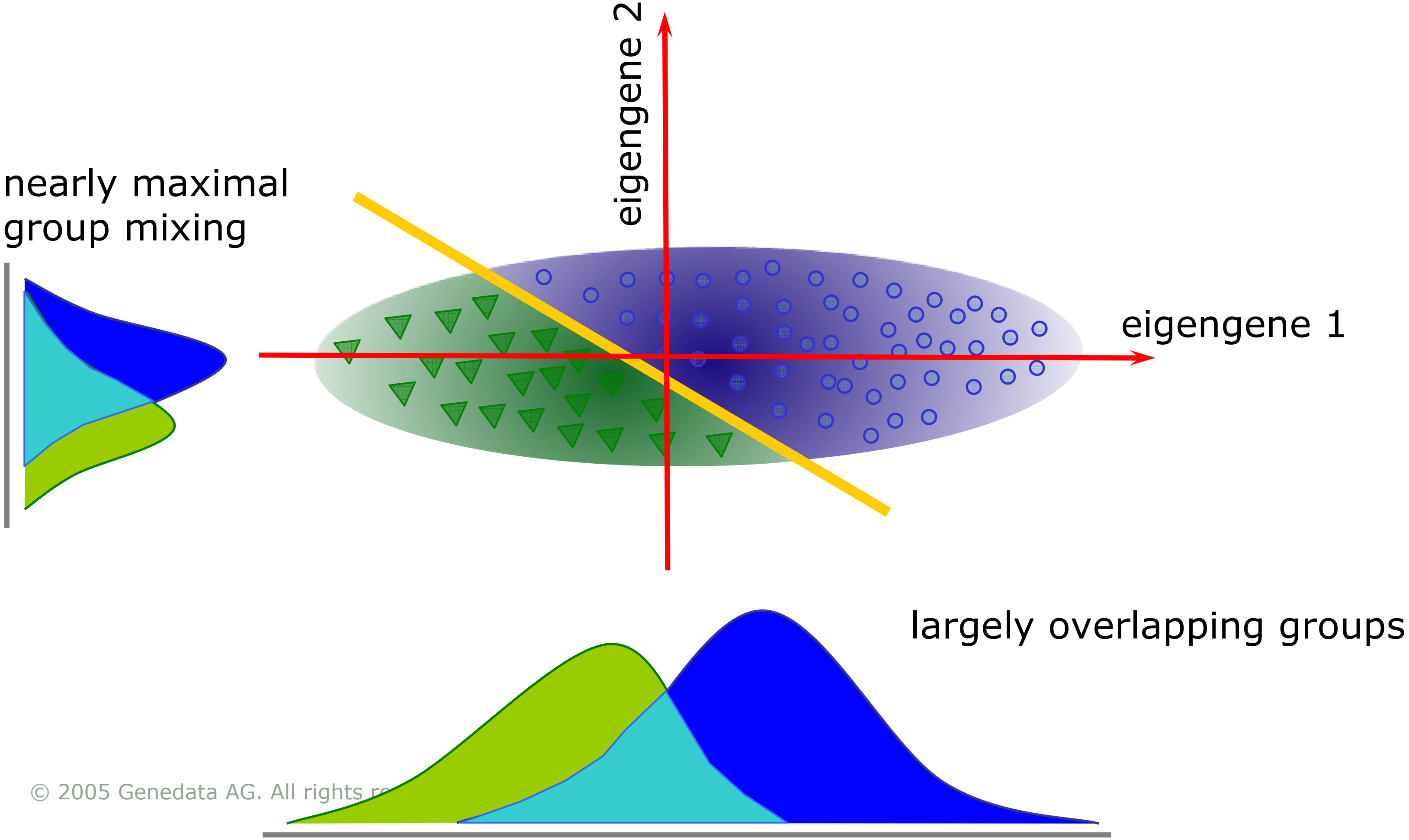
PCA vs. supervised learning

- PCA is in general not the best method to classify experiment groups as can be seen from the following example
- The group separation line (decision plane) is in general not parallel to any of the genes or eigen-directions



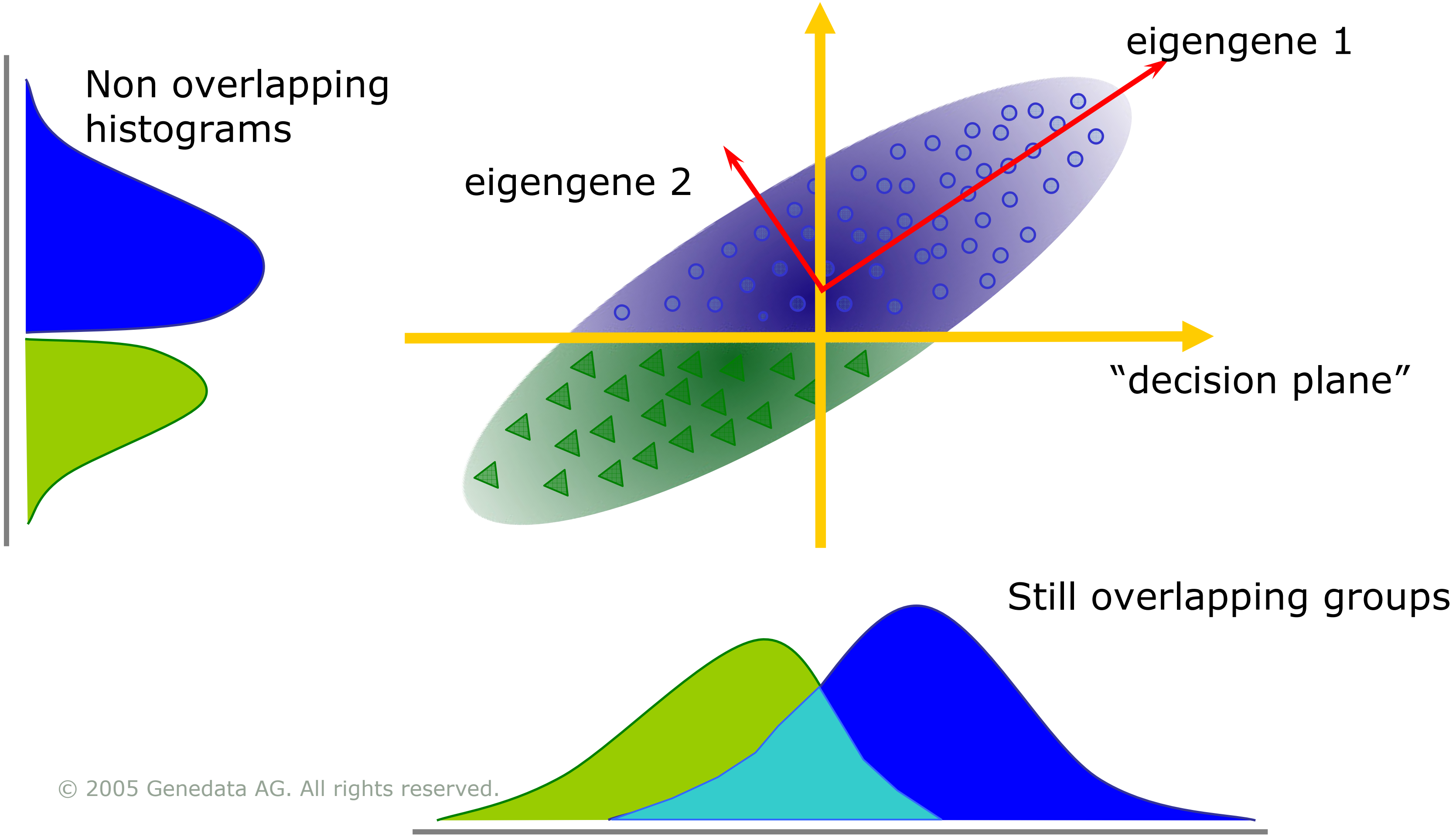
PCA: Classification using PCA

- A rotation to the eigenspace therefore does not solve the classification problem

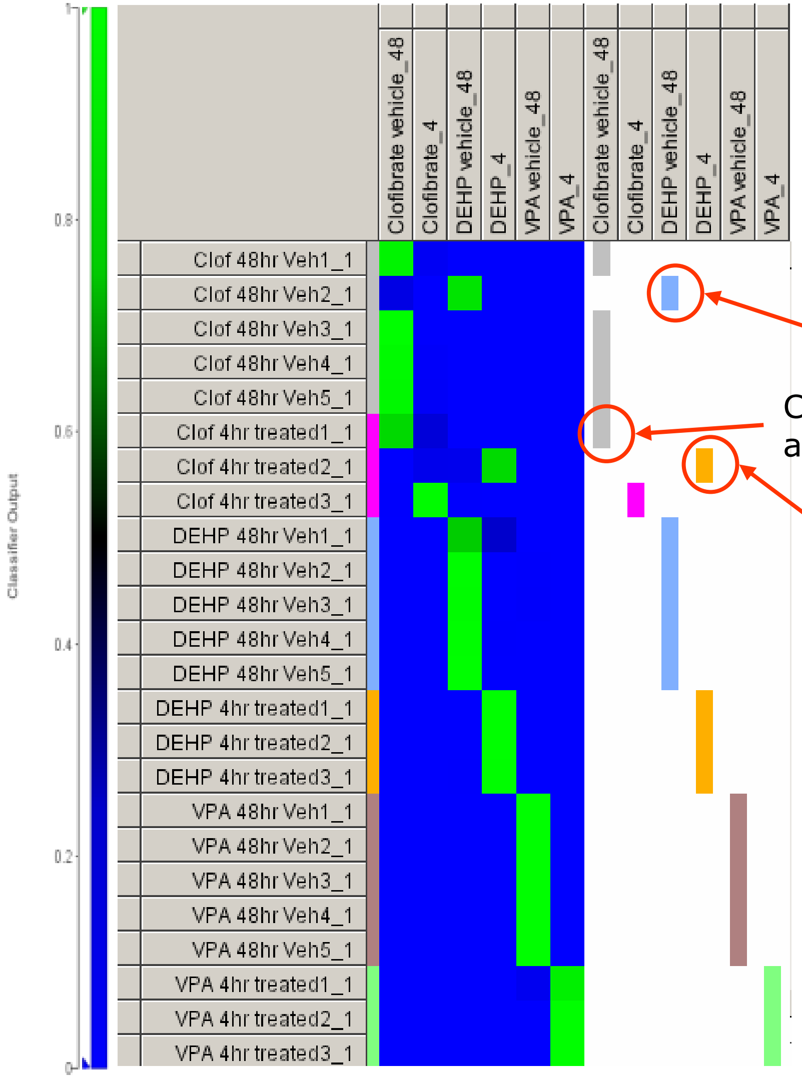


PCA vs. supervised learning

- A complete separation of groups is possible by using the coordinate system obtained from the classifier



Cross-validation of vehicle and 4h experiments



Clof 48h vehicle classified as DEHP vehicle 48h

Clof 4h classified as Clof vehicle 48h

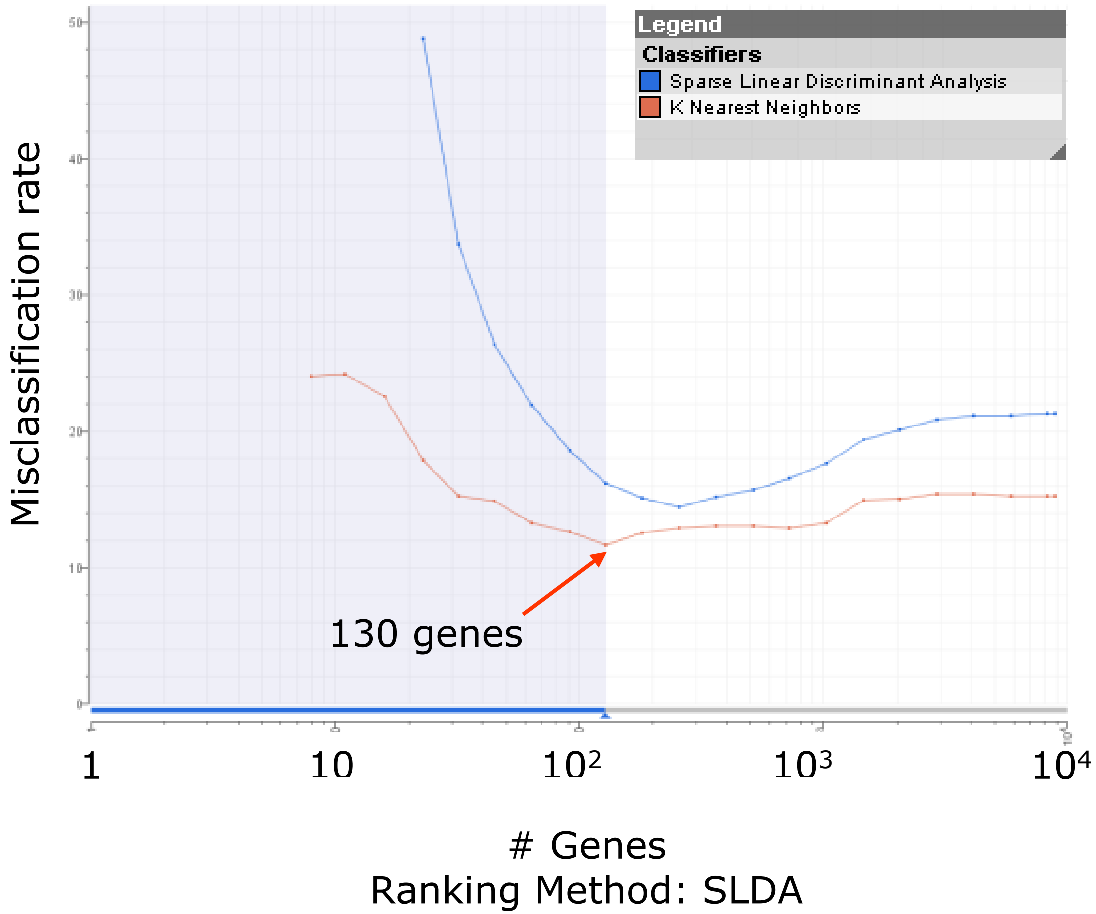
Clof 4h classified as DHEP 4h

- Classifier: KNN
  all genes (~ 8800)
  Distance: Positive Correlation
  K: 1
  Test Set Fraction: 25
  Number of Repeats: 500
- Misclassification rate ~ 13%

- Acceptable misclassification error in reference compendium consisting of vehicle and 4h experiments
- Reference compendium can be used for predicting the output variable of a compound based on the expression profile

Gene ranking to define optimal marker gene set

Gene ranking and cross validation

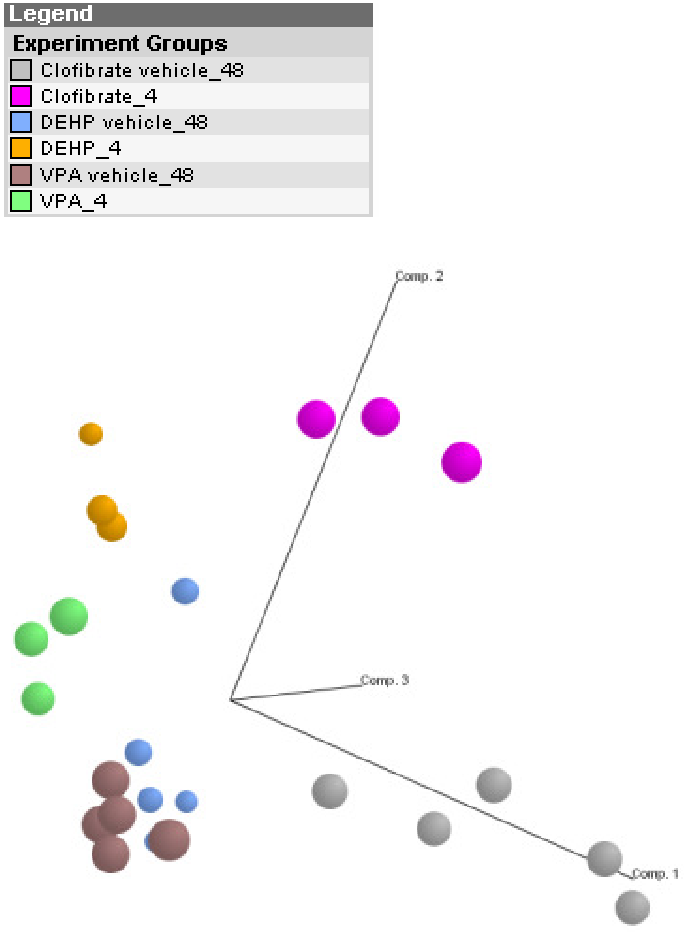


- Classifiers: KNN and SLDA
  all genes (~ 8800)
  Test Set Fraction: 25
  Number of Repeats: 500
  Ranking: Sparse Linear Ranking
  Kernel for SLDA: Linear
  Distance for KNN: Positive Correlation
  K: 1
- Prediction error ~ 15 %

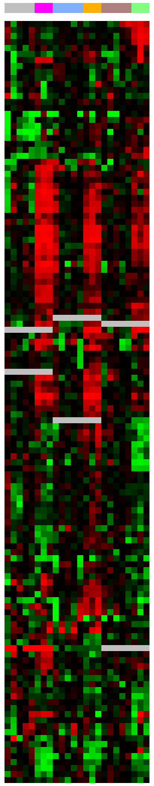
- Optimal marker gene set consists of ~ 130 genes
- Prediction error ~ 12 %

- No significant reduction of prediction error but reduction in number of significant transcripts

Analyses of vehicle and 4h experiments with optimal marker gene set

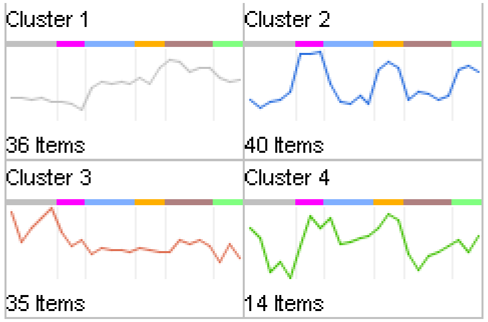


Hierarchical clustering



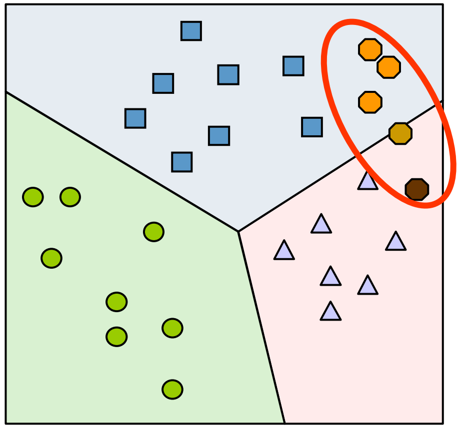
- PCA, K-means and hierarchical clustering with optimal marker gene set consisting of 130 transcripts (detection p-value < 0.04, 50% valid values per group)

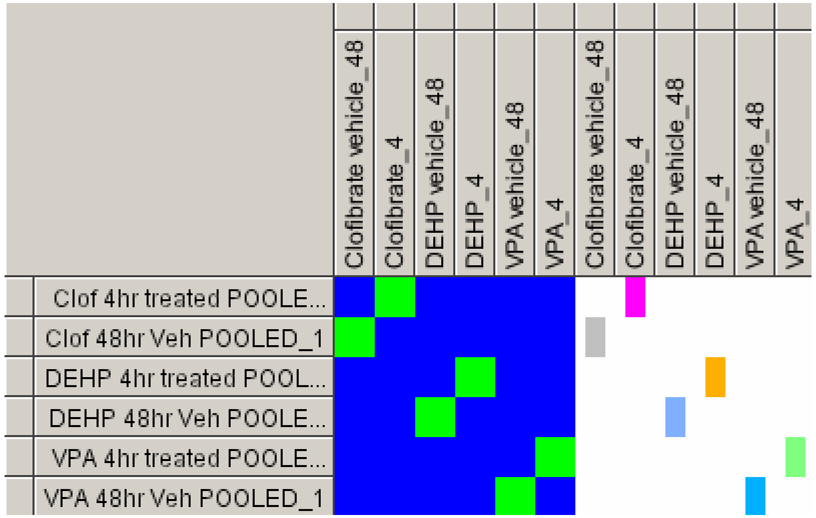
K-means clustering



Classification of 'Unknown' compounds

- Pooled vehicle and 4h experiments classified into compendium consisting of individual vehicle and 4h experiments
- Experiment is assigned to the group with highest affinity



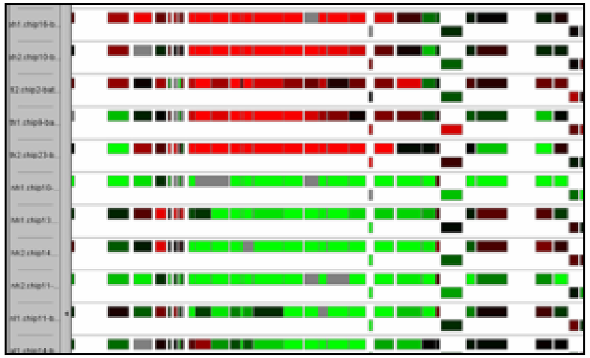


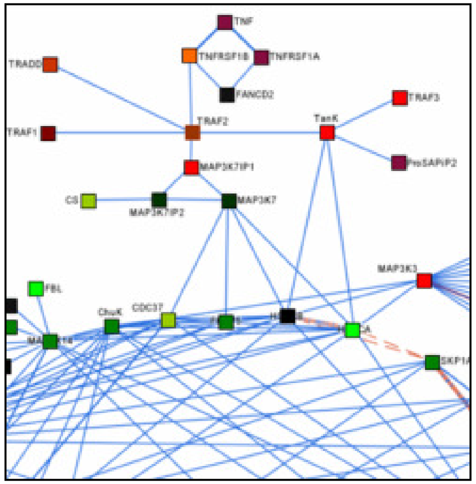
	Clofibrate vehicle_48	Clofibrate_4	DEHP vehicle_48	DEHP_4	VPA vehicle_48	VPA_4	Clofibrate vehicle_48	Clofibrate_4	DEHP vehicle_48	DEHP_4	VPA vehicle_48	VPA_4
Clof 4hr treated POOLE...	🟦	🟩	🟦	🟦	🟦	🟦		🟪				
Clof 48hr Veh POOLED_1	🟩	🟦	🟦	🟦	🟦	🟦	▨					
DEHP 4hr treated POOL...	🟦	🟦	🟦	🟩	🟦	🟦				🟧		
DEHP 48hr Veh POOLE...	🟦	🟦	🟩	🟦	🟦	🟦			🟦			
VPA 4hr treated POOLE...	🟦	🟦	🟦	🟦	🟦	🟩						🟩
VPA 48hr Veh POOLED_1	🟦	🟦	🟦	🟦	🟩	🟦					🟦	

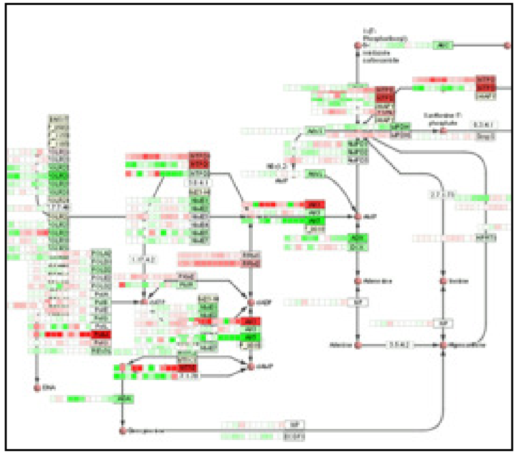
- Classifier: K Nearest Neighbours
  Distance: Positive Correlation
  K: 1
- No misclassified experiment

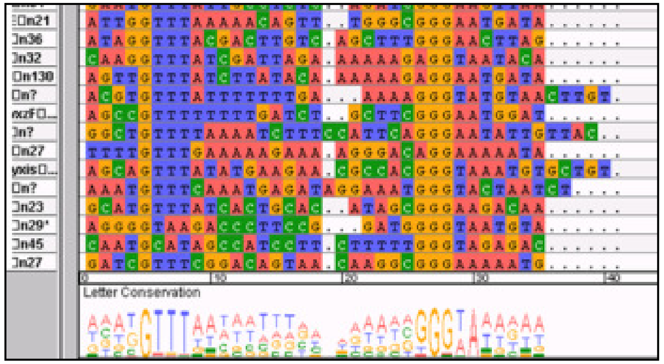
Pathway characterization and biomarker characterization

- The reference compendium and the optimal gene set provides the ideal foundation for developing sophisticated MOA models and potential biomarker identification
  - Pathway analysis
  - Genomic analysis
  - Promoter analysis
  - Protein interaction analysis, etc.









Visualization of hybridization values on individual Affymetrix Chips

- Visualization of vehicle and 4h experiments hybridization values on probe sets of some transcripts from optimal gene set
- Highly reproducible expression signals
- Small differences between vehicle and 4h experiments

Comt Cyp4a22 Acaa1 Pdk3 Slc25a30

Clof 48h vehicle

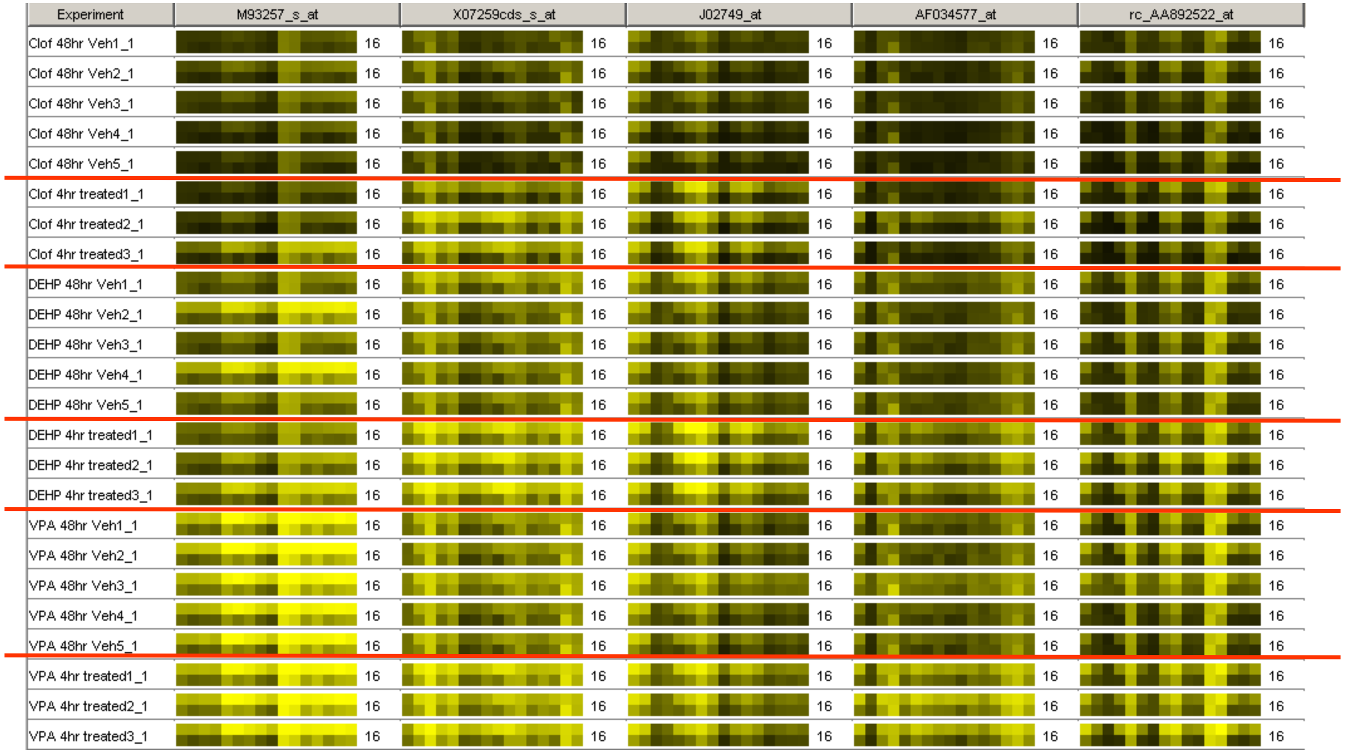
Clof 4h

DEHP 48h vehicle

DEHP 4h

VPA 48h vehicle

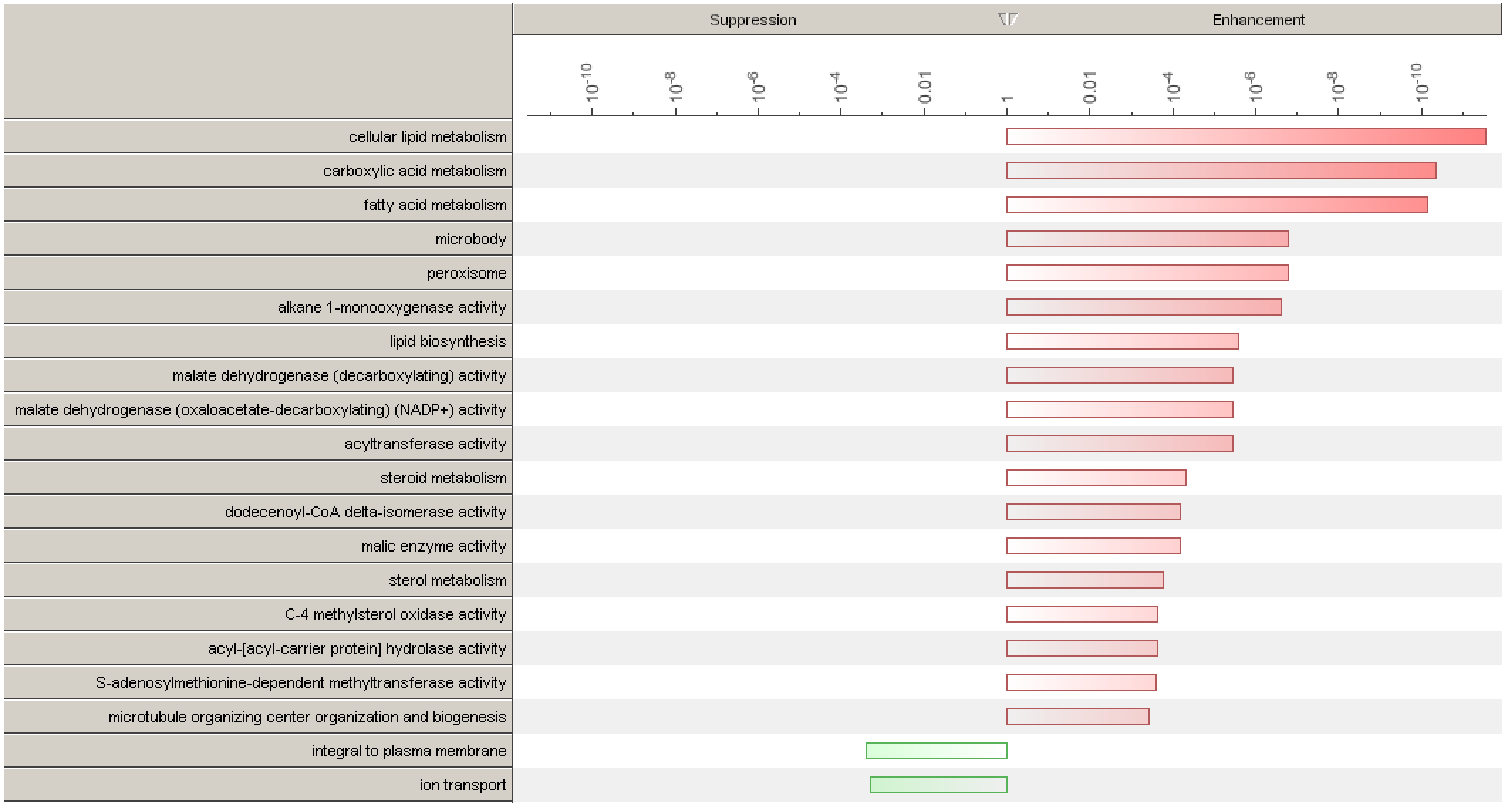
VPA 4h



Experiment	M93257_s_at	X07259cds_s_at	J02749_at	AF034577_at	rc_AA892522_at
Clof 48hr Veh1_1	16	16	16	16	16
Clof 48hr Veh2_1	16	16	16	16	16
Clof 48hr Veh3_1	16	16	16	16	16
Clof 48hr Veh4_1	16	16	16	16	16
Clof 48hr Veh5_1	16	16	16	16	16
Clof 4hr treated1_1	16	16	16	16	16
Clof 4hr treated2_1	16	16	16	16	16
Clof 4hr treated3_1	16	16	16	16	16
DEHP 48hr Veh1_1	16	16	16	16	16
DEHP 48hr Veh2_1	16	16	16	16	16
DEHP 48hr Veh3_1	16	16	16	16	16
DEHP 48hr Veh4_1	16	16	16	16	16
DEHP 48hr Veh5_1	16	16	16	16	16
DEHP 4hr treated1_1	16	16	16	16	16
DEHP 4hr treated2_1	16	16	16	16	16
DEHP 4hr treated3_1	16	16	16	16	16
VPA 48hr Veh1_1	16	16	16	16	16
VPA 48hr Veh2_1	16	16	16	16	16
VPA 48hr Veh3_1	16	16	16	16	16
VPA 48hr Veh4_1	16	16	16	16	16
VPA 48hr Veh5_1	16	16	16	16	16
VPA 4hr treated1_1	16	16	16	16	16
VPA 4hr treated2_1	16	16	16	16	16
VPA 4hr treated3_1	16	16	16	16	16

Fisher’s exact test using GO annotation

- The Fisher's Exact Test is a statistical counting test that assigns statistical significance to statements about the over- or under-representation of properties in a selection group when compared to a so-called universe group.
- Comparison of optimal gene set from gene ranking against all transcripts



Genedata
solutions in-silico

Thank you

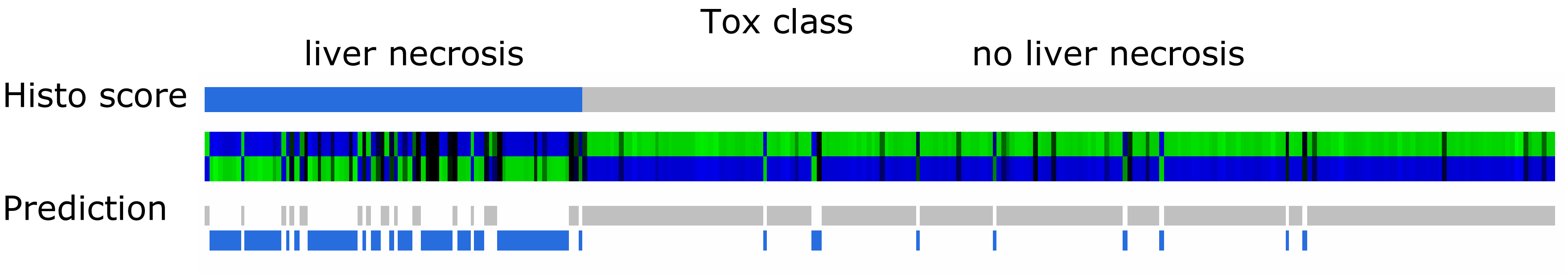
Genedata AG
Maulbeerstrasse 46
CH-4016 Basel, Switzerland
Tel +41 61 697 7651
Fax +41 61 697 7244

hans.gmuender@genedata.com

www.genedata.com

Consulting study: Prediction of histopathology 'liver necrosis'

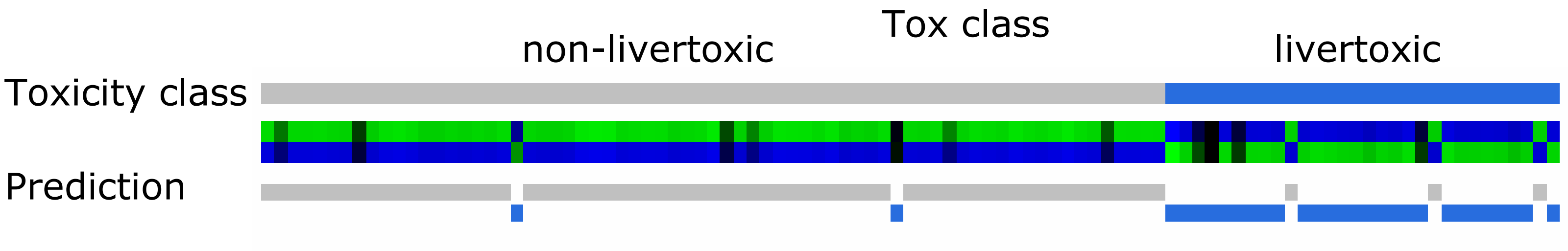
- 52 compounds tested
- Compounds applied at a low and a high concentration
- Samples taken after 6h, 24h and 72h
- Experimental data set included 1597 experiments
- Histopathological scores assigned to each experiment



- **Prediction error: ~ 10%**

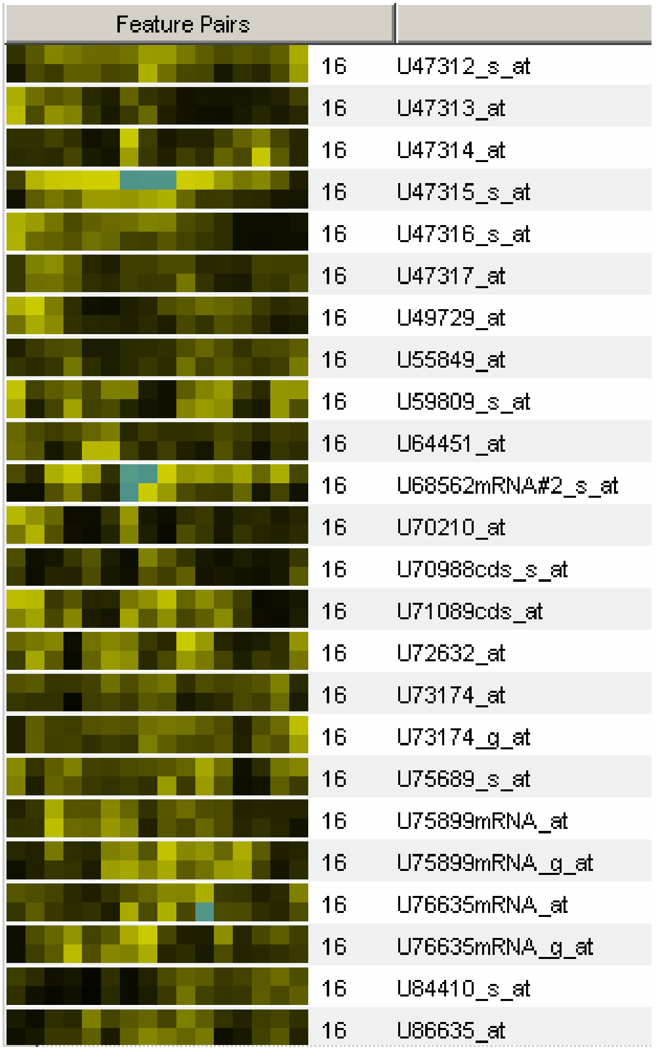
Consulting study:
Prediction of liver toxicity

- 33 compounds
- Compounds applied at a low and a high concentration
- Samples taken after 6h, 24h and 72h
- Experimental data set included 958 experiments
- Toxicity class assigned to each experiment



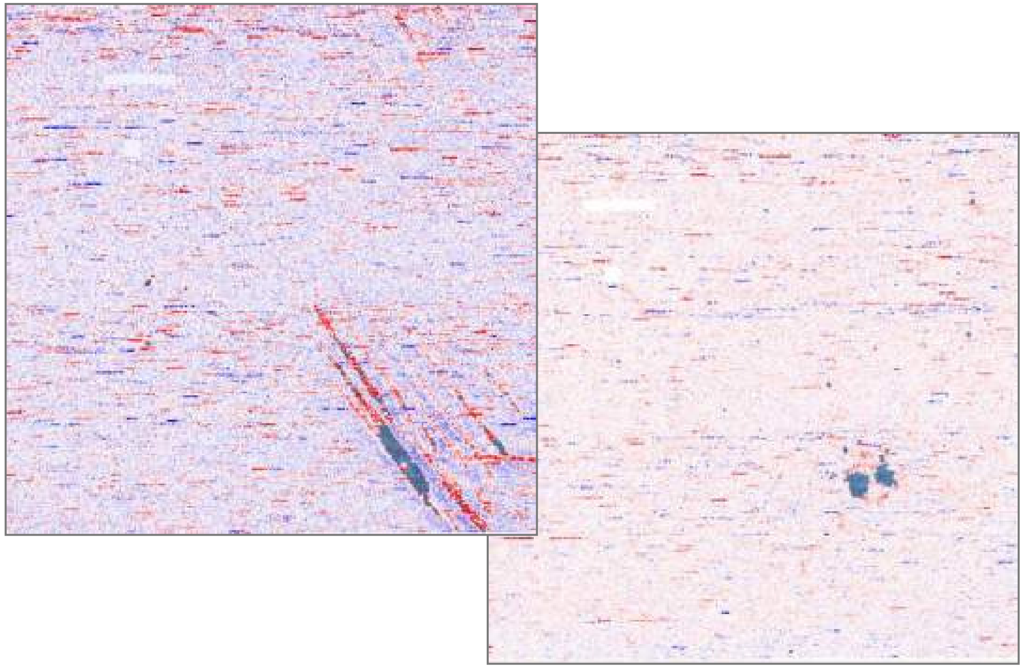
- **Prediction error: ~ 5%**

Defect detection and masking (perceived as 'spots')



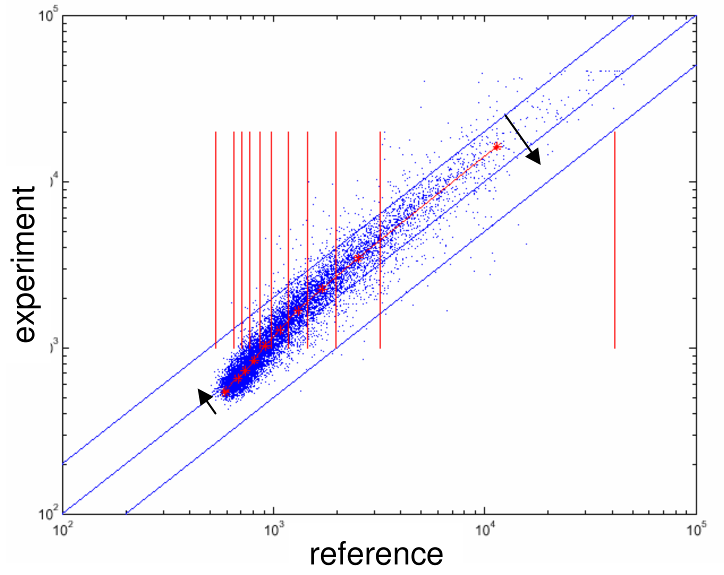
Feature Pairs	
16	U47312_s_at
16	U47313_at
16	U47314_at
16	U47315_s_at
16	U47316_s_at
16	U47317_at
16	U49729_at
16	U55849_at
16	U59809_s_at
16	U64451_at
16	U68562mRNA#2_s_at
16	U70210_at
16	U70988cds_s_at
16	U71089cds_at
16	U72632_at
16	U73174_at
16	U73174_g_at
16	U75689_s_at
16	U75899mRNA_at
16	U75899mRNA_g_at
16	U76635mRNA_at
16	U76635mRNA_g_at
16	U84410_s_at
16	U86635_at

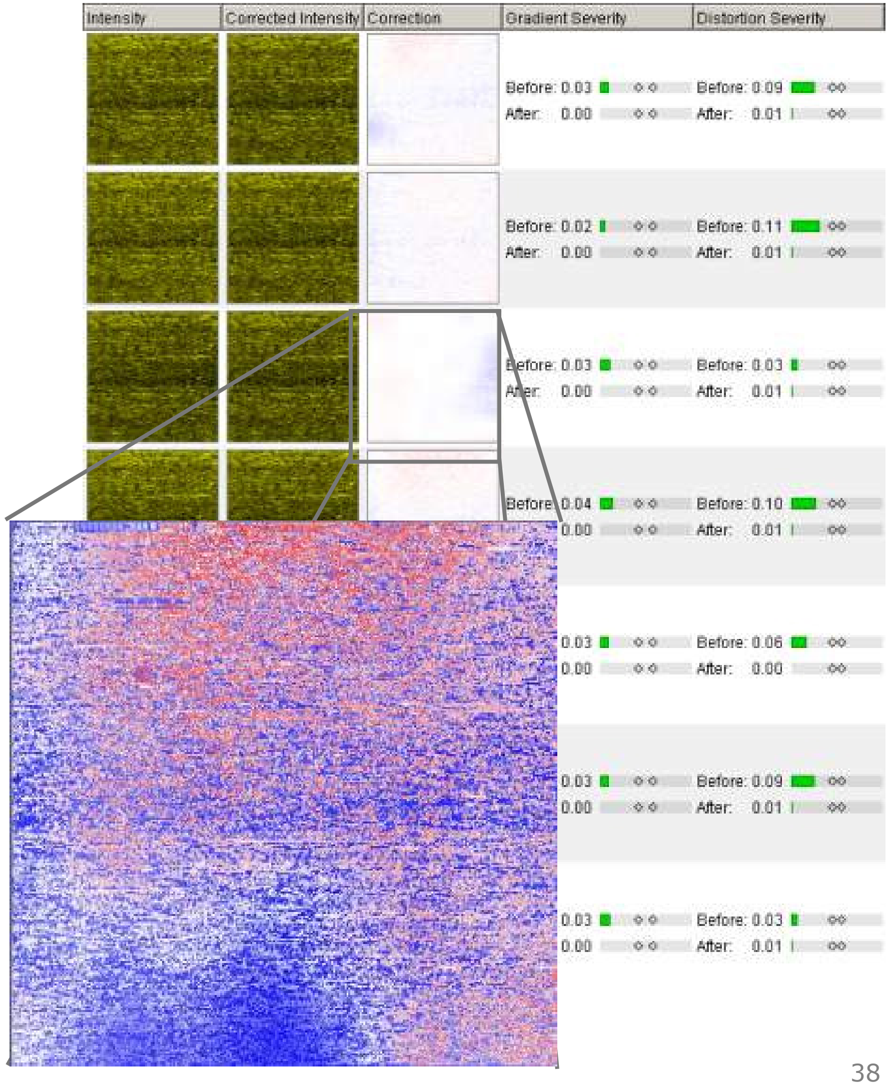
- Compares the experiments against the corresponding 'reference', identifies systematic deviations as defects and masks them
- Both dark and bright defects are detected and masked



Signal correction

- Divides the reference signal range into stripes
- In each stripe, determines the median of experiment signals
- Force this median line to be the diagonal of the new point cloud

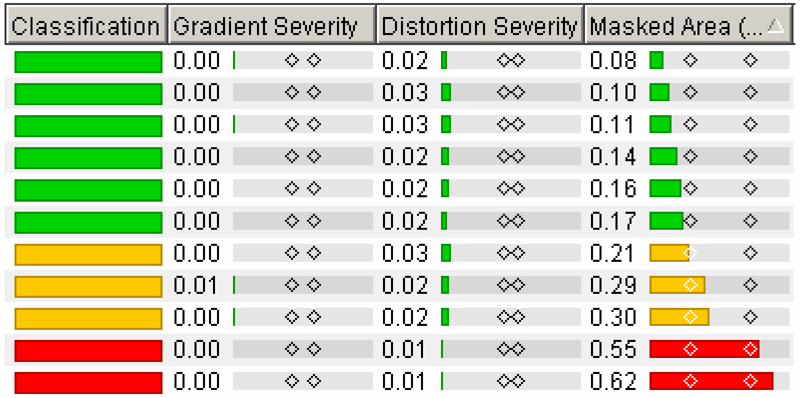




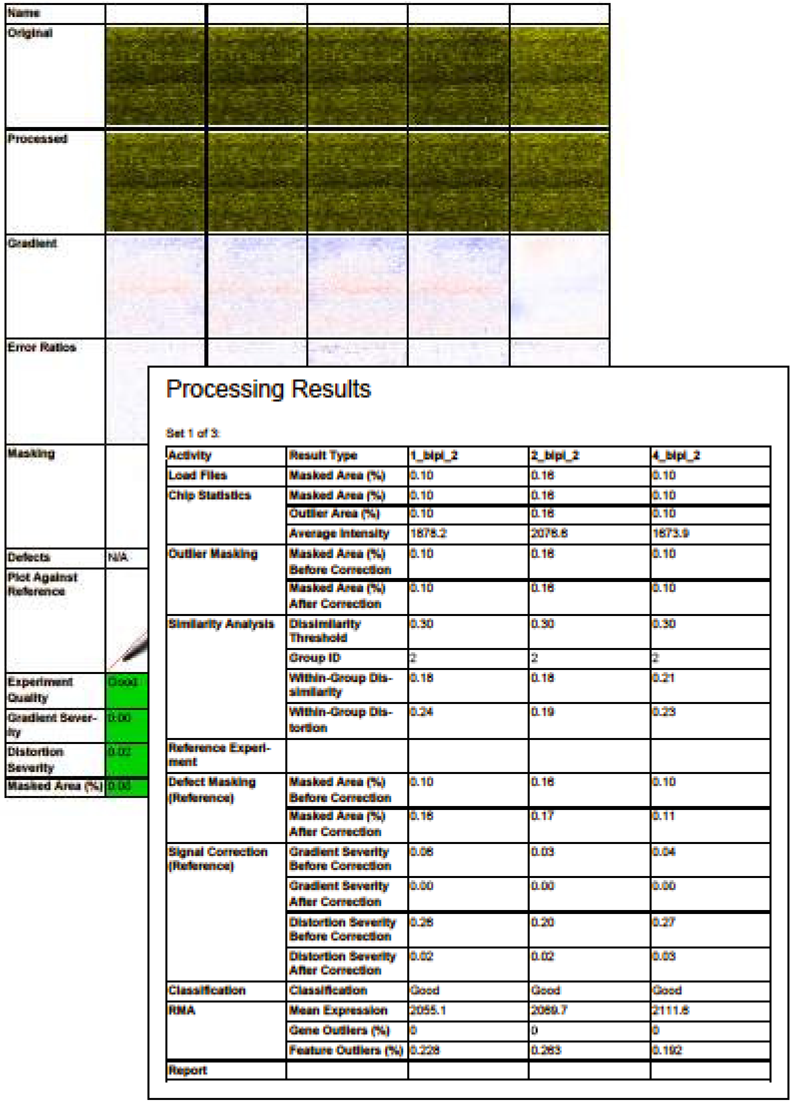
Intensity	Corrected Intensity	Correction	Gradient Severity	Distortion Severity
			Before: 0.03 After: 0.00	Before: 0.09 After: 0.01
			Before: 0.02 After: 0.00	Before: 0.11 After: 0.01
			Before: 0.03 After: 0.00	Before: 0.03 After: 0.01
			Before: 0.04 0.00	Before: 0.10 After: 0.01
			0.03 0.00	Before: 0.06 After: 0.00
			0.03 0.00	Before: 0.09 After: 0.01
			0.03 0.00	Before: 0.03 After: 0.01

Quality classification and reports

- Quality control reports are archived and directly accessible
- Quality control values are stored in database and accessible for data analysis



Classification	Gradient Severity	Distortion Severity	Masked Area (...
🟩	0.00	0.02	0.08
🟩	0.00	0.03	0.10
🟩	0.00	0.03	0.11
🟩	0.00	0.02	0.14
🟩	0.00	0.02	0.16
🟩	0.00	0.02	0.17
🟨	0.00	0.03	0.21
🟨	0.01	0.02	0.29
🟨	0.00	0.02	0.30
🟥	0.00	0.01	0.55
🟥	0.00	0.01	0.62



Name	
Original	
Processed	
Gradient	
Error Ratios	
Masking	
Defects	N/A
Plot Against Reference	
Experiment Quality	Good
Gradient Severity	0.00
Distortion Severity	0.02
Masked Area (%)	0.08

Processing Results

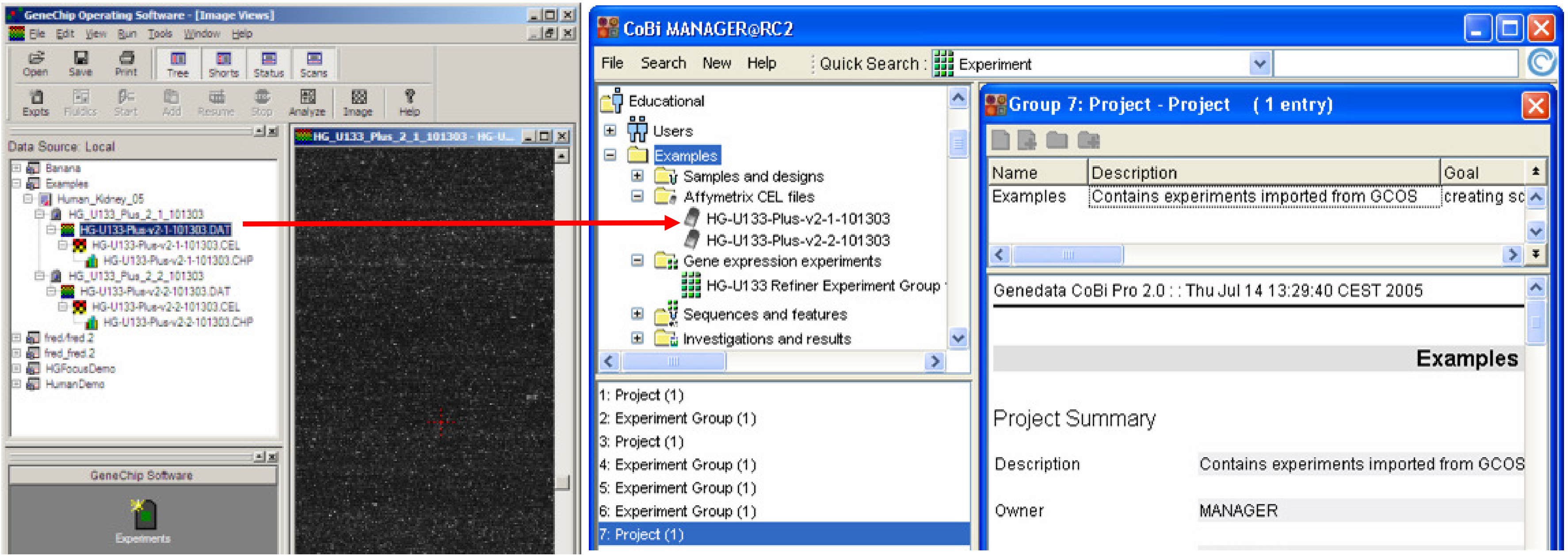
Set 1 of 3:

Activity	Result Type	1_bipi_2	2_bipi_2	4_bipi_2
Load Files	Masked Area (%)	0.10	0.16	0.10
Chip Statistics	Masked Area (%)	0.10	0.16	0.10
	Outlier Area (%)	0.10	0.16	0.10
	Average Intensity	1878.2	2076.6	1673.9
Outlier Masking	Masked Area (%) Before Correction	0.10	0.16	0.10
	Masked Area (%) After Correction	0.10	0.16	0.10
Similarity Analysis	Dissimilarity Threshold	0.30	0.30	0.30
	Group ID	2	2	2
	Within-Group Dissimilarity	0.18	0.18	0.21
	Within-Group Distortion	0.24	0.19	0.23
Reference Experiment				
Defect Masking (Reference)	Masked Area (%) Before Correction	0.10	0.16	0.10
	Masked Area (%) After Correction	0.16	0.17	0.11
Signal Correction (Reference)	Gradient Severity Before Correction	0.08	0.03	0.04
	Gradient Severity After Correction	0.00	0.00	0.00
	Distortion Severity Before Correction	0.28	0.20	0.27
	Distortion Severity After Correction	0.02	0.02	0.03
Classification	Classification	Good	Good	Good
RMA	Mean Expression	2055.1	2089.7	2111.8
	Gene Outliers (%)	0	0	0
	Feature Outliers (%)	0.228	0.263	0.192
Report				

Tight integration with Affymetrix GCOS

- Expressionist offers seamless automated import, pre-processing and analysis of data from GCOS enabling streamlined efficient microarray research

- Project structure from GCOS is conserved in Expressionist for automated standardized management of huge data sets



Name	Description	Goal
Examples	Contains experiments imported from GCOS	creating sc

Genedata CoBi Pro 2.0 : : Thu Jul 14 13:29:40 CEST 2005

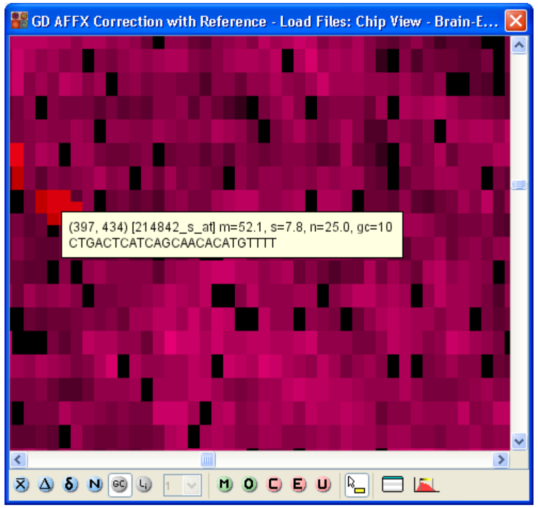
Examples

Project Summary

Description	Contains experiments imported from GCOS
Owner	MANAGER

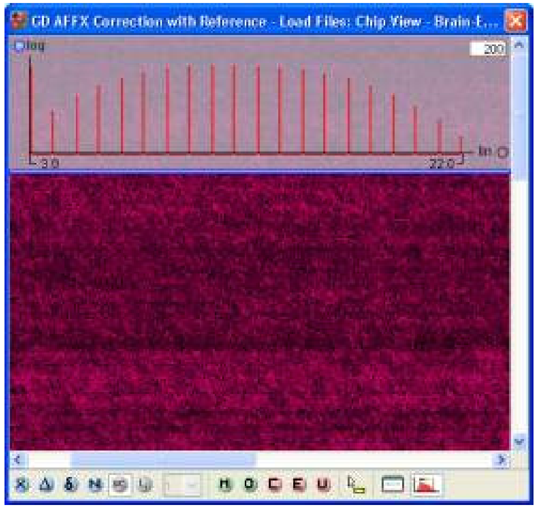
FASTA information for Affymetrix GeneChips

- Expressionist provides FASTA information for all Affymetrix GeneChips for in-depth analysis and interpretation of gene expression measurements on the nucleotide level

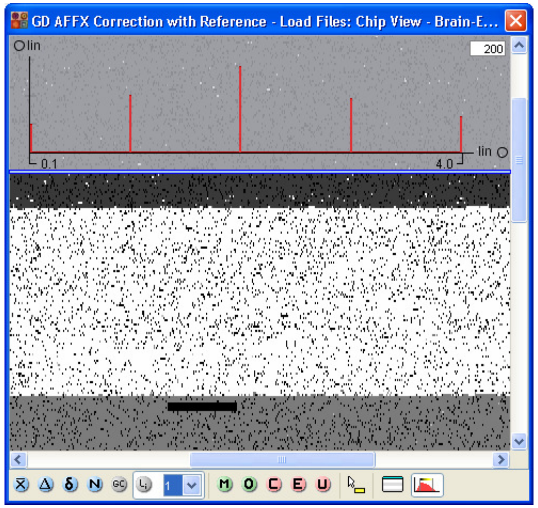


(397, 434) [214842_s_at] m=52.1, s=7.8, n=25.0, gc=10
CTGACTCATCAGCAACACATGTTTT

Tooltip view of probe set nucleotide sequence

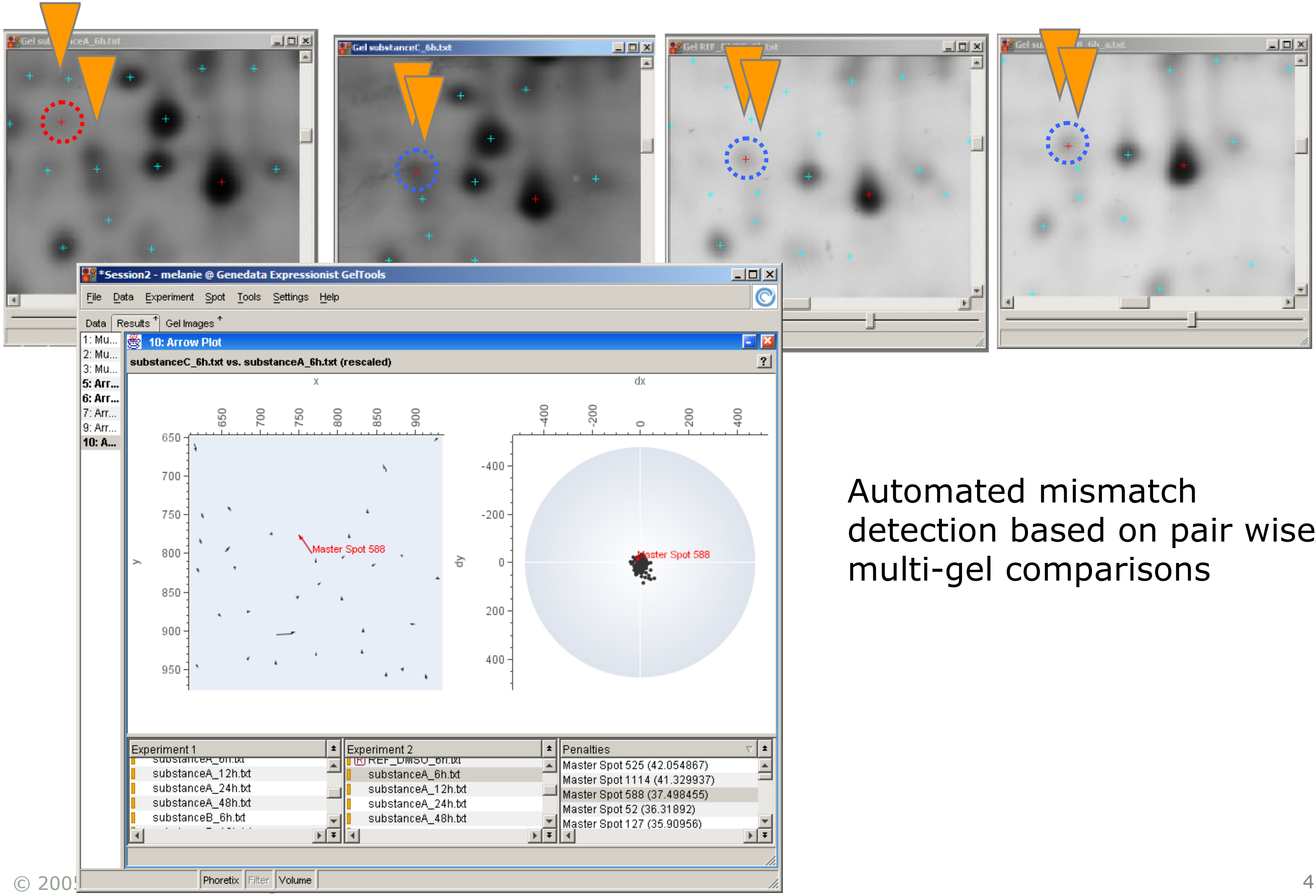


GC content view



Oligo layer view

Data quality improvement

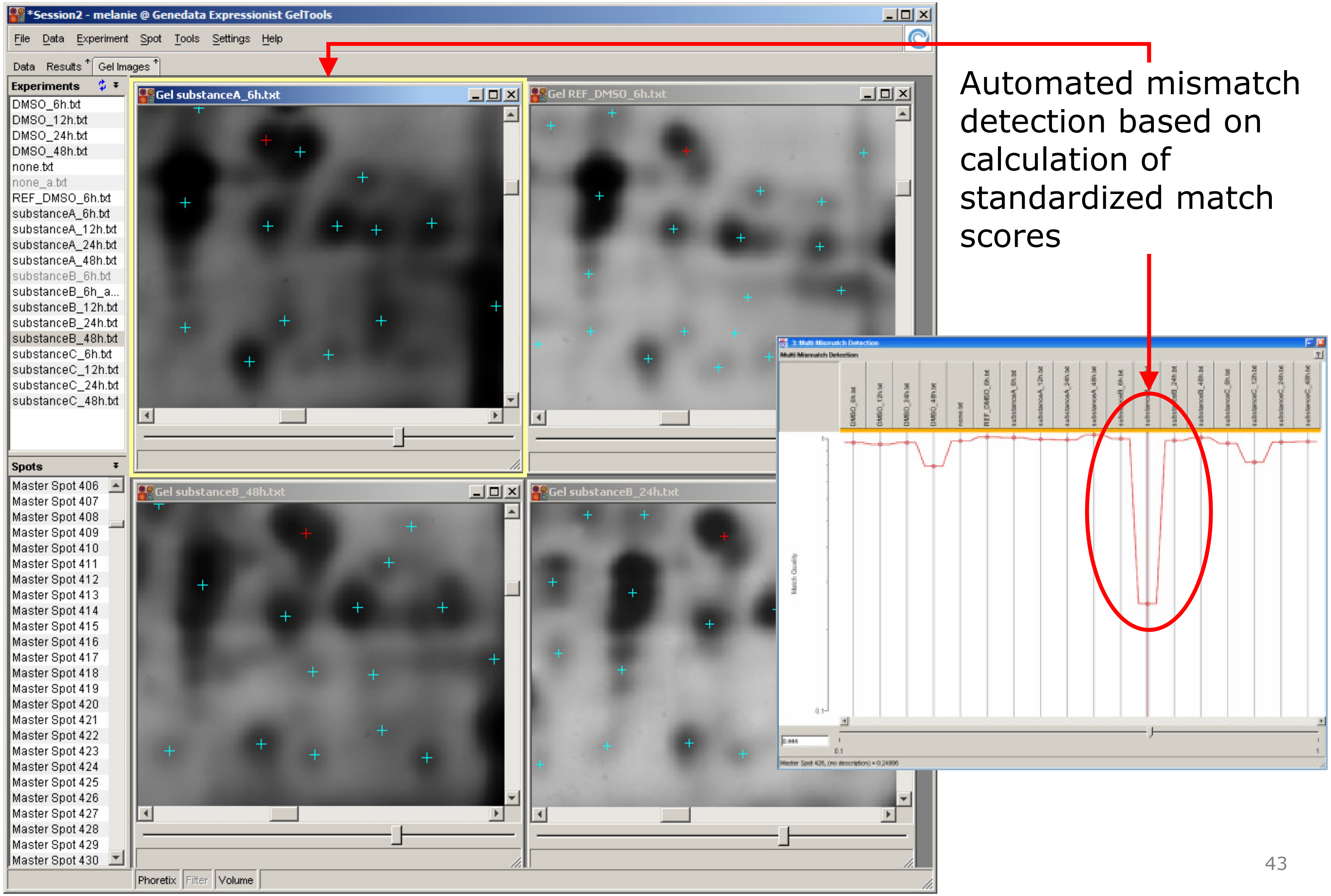


substanceC_6h.txt vs. substanceA_6h.txt (rescaled)

Experiment 1	Experiment 2	Penalties
substanceA_6h.txt	REF_DMSO_6h.txt	Master Spot 525 (42.054867)
substanceA_12h.txt	substanceA_6h.txt	Master Spot 1114 (41.329937)
substanceA_24h.txt	substanceA_12h.txt	Master Spot 588 (37.498455)
substanceA_48h.txt	substanceA_24h.txt	Master Spot 52 (36.31892)
substanceB_6h.txt	substanceA_48h.txt	Master Spot 127 (35.90956)

Automated mismatch detection based on pair wise multi-gel comparisons

Comprehensive mismatch analysis

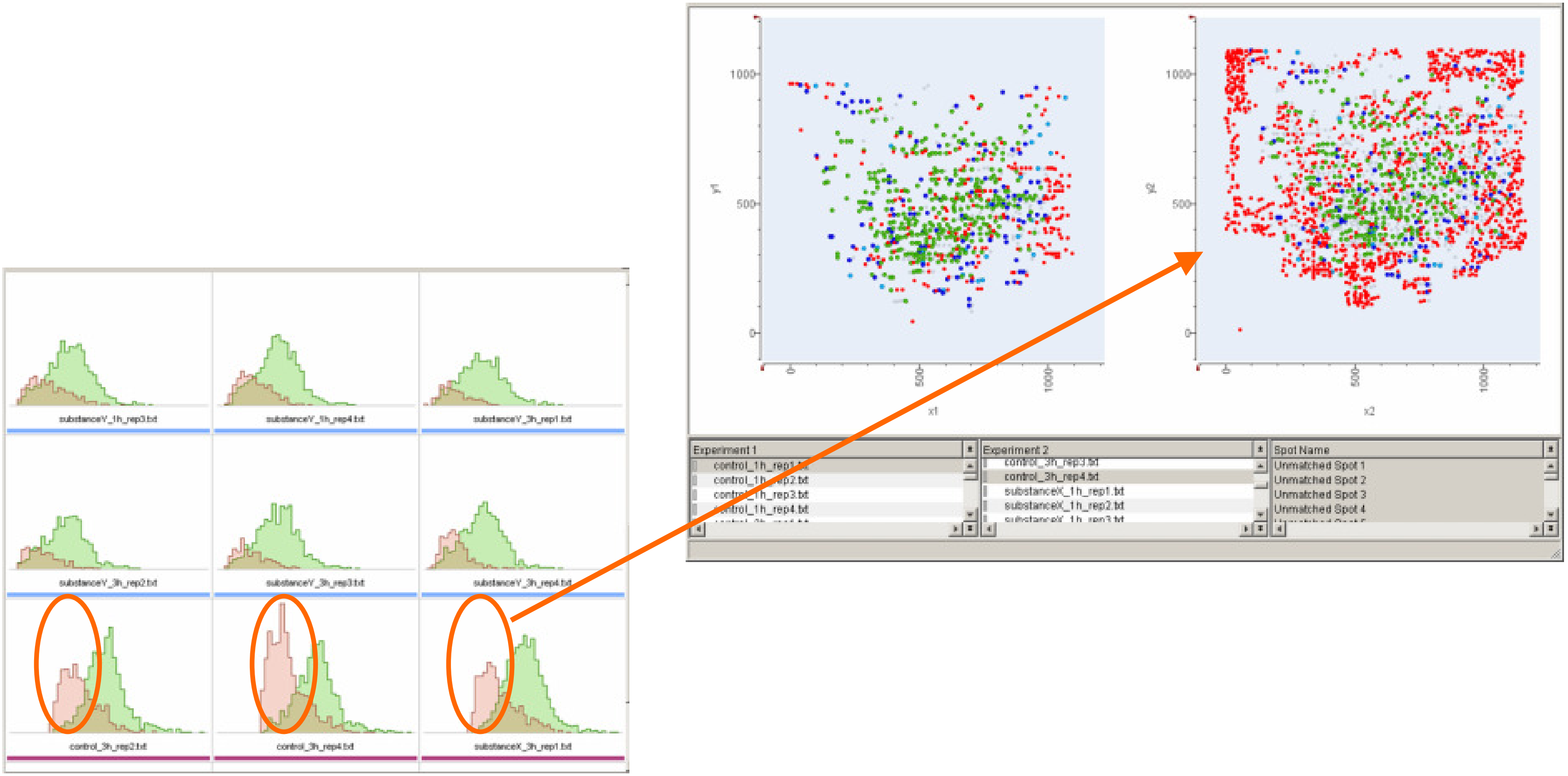


Master Spot 426, (no description) = 0.24896

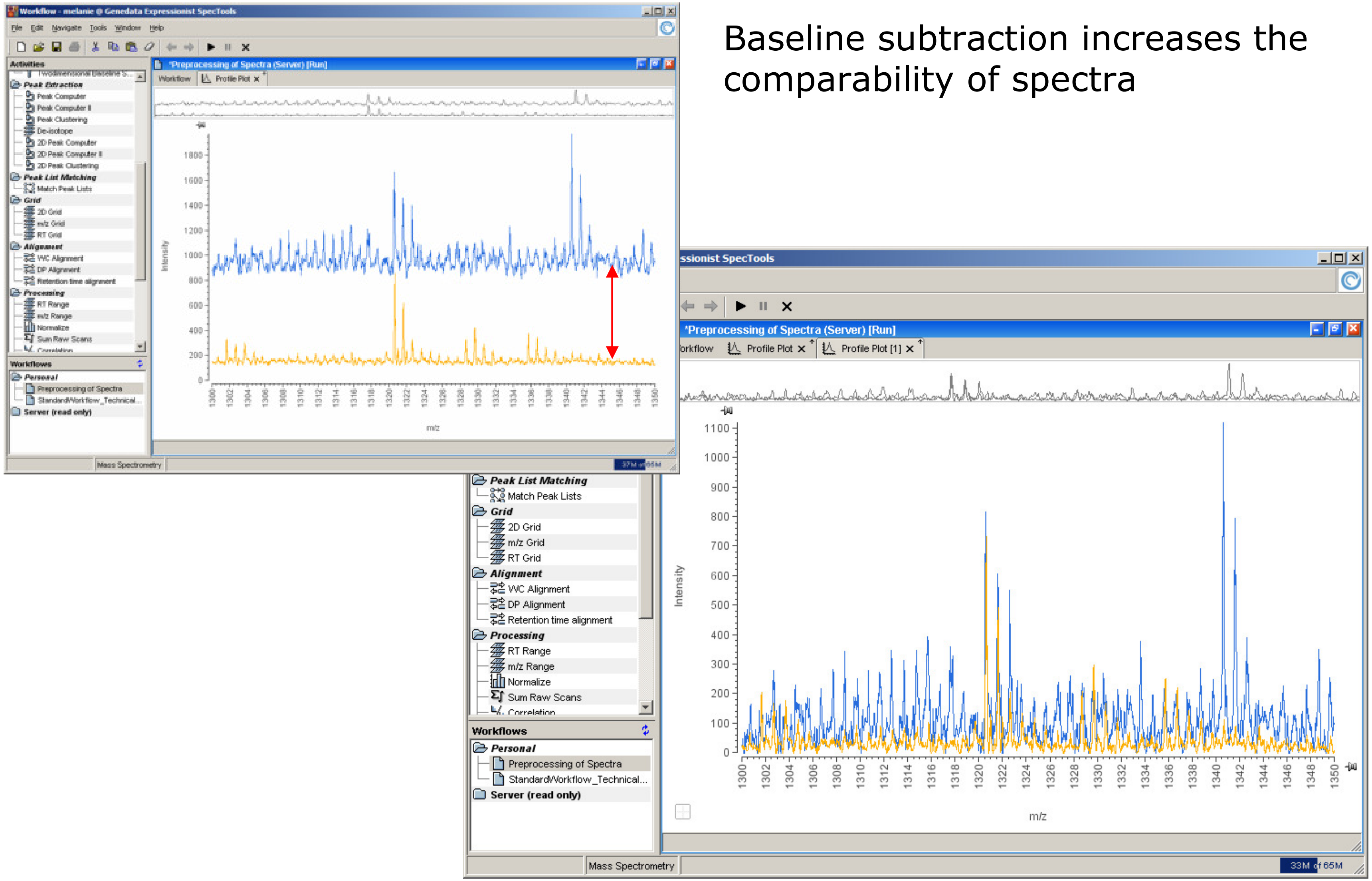
Automated mismatch detection based on calculation of standardized match scores

Signal distribution

- Localization of experimental peculiarities:
  - Low match rates due to accumulation of unmatched faint spots in a subset of gels

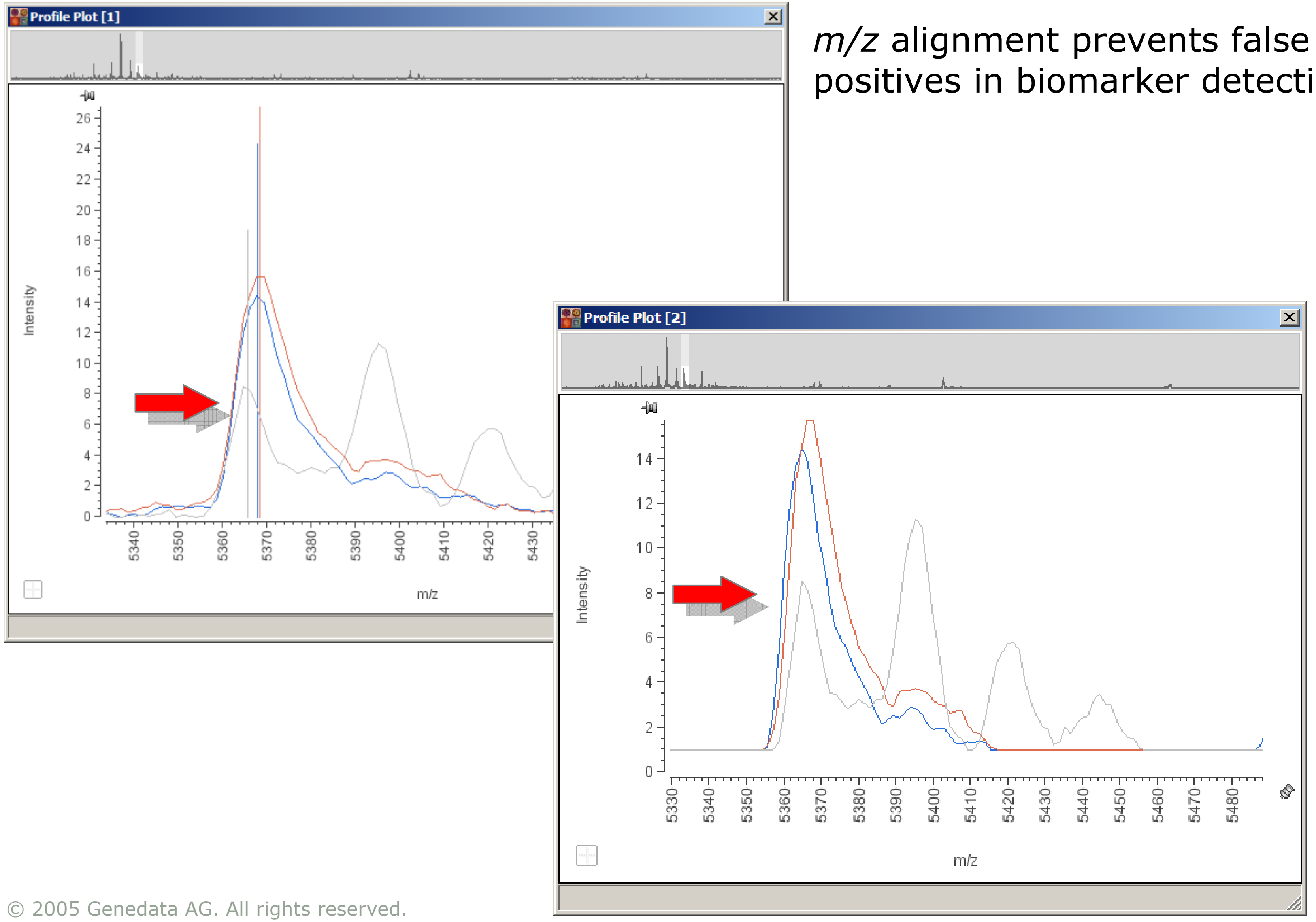


Baseline adjustment



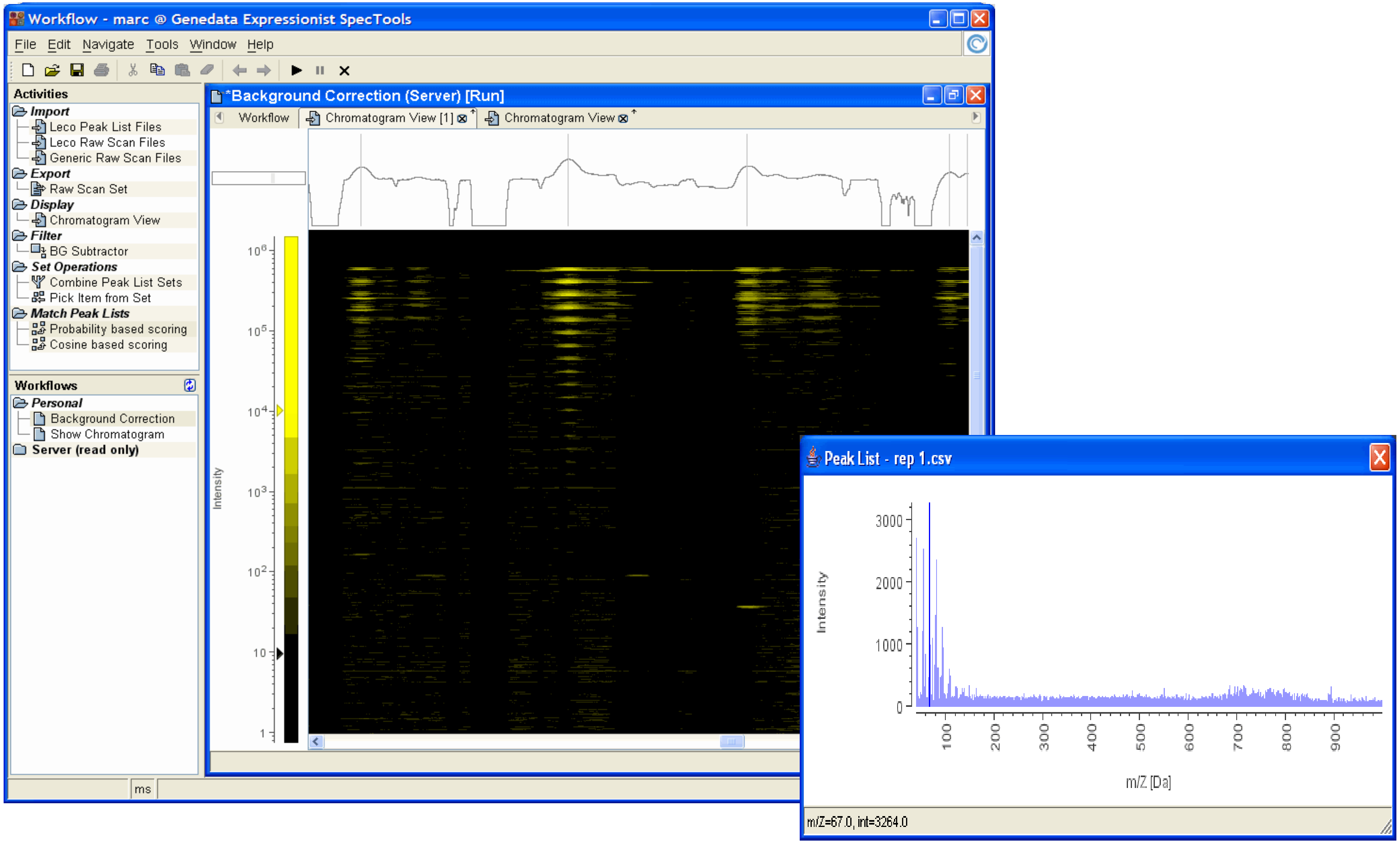
Baseline subtraction increases the comparability of spectra

m/z alignment



m/z alignment prevents false positives in biomarker detection

Background correction



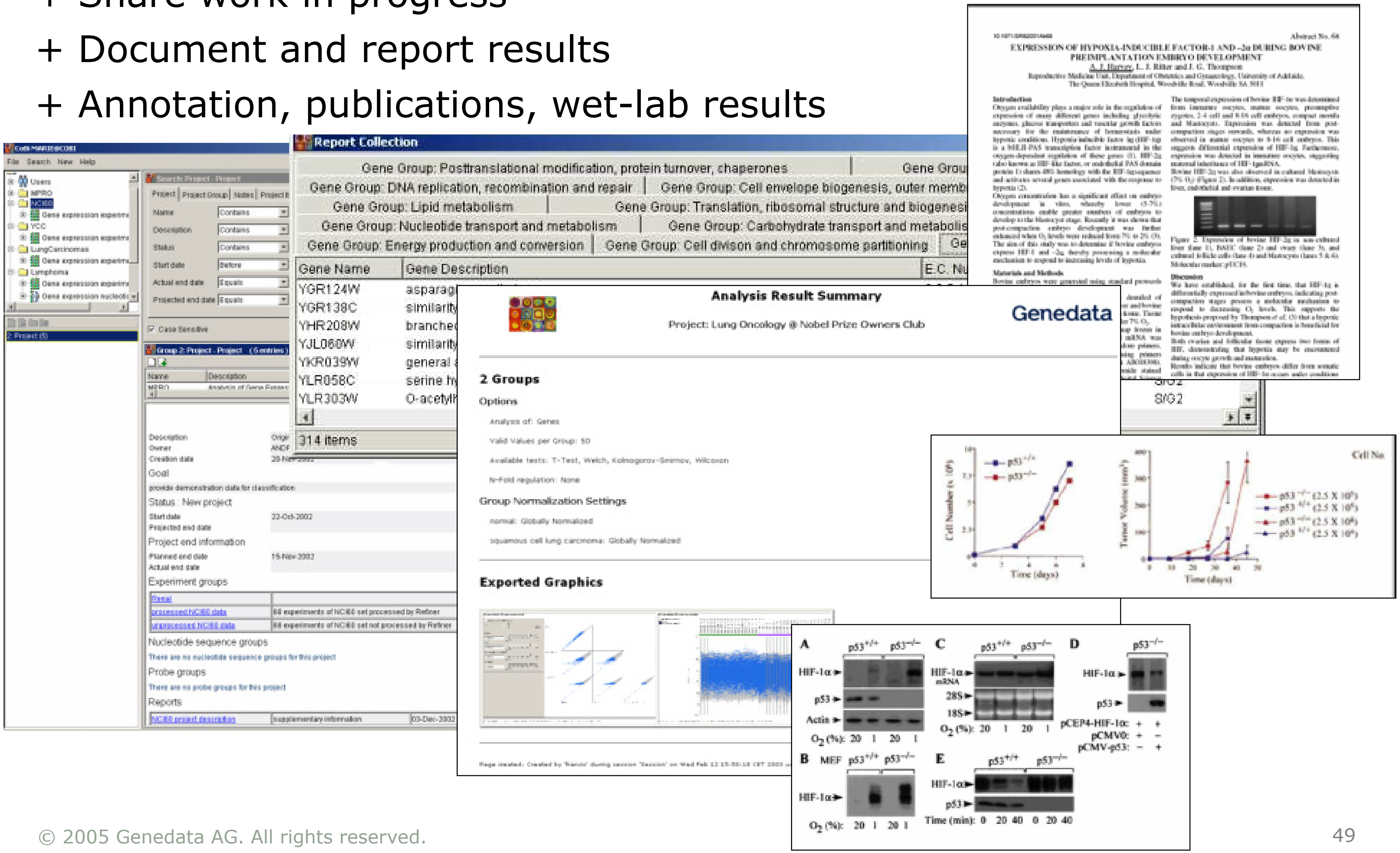
Activities in toxicogenomics

- Collaborative projects with pharmaceuticals

- Bioinformatics Partner of several EU funded Tox-related Consortia, including
  - BioCop: New technologies to screen multiple chemical contaminants in foods
  - NewGeneris: Newborns and genotoxic exposure risks
  - InnoMed: Predictive toxicology using systems biology approach

Facilitation of teamwork & project management

- + Share work in progress
- + Document and report results
- + Annotation, publications, wet-lab results



Gene Name	Gene Description
YGR124W	asparag
YGR138C	similarity
YHR208W	branched
YJL060W	similarity
YKR039W	general
YLR058C	serine h
YLR303W	O-acetyl

Description
Owner
Creation date

Goal

provide demonstration data for classification

Status : New project

Start date 22-Oct-2002
Projected end date

Project end information

Planned end date 15-Nov-2002
Actual end date

Experiment groups

Renal	
processed NCI60 data	60 experiments of NCI60 set processed by Refiner
unprocessed NCI60 data	60 experiments of NCI60 set not processed by Refiner

Nucleotide sequence groups

There are no nucleotide sequence groups for this project

Probe groups

There are no probe groups for this project

Reports

NCI60 project description	supplementary information	03-Dec-2002

Analysis Result Summary

Project: Lung Oncology @ Nobel Prize Owners Club

Genedata

2 Groups

Options

Analysis of: Genes

Valid Values per Group: 50

Available tests: T-Test, Welch, Kolmogorov-Smirnov, Wilcoxon

N-Fold regulation: None

Group Normalization Settings

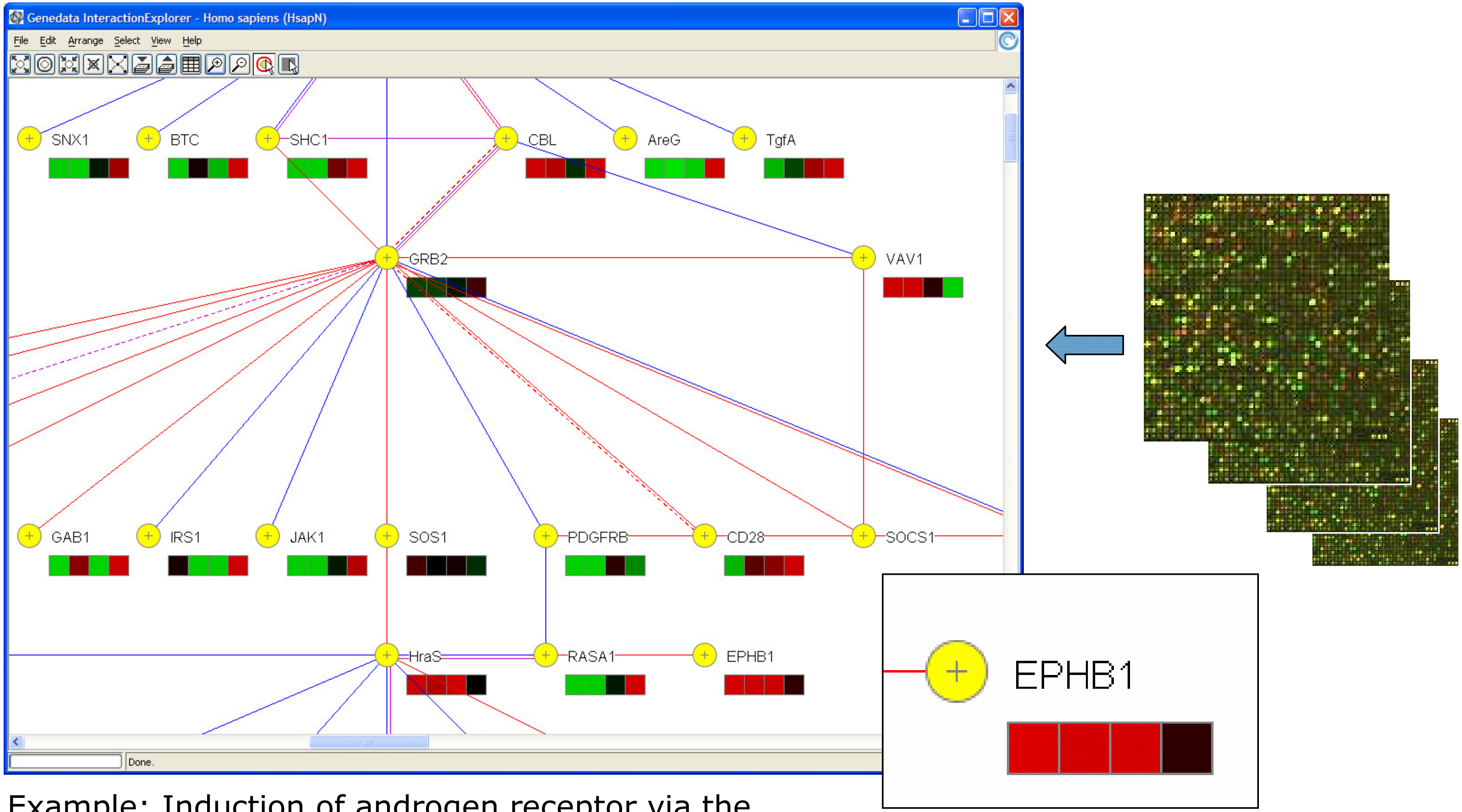
normal: Globally Normalized

squamous cell lung carcinoma: Globally Normalized

Exported Graphics

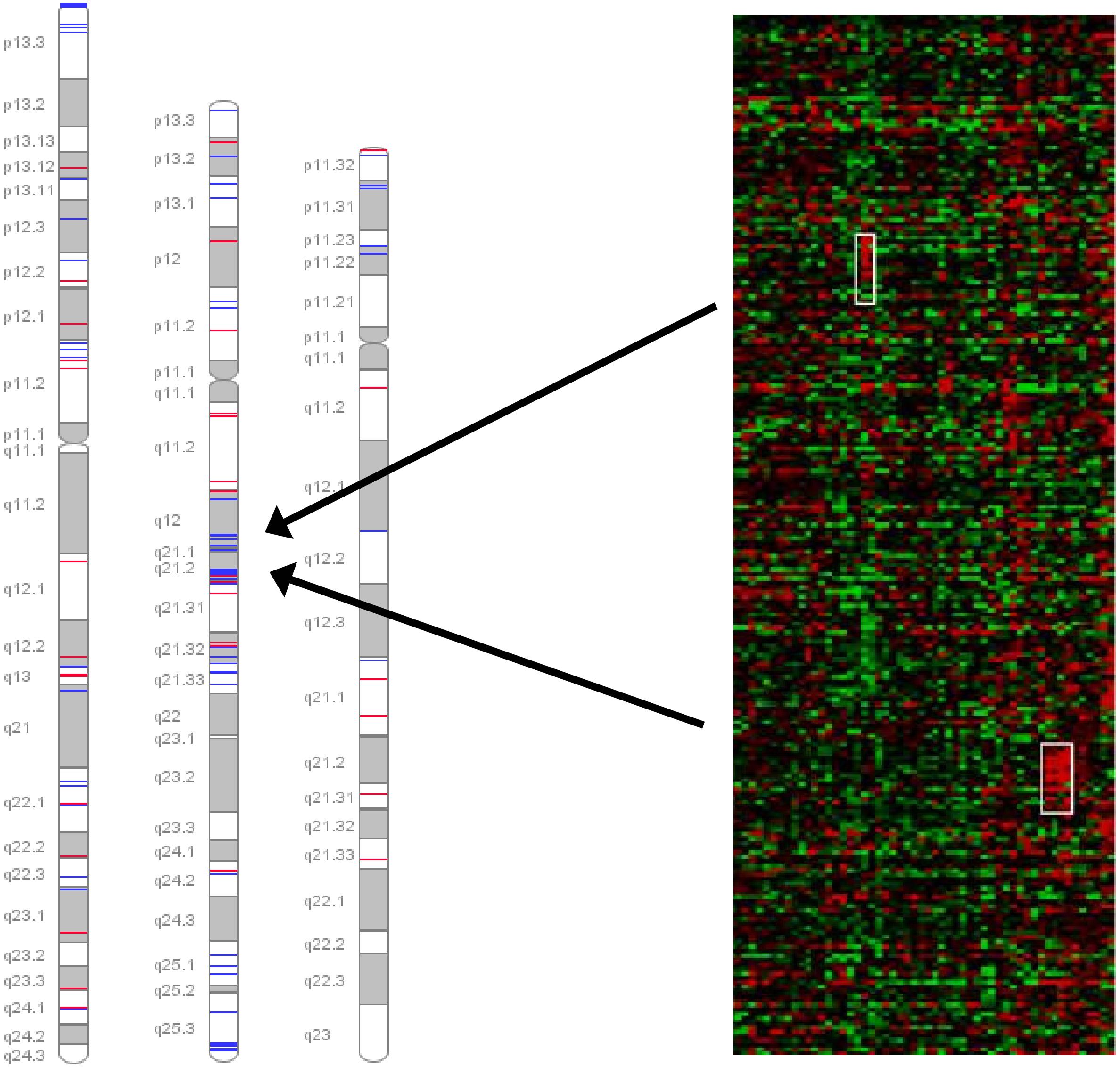
EXPRESSION OF HYPOXIA-INDUCIBLE FACTOR-1 AND -2α DURING BOVINE PREIMPLANTATION EMBRYO DEVELOPMENT

Elucidate molecular disease mechanisms - Cellular regulation



Example: Induction of androgen receptor via the MAPK signalling pathway

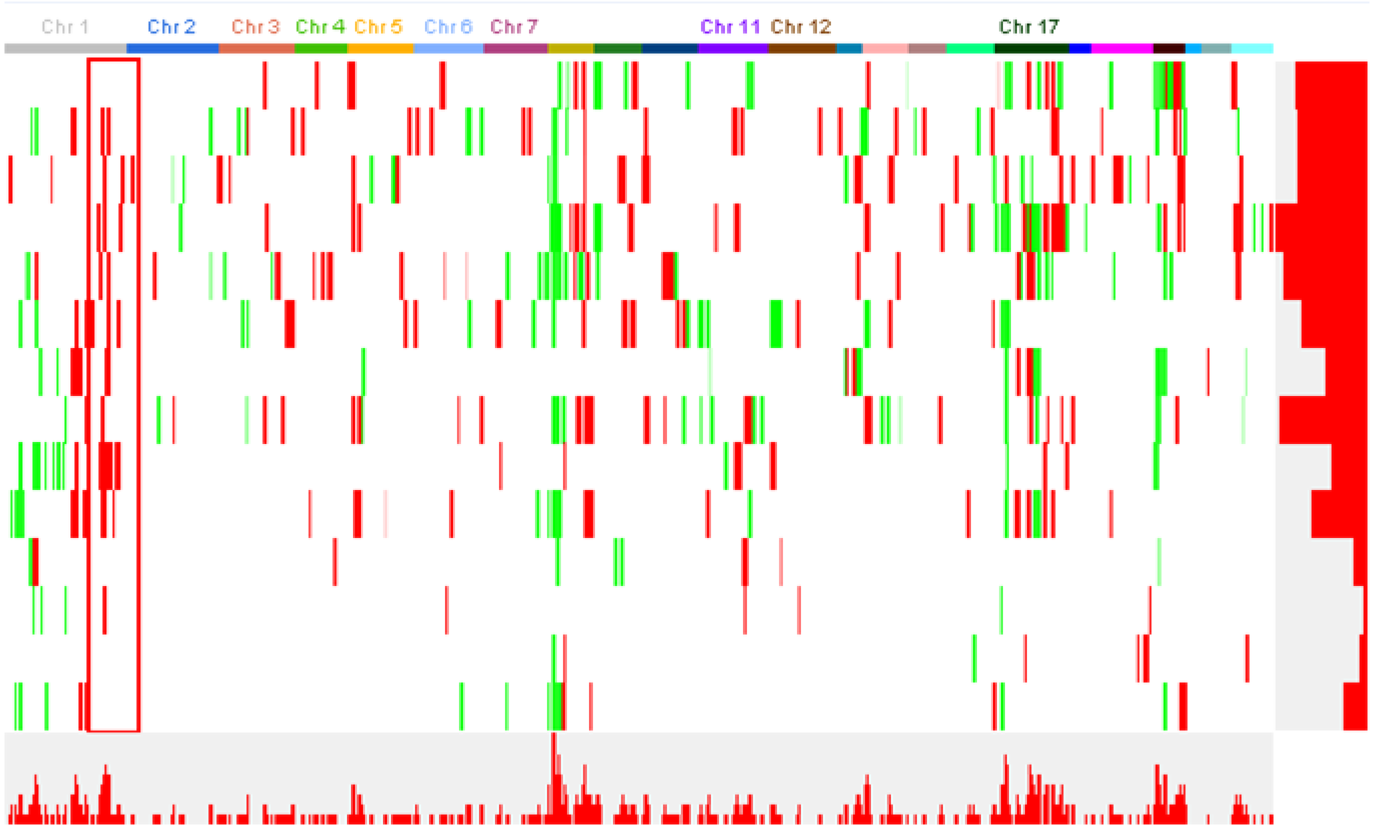
Elucidate molecular disease mechanisms - Genomic duplications



Amplicon identification

- Global view on copy number alterations

Simultaneous Amplicon identification of several experiments



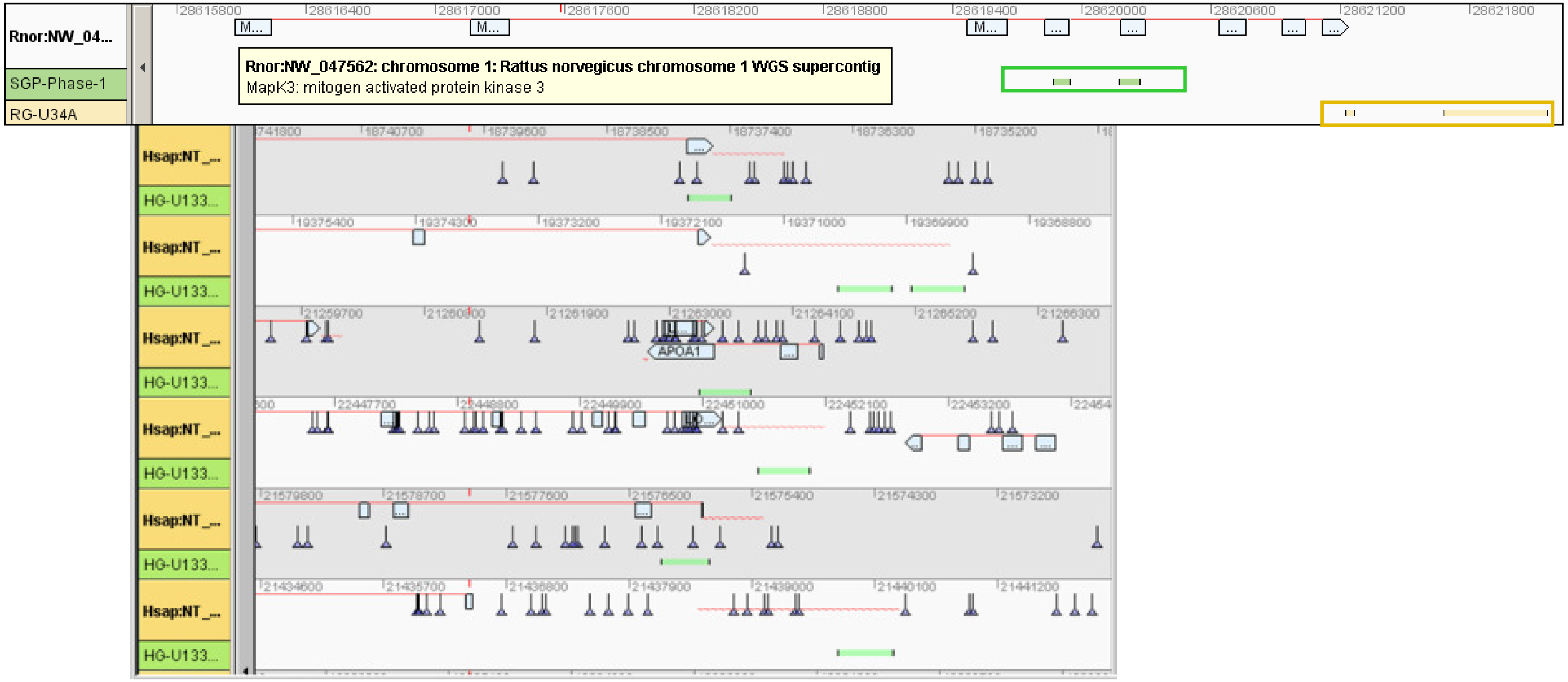
Experiments (patients)

Sum of the changes over each sample

Sum of the changes in specific chromosomal regions over all experiments

Gene expression values may depend on selected microarray sequences and on SNPs

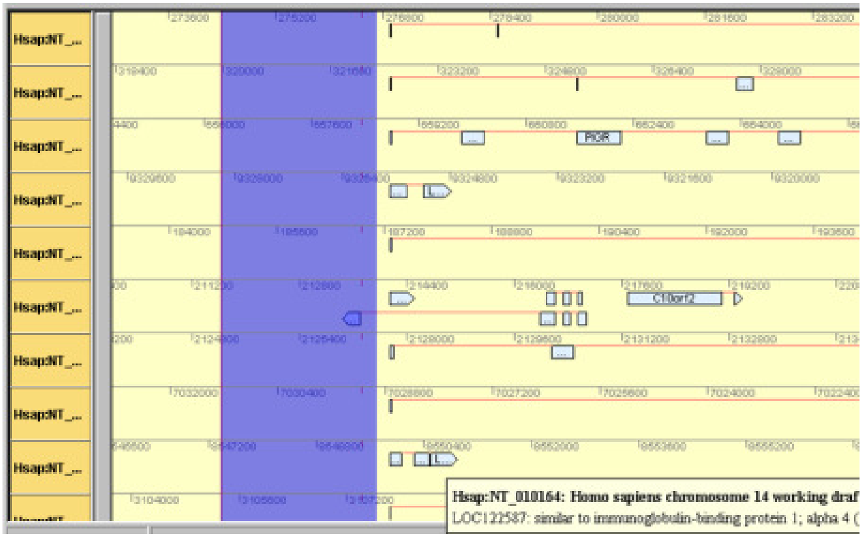
- Match all sequences represented on the microarray onto the genome
- Investigation of SNPs that may affect the oligonucleotide binding specificity, resulting in reduced hybridization signals



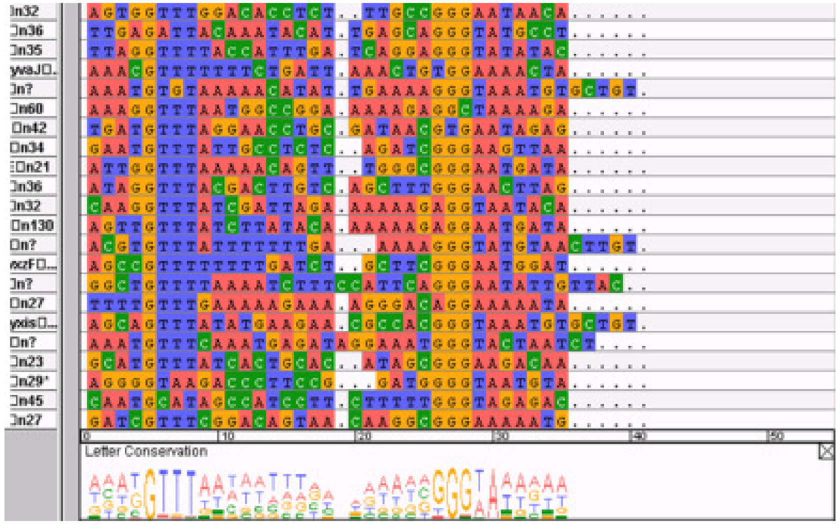
Rnor:NW_047562: chromosome 1: Rattus norvegicus chromosome 1 WGS supercontig
MapK3: mitogen activated protein kinase 3

Potential transcription factor binding sites

- Mapping and alignment onto genome and identification of upstream regions
- Selection of upstream regions of co-regulated genes for further analysis
- Identification of motifs for transcription factor binding sites

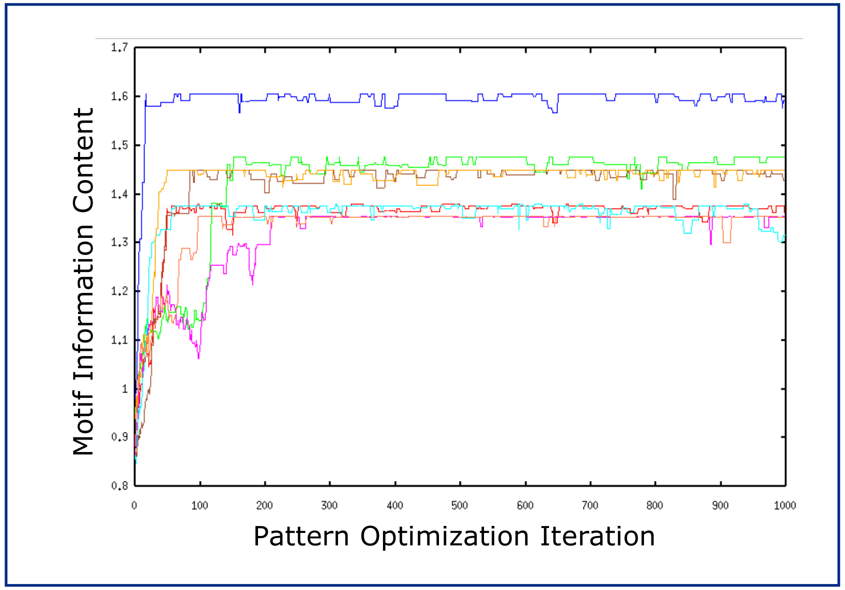


Hsap:NT_010164: Homo sapiens chromosome 14 working draf
LOC122587: similar to immunoglobulin-binding protein 1; alpha 4 (



```
n32      AGTGGTTTGGACACCTCT..TTGCCGGGAATAACA......
n36      TTGAGATTACAAATACAT.TGAGCAGGGTATGCCT......
n35      TTAGGTTTTACCATTTGA.TCAGGAGGGTATATAC......
ywaJ□..  AAACGTTTTTTTCTGATT.AAACTGTGGAAAACTA......
n?       AAATGTGTAAAAACATAT.TGAAAAGGGTAAATGTGCTGT.
n60      AAAGGTTTAATGGCCGGA.AAAAGAGGCTAAAAGA......
□n42     TGATGTTTAGGAACCTGC.GATAACGTGAATAGAG......
□n34     GAATGTTTATTGCCTCTC..AGATCGGGAAGTTAA......
□n21     ATTGGTTTAAAAACAGTT..TGGGCGGGAATGATA......
n36      ATAGGTTTACGACTTGTC.AGCTTTGGGAACTTAG......
n32      CAAGGTTTATCGATTAGA.AAAAAGAGGTAATACA......
□n130    AGTTGTTTATCTTATACA.AAAAAGAGGAATGATA......
□n?      ACGTGTTTATTTTTTTGA...AAAAGGGTATGTAACTTGT.
zF□..    AGCCGTTTTTTTTGATCT..GCTTCGGGAATGGAT......
n?       GGCTGTTTTAAAATCTTTCCATTCAGGGAATATTGTTAC..
□n27     TTTTGTTTGAAAAAGAAA.AGGGACAGGAAAAATA......
ydis□..  AGCAGTTTATATGAAGAA.CGCCACGGGTAAATGTGCTGT.
□n?      AAATGTTTCAAAATGAGATAGGAAATGGGTACTAATCT....
n23      GCATGTTTATCACTGCAC..ATAGCGGGAAGACAA......
n29'     AGGGGTAAGACCCTTCCG...GATGGGGTAATGTA......
n45      CAATGCATAGCCATCCTT.CTTTTTGGGTAGAGAC......
n27      GATCGTTTCGGACAGTAA.CAAGGCGGGAAAAATG......
```

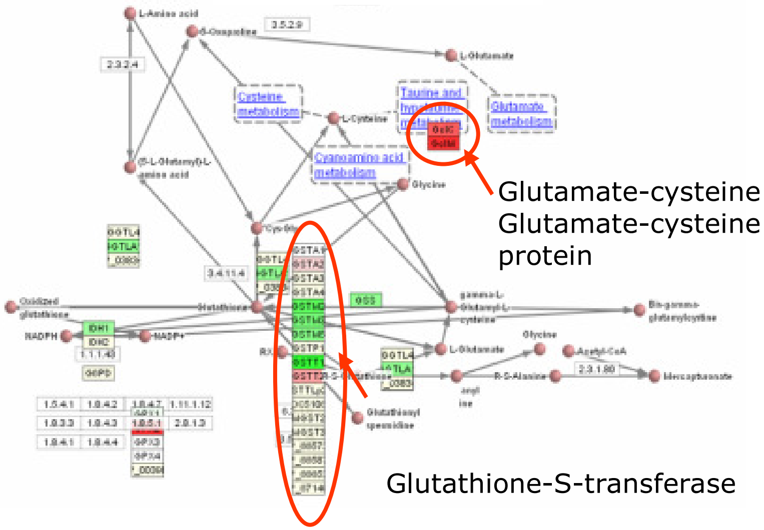
GTTTAA-N(12-15)-GGGTA



Changes of gene expression of enzymes in metabolic pathways

- Storing, editing and viewing standard and custom pathway maps, as well as mapping mRNA profiles onto metabolic or regulatory pathways
- Time series expression profiling

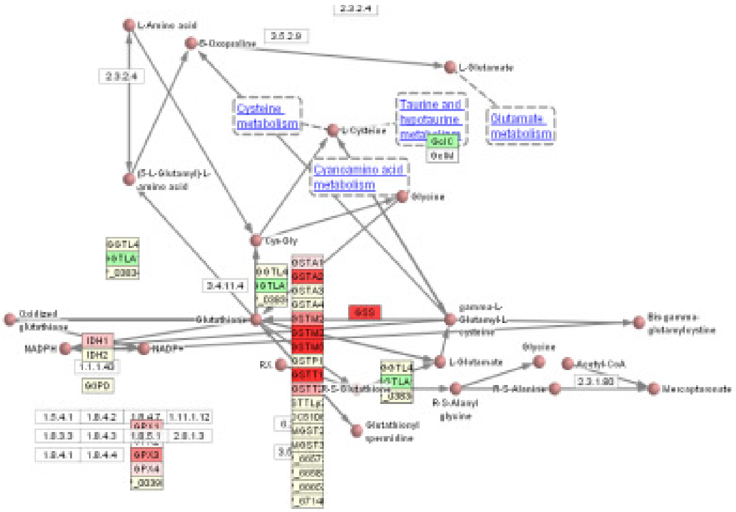
Paracetamol 6h



Glutamate-cysteine ligase
Glutamate-cysteine ligase regulatory protein

Glutathione-S-transferase

Paracetamol 48h



Analysis of co-expression and co-regulation

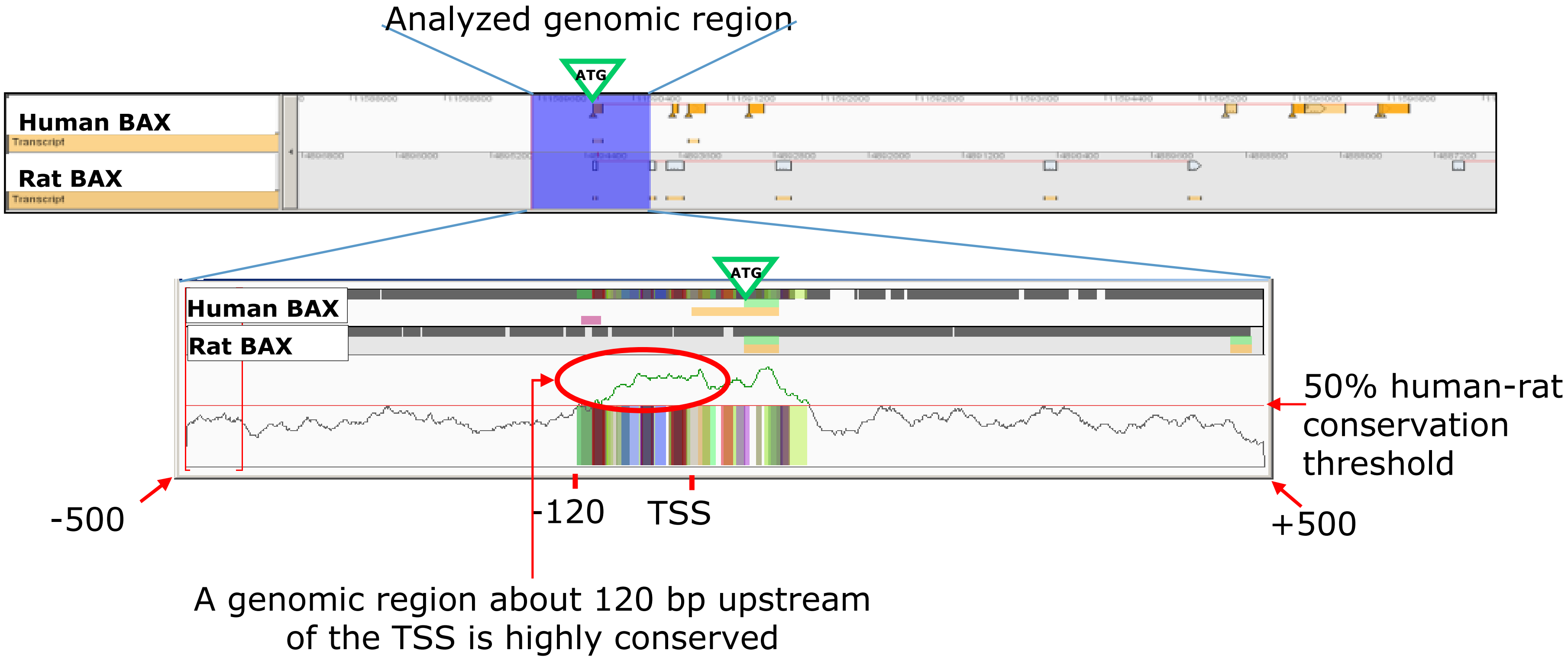
- A number of genes in rat liver were found to be de-regulated by genotoxic hepatocarcinogens [Ellinger et al. 2004]
- Transcription of a small set of genes behaves similarly (co-expression), suggesting a common molecular mechanism for gene regulation (co-regulation)
- A subset of the co-expressed genes are known p53 targets

- Are there other transcription factors that might synergize with p53 to coordinate the expression of genes that are induced by genotoxic hepatocarcinogens?

- To generate new hypotheses different in silico-approaches were used to characterize the promoters of those genes
  - Genome-genome comparisons (“phylogenetic footprinting”) a powerful method to deduce regulatory regions in orthologous regions from different species
  - Use of libraries of experimentally derived Transcription Factor Binding Site (TFBS) models for predicting putative TFBSs

Comparison of the human and rat BAX gene and identifying conserved upstream regions

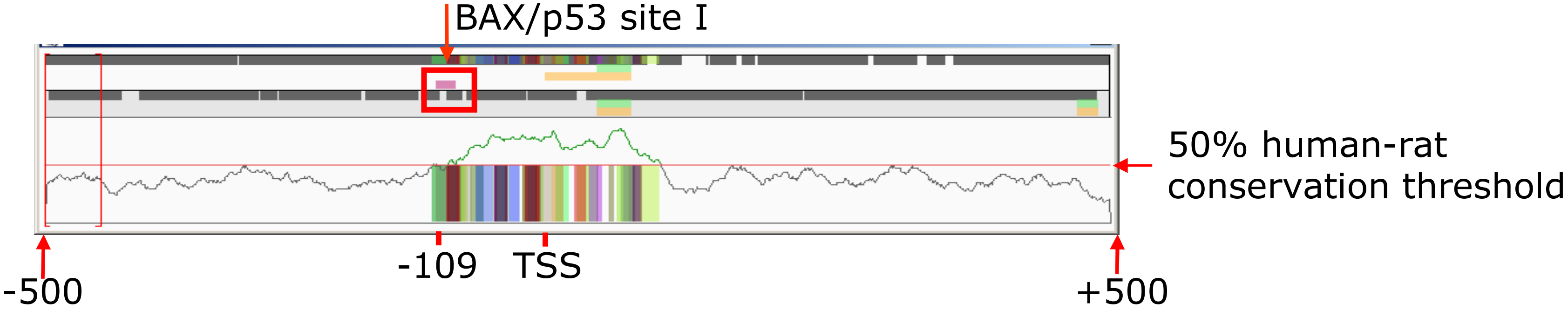
- One major application of phylogenetic footprinting is to screen for biologically relevant Transcription Factor Binding Sites (TFBS) based on Position Weight Matrices (PWMs)



A genomic region about 120 bp upstream of the TSS is highly conserved

Identification of human-rodent conserved p53 DNA-binding sites

- In the BAX promoter region one putative well conserved p53 binding site could be identified upstream of the TSS (site I)

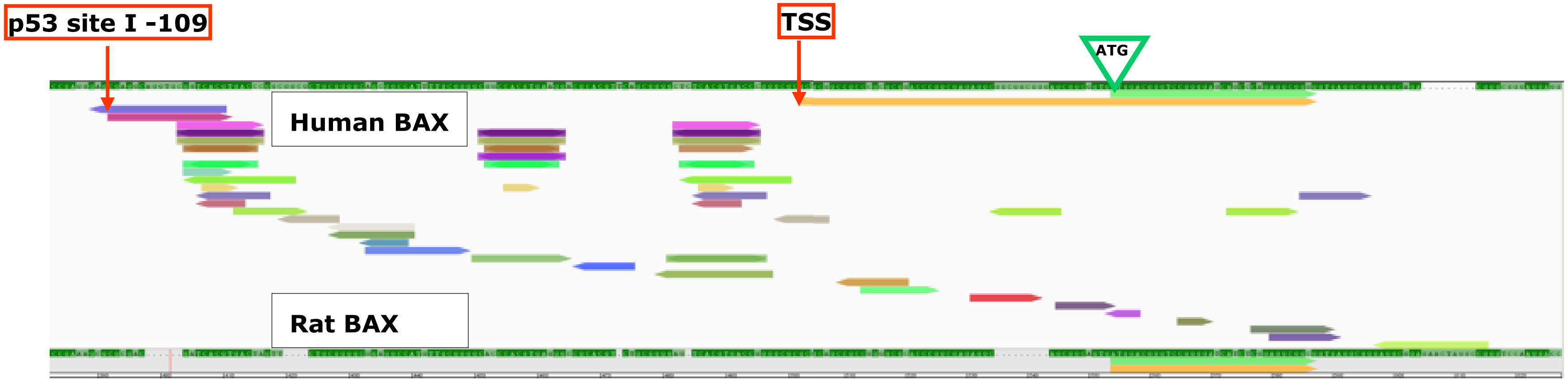


- Two other p53 binding sites can be identified in less conserved regions. One is located further upstream of the TSS (-421 bp; site II), and another in the first intron (+329 bp; site III)

- **Phylogenetic footprinting pinpointed an additional p53 binding site candidate (site III)**
- **Future investigation might reveal the functional relevance of this site**

Identification of additional relevant TFBSs

- The in-silico analysis suggests that besides p53 other mammalian transcription factors that bind in the vicinity of the p53 site might be involved in the regulation of BAX



- **At least 16 TFBS sites could be found in the vicinity of p53 sites that are significantly overrepresented in the regulatory regions of genes shown to be co-expressed under genotoxic stress**
- **These factors might cooperate with p53 in the transcriptional activation caused by genotoxic hepatocarcinogens**

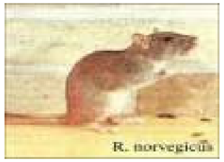
Predictive Tox Database

Compounds

- Compound
- Compound class
- Concentration
- Treatment time
- Dosing route
- Dosing frequency
- Vehicle
- Endpoints
- …

Animals

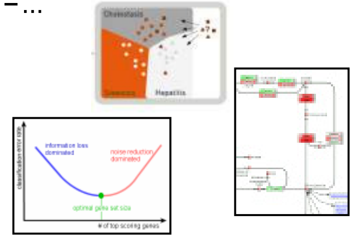
- Species
- Strain
- Sex
- Age
- Weight
- Observations
- …



MOA, MOT
Biomarkers
Predictability

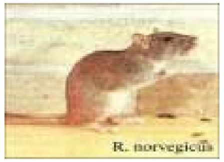
MOA, Tox Prediction

- Transcript/protein/metabolite annotation
- Experiment annotation
- MOA classification
- Biomarker candidates
- Tox Mechanism
- Tox Prediction
- …



Treated animals

- Tissue
- Histopathology
- Clinical endpoints
- Clinical chemistry
- Serum chemistry
- Urine chemistry
- Hematology
- …



-omics data

- 1 channel data
- 2 channel data
- 2D gel
- LC/MS, GC/MS
- NMR
- Raw data
- Processed data
- Expression values
- Classification
- …