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

May 2006

Dr. Hans Gmuender
Scientific Consultant, Genedata AG, CH 4016 Basel, Switzerland

Drug safety and interest in toxicogenomics

- Pre-clinical and clinical safety issues account for 20% of new drug failures
- 80% of drug development costs occur in late stages of drug development
- Yet only after marketing drugs rare side effects are discovered
- The public is increasingly intolerant towards side effects

- Rat toxicity experiments accurately predict human toxicity in less than 50%
- Using two model organisms the accuracy is less than 70%

➔ **Increase predictability**

➔ **Characterize toxic MoA**

➔ **Identify cross-species biomarkers**

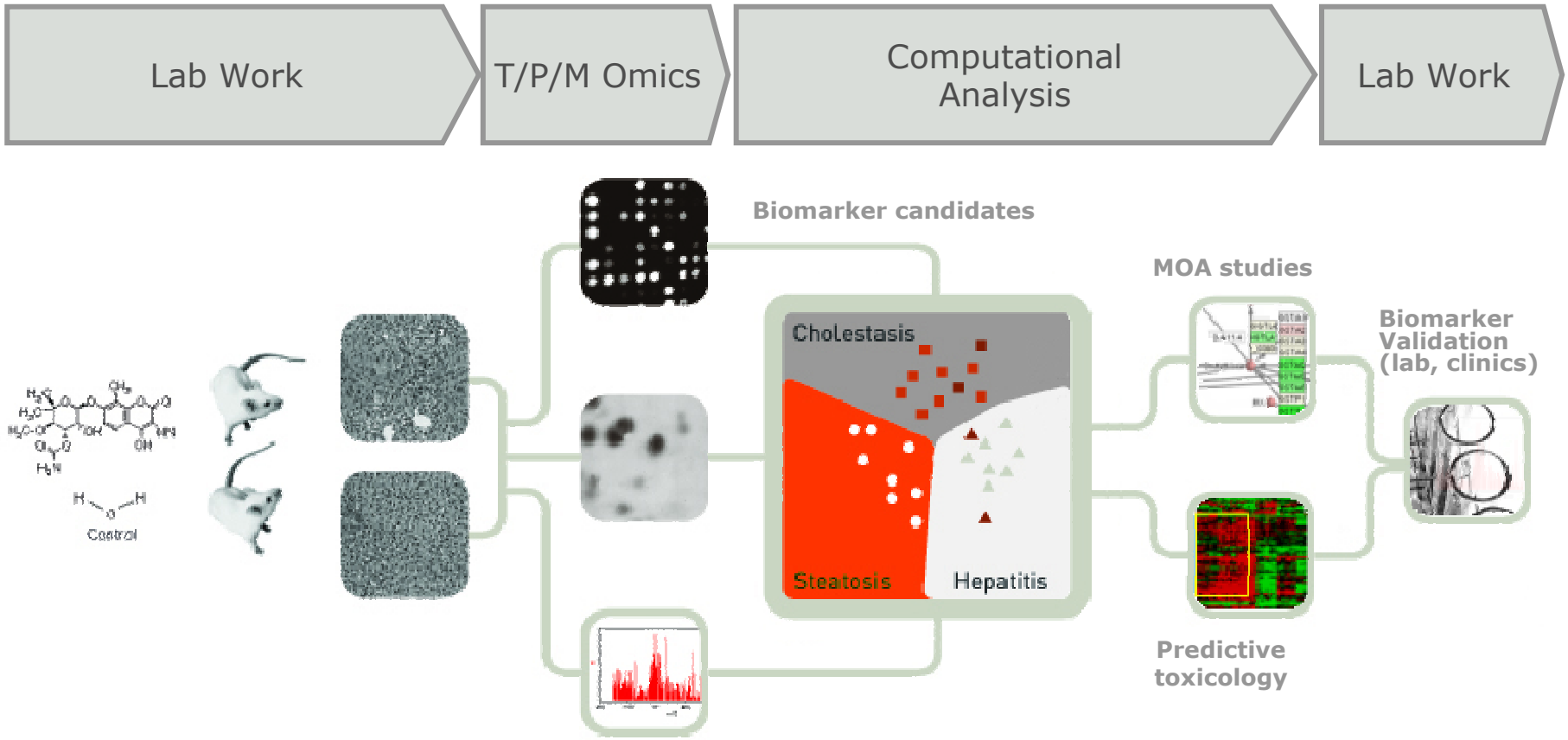
Toxicogenomics

- 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

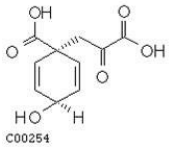
Research process for Toxicogenomics



Tox Database

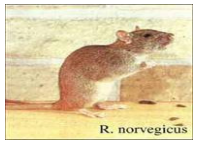
Compounds

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



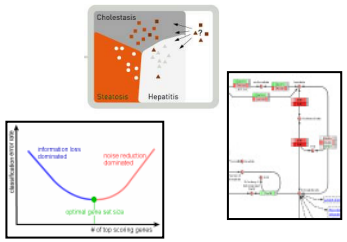
Animals

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



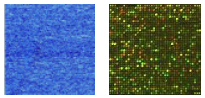
MOA, Tox Prediction

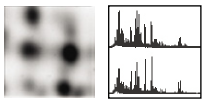
- MOA
- Biomarker candidates
- Tox Mechanism
- Tox Prediction
- …

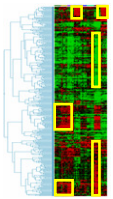


-omics data

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

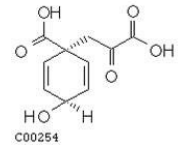




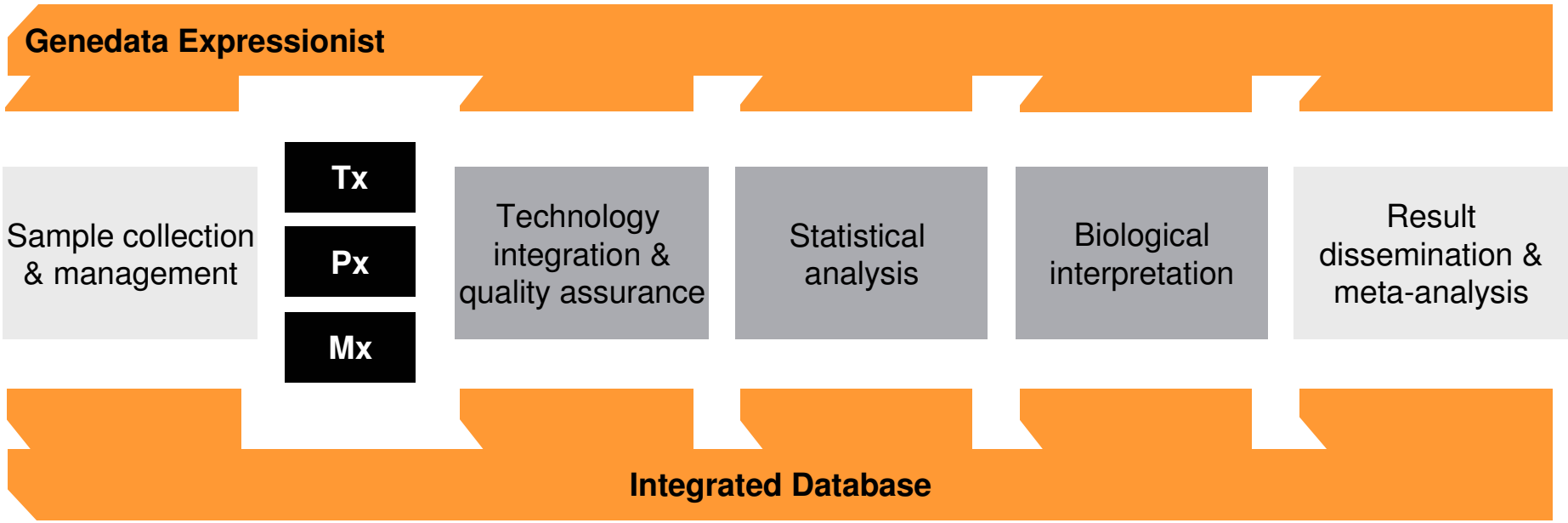


Treated animals

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



Genedata Expressionist®



Consortia

- **Innomed**
  - Prediction of human toxicity using T',P',M'omics technologies and conventional toxicology parameters
  - Thirteen Europe-based pharma companies, three academic institutions

- **BioCop**
  - Improvement in the ability to monitor for many classes of chemical contaminants present in cereals, meats, seafood and processed foods
  - Thirty three organizations (university, food industry, SMEs)

- **NewGeneris**
  - Development of biomarkers of dietary exposure to genotoxic chemicals and biomarkers of precarcinogenic and immunomodulatory effects in newborns
  - Twenty five organizations (university, SMEs)

InnoMed: PredTox consortium

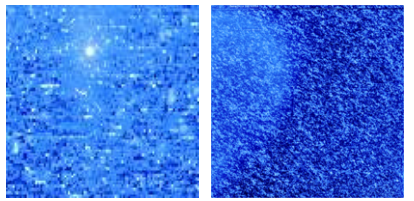
- Thirteen Europe-based pharma companies, three academic institutions, one bioinformatics company
  - Altana, Bayer, Boehringer Ingelheim, J & J, Merck KG, Novartis, Novo Nordisk, Organon, Roche, Schering AG, Sanofi-Aventis, Serono, Servier,
  - U Dublin, U Hacettepe, U Wuerzburg
  - Genedata
- Basic information
  - 3-year project, beginning in October 2005
  - Funded by participating companies and the EC
- Goals
  - Prediction of human toxicity using T',P',M' omics technologies and conventional toxicology parameters
  - 15 pharmaceuticals of which a portion should preferentially be from compounds that failed in clinical development but were not caught by preclinical toxicology
- Genedata involvement
  - Generation of a database for toxicogenomics data together with conventional toxicological endpoints
  - scientific data analysis; systems biology and pathway analysis; elucidation of biomarker candidates

Data quality control and data normalization

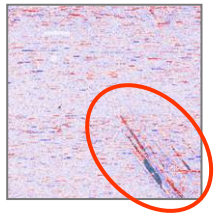
- Toxicogenomics is crucially dependent on high quality expression data

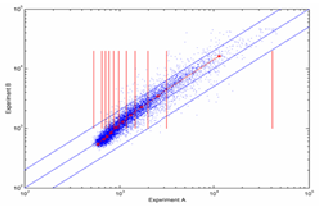
- Data quality control has to ensure:
  - Data quality assurance over large experimental series
  - High throughput analysis with standardized data processing
  - Diagnosis of technically conditioned effects
  - Enabling of consortial work and the submission of toxicogenomics data

Refiner Transcriptomics



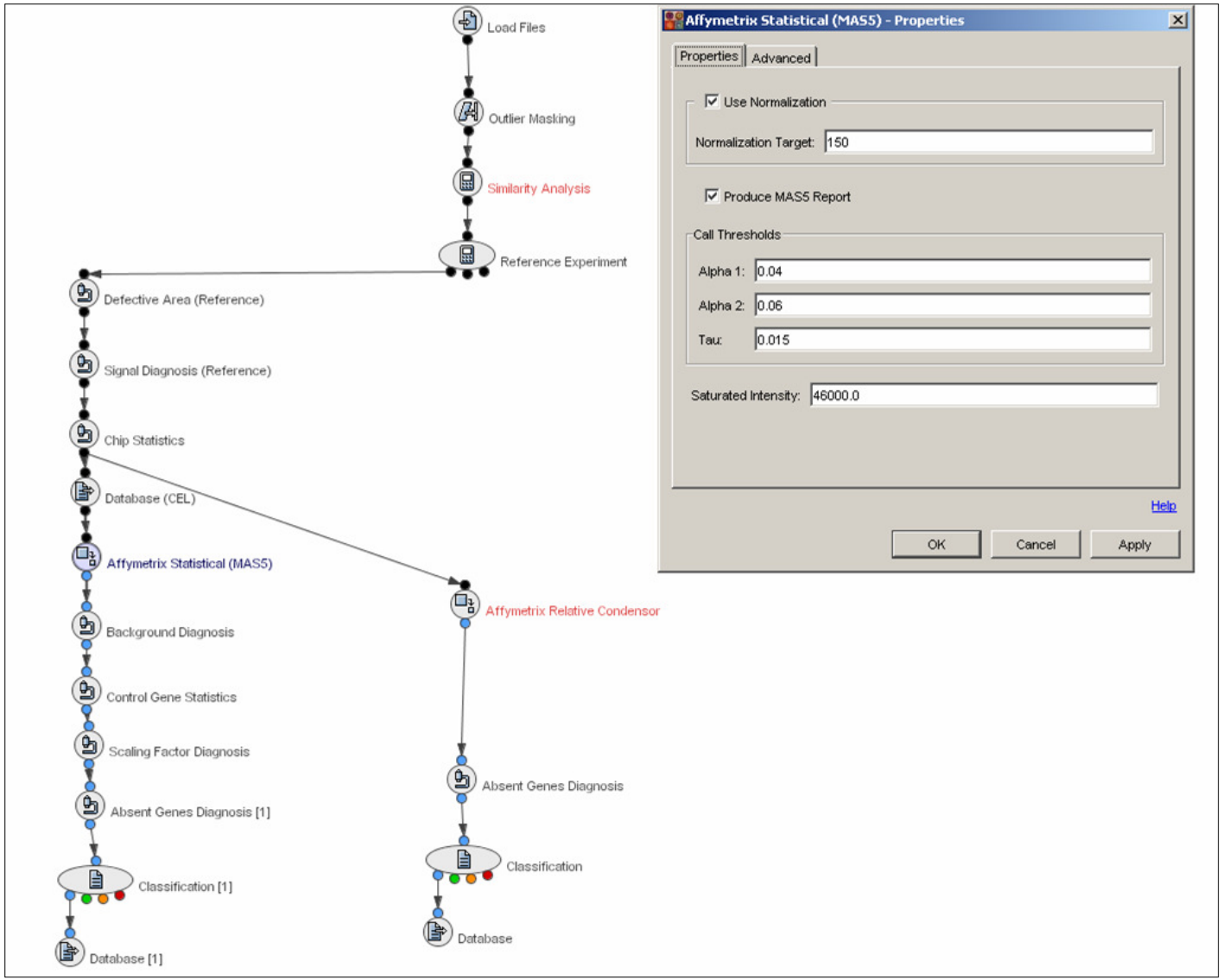
- Detection and correction of defects on microarrays
- Automated data quality control:
  - Loads uncondensed raw data
  - Detects and masks defective regions
  - Detects and corrects gradients and distortions
  - Condenses the data (MAS5, Li-Wong, RMA, GC-RMA)
  - Generates a quality classification for each chip
  - Saves condensed data into database





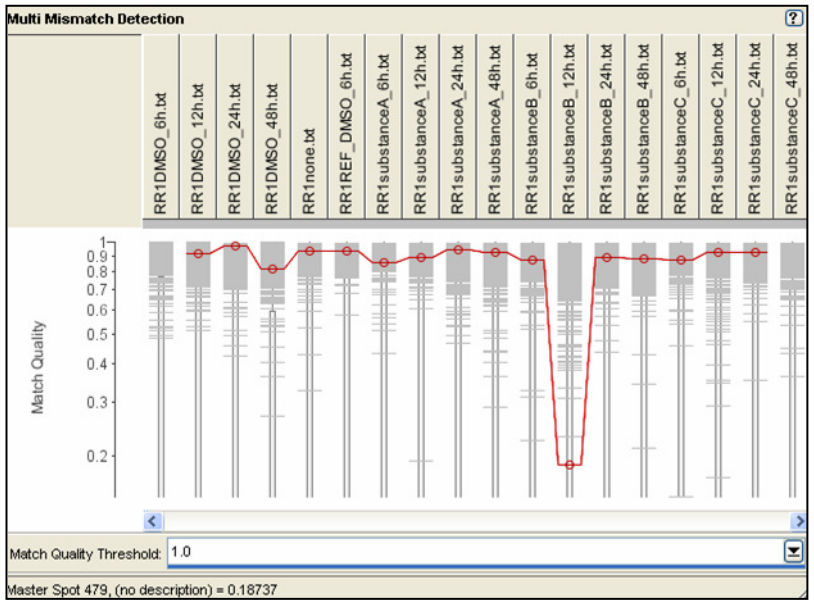
Classification	Gradient Severity	Distortion Severity	Masked Area (...
	0.00	0.02	0.08
	0.00	0.03	0.10
	0.00	0.03	0.11
	0.00	0.02	0.14
	0.00	0.02	0.16
	0.00	0.02	0.17
	0.00	0.03	0.21
	0.01	0.02	0.29
	0.00	0.02	0.30
	0.00	0.01	0.55
	0.00	0.01	0.62

Workflow for data quality assessment for one-channel data

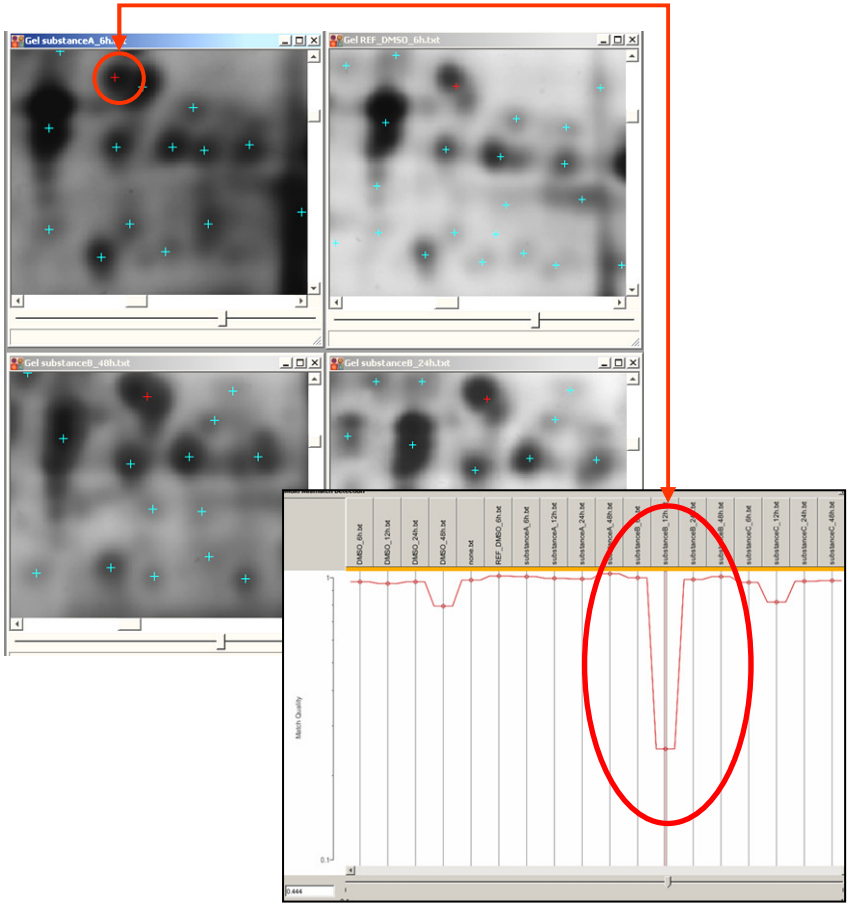


Refiner Proteomics

Compares location of spots over complete gel data set

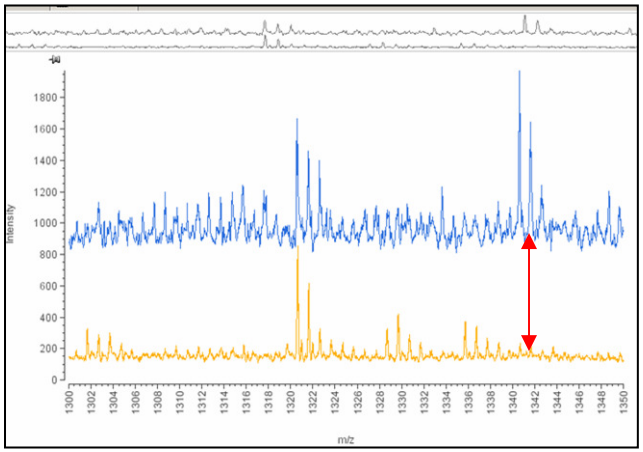


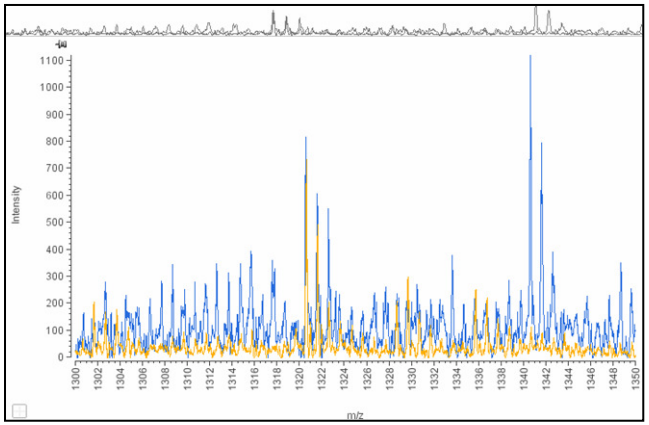
Automated mismatch detection based on calculation of standardized match scores



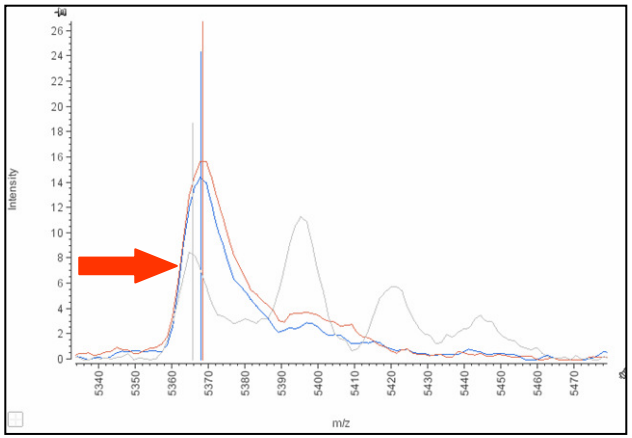
Refiner Metabolomics

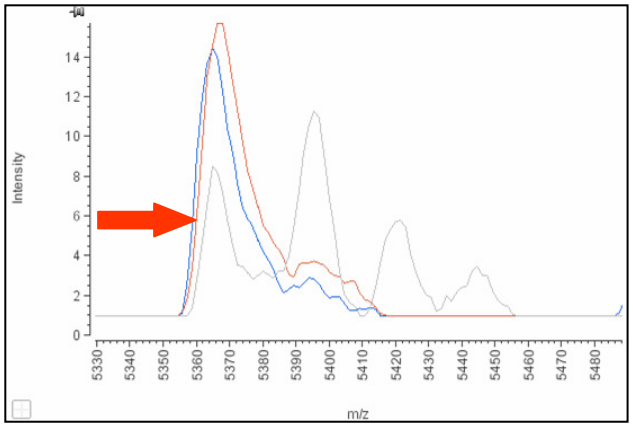
Baseline subtraction increases the comparability of spectra





m/z alignment prevents false positives in biomarker detection

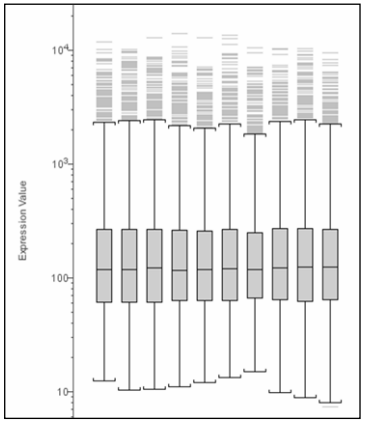


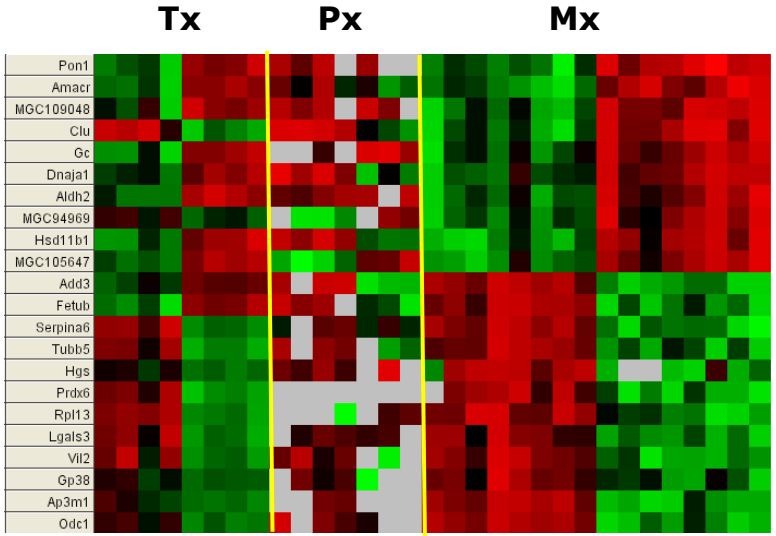


Mapping and normalization

- 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.

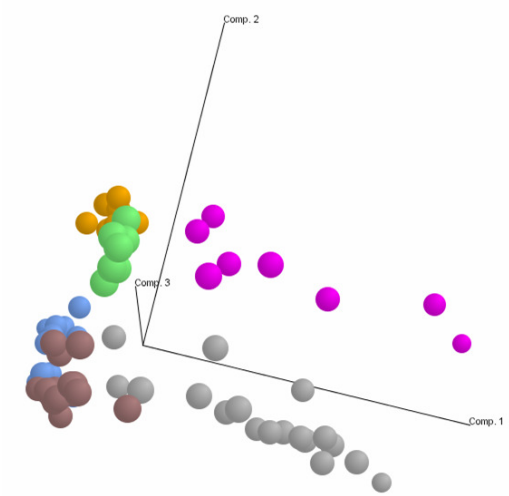


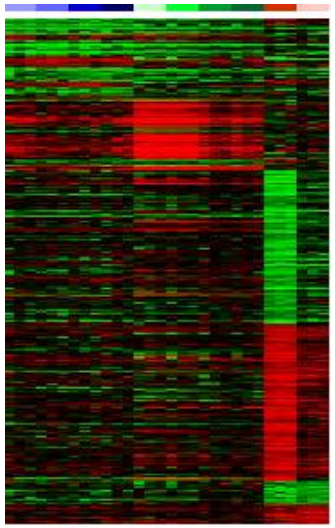


Mapping of data into a gene symbol space

Data overview

- After the basic data preprocessing steps, a first overview of the data can be achieved with unsupervised learning algorithms, e.g.
  - Principal Components Analysis
  - Hierarchical clustering
  - Self-organizing maps
  - K-means clustering





- These methods can be used to arrange transcripts/proteins/metabolites in groups or clusters based solely on the similarity of their expression
- The results can be compared with possible information on phenotypes or already be used for predictions

Biomarker discovery

- Depending on the experimental design, different methods to discover differentially expressed genes can be applied:

- Two experiment groups with very few (1 to ~3) experiments:
  - **N-fold regulation**: most basic method, no statistical test

- Two experiment groups with 'sufficient' number of experiments per group:
  - **T-test** and variants of it (Welch, Kolmogorov-Smirnov, Wilcoxon, Paired T-test)

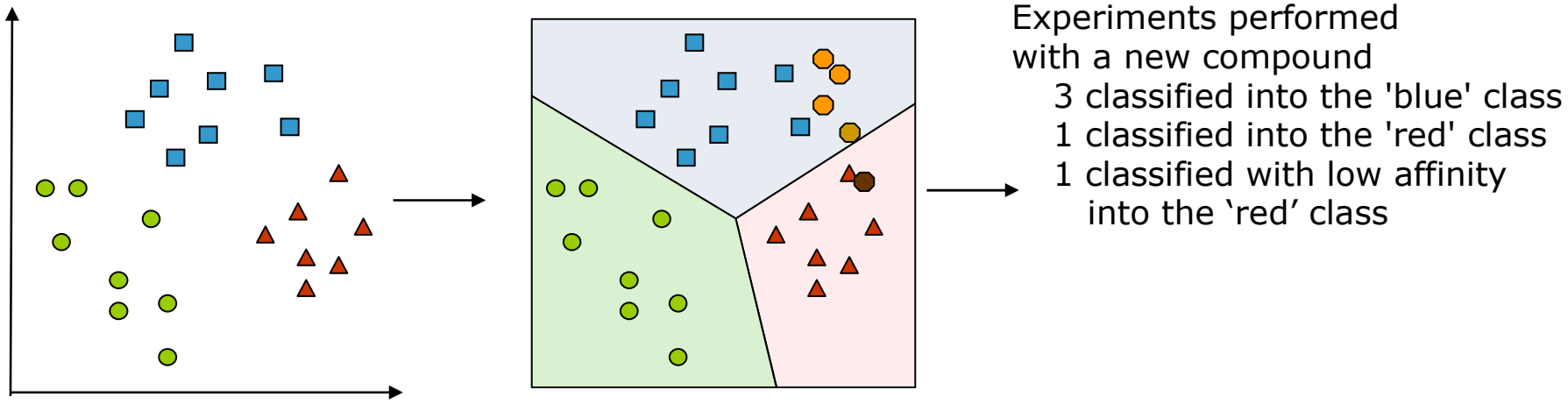
- More than two experiment groups and one factor
  - **ANOVA, Kruskal-Wallis, Contrasts**

- More than two experiment groups, more than one factor
  - **N-way ANOVA**

- Estimates of **False Discovery Rate** and **Sensitivity** help to judge reliability of list of biomarkers

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 in gene space annotated with clinical endpoints

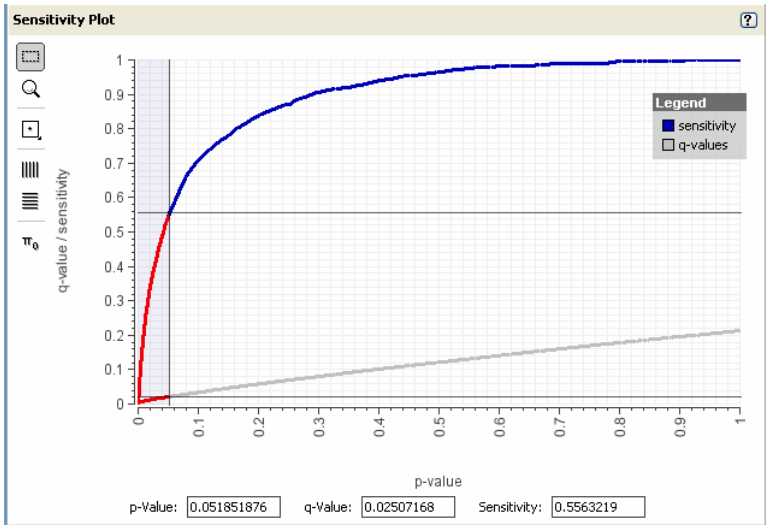
Reference compendium

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

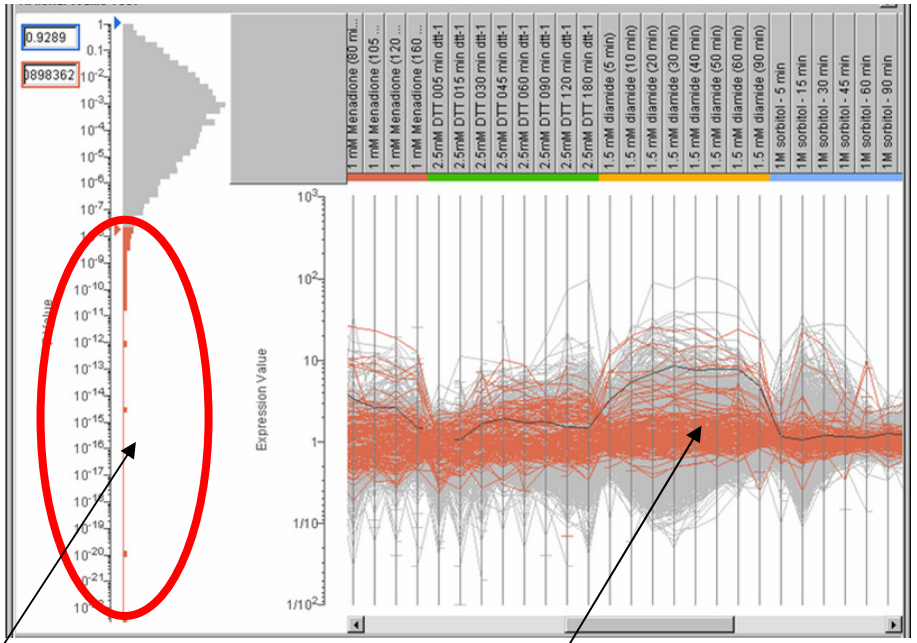
Building of a reference compendium by unsupervised learning algorithms

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

ROC curve



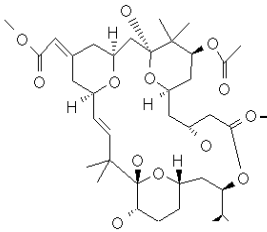
ANOVA



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

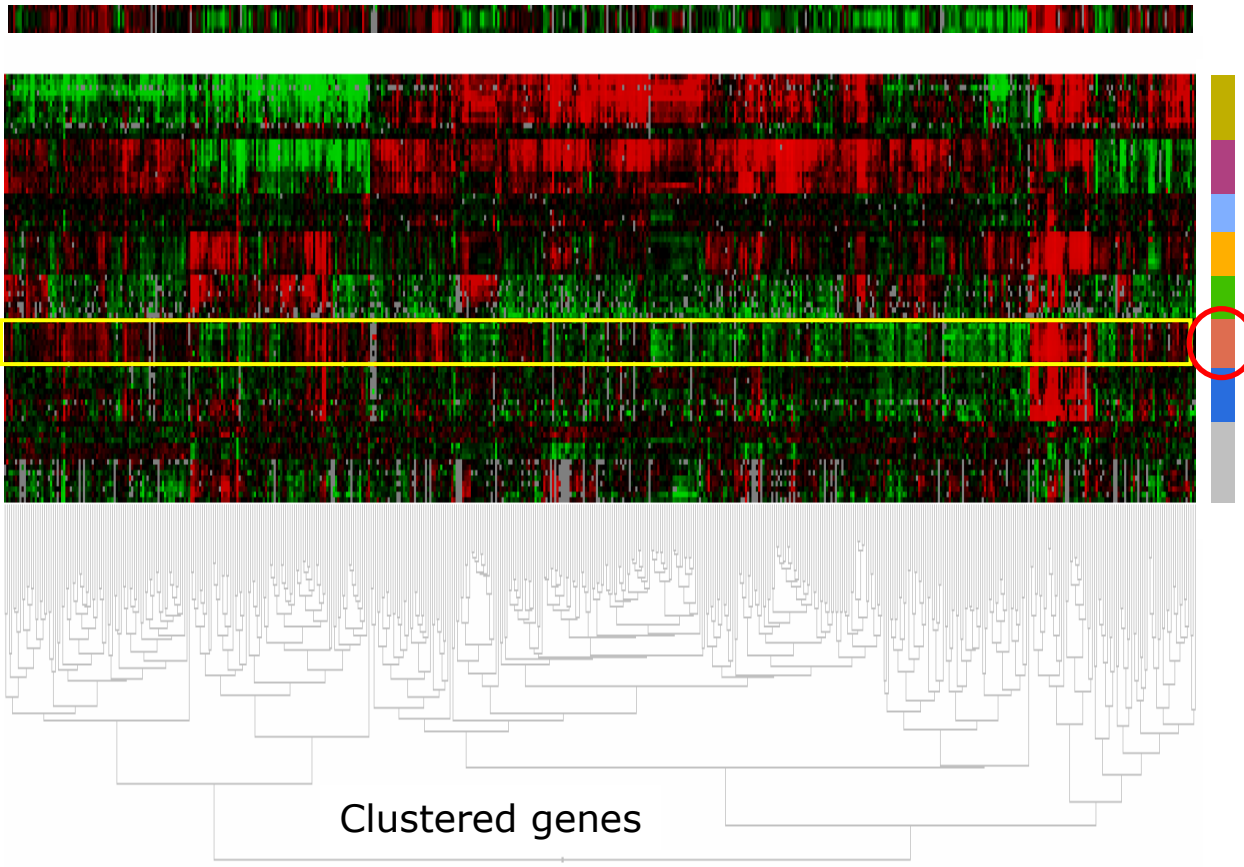
mRNA changes characteristic for a certain toxic mechanism

Prediction of toxicity of a new compound using unsupervised learning methods



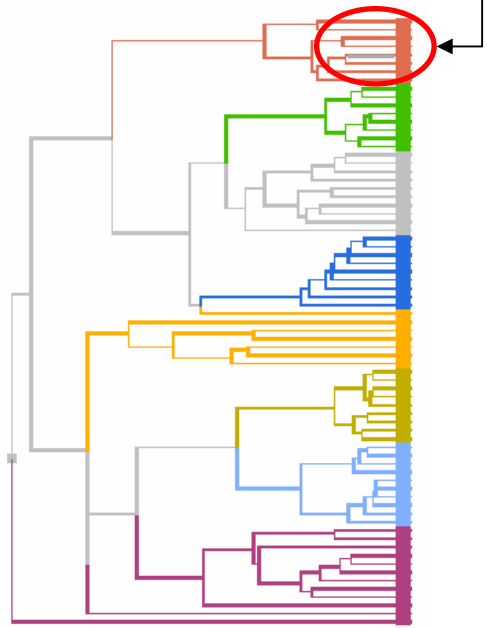
Drug candidate

Expression profile of drug candidate



Clustered genes

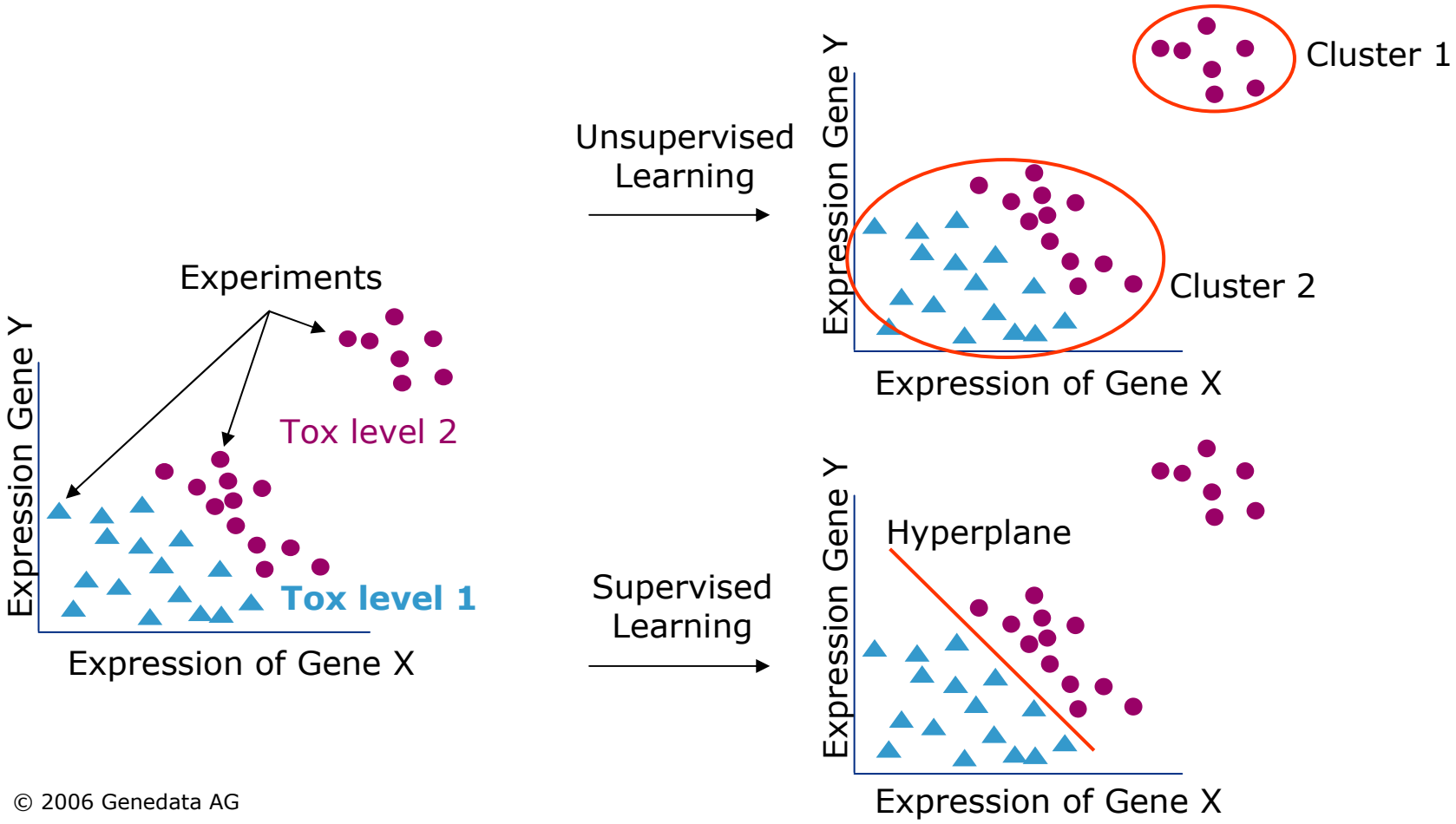
1 row = 1 experiment
1 colour = 1 toxin



Clustering of experiments

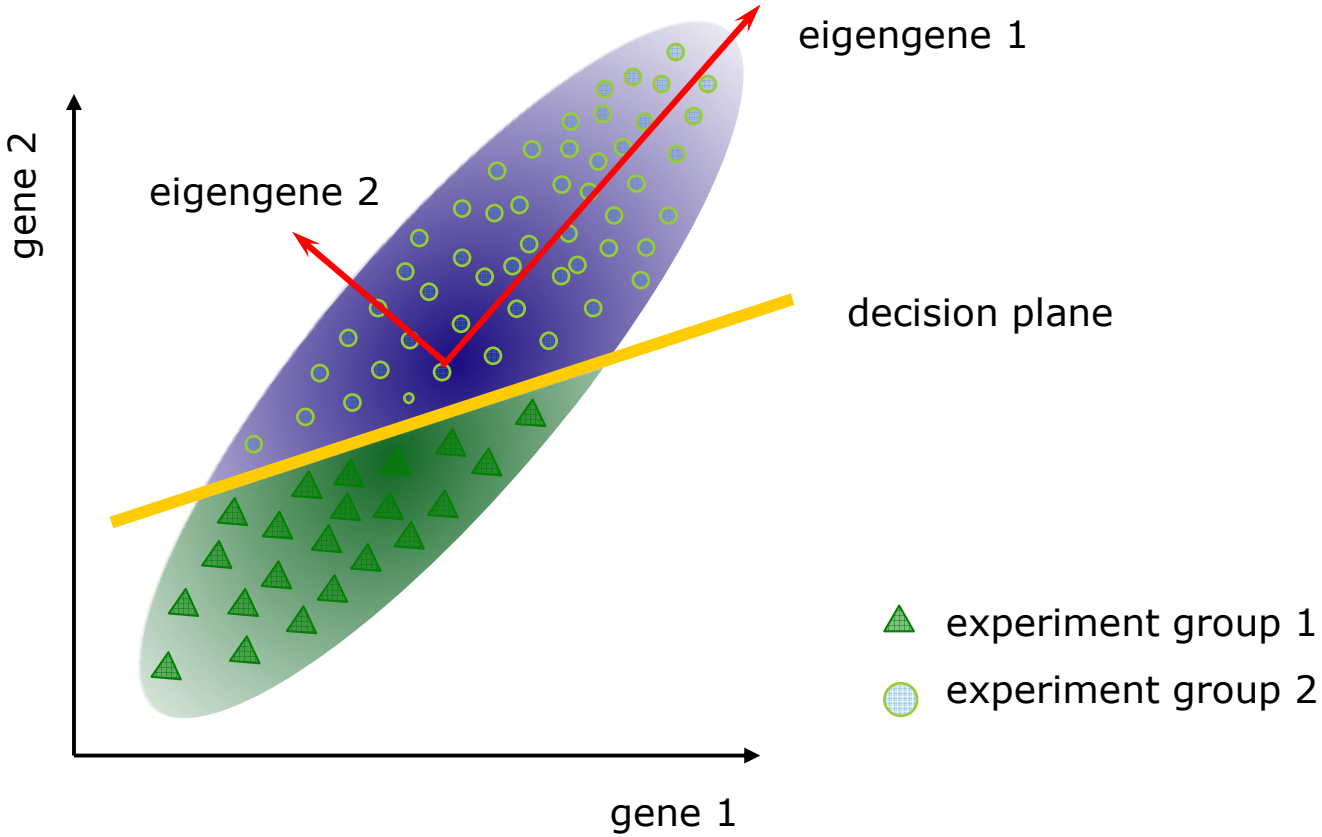
Reference compendium for toxicity prediction by supervised learning methods

- 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



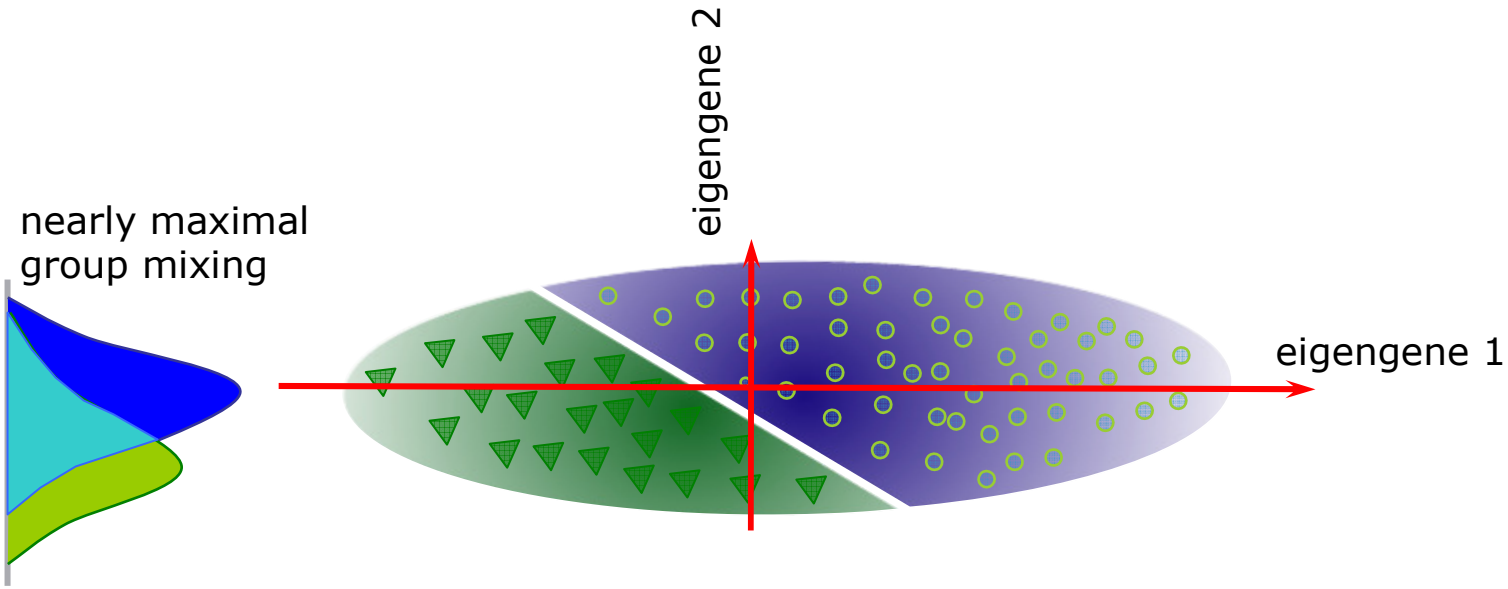
PCA vs. supervised learning

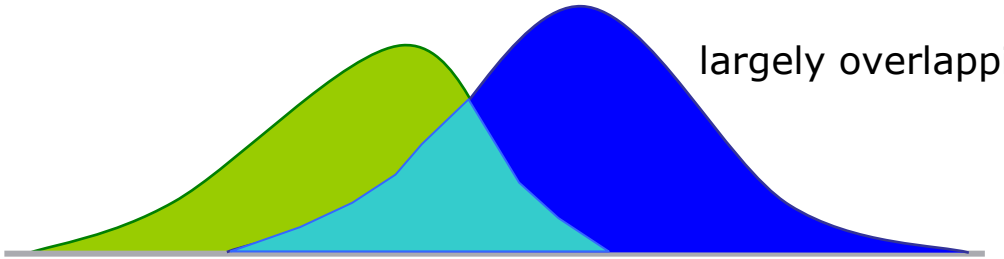
- PCA is in general also not the best method to classify experiment groups as can be seen from the following example
- The group separation line (decision plane) is in general not parallel to any of the genes or eigen-directions



Classification using PCA

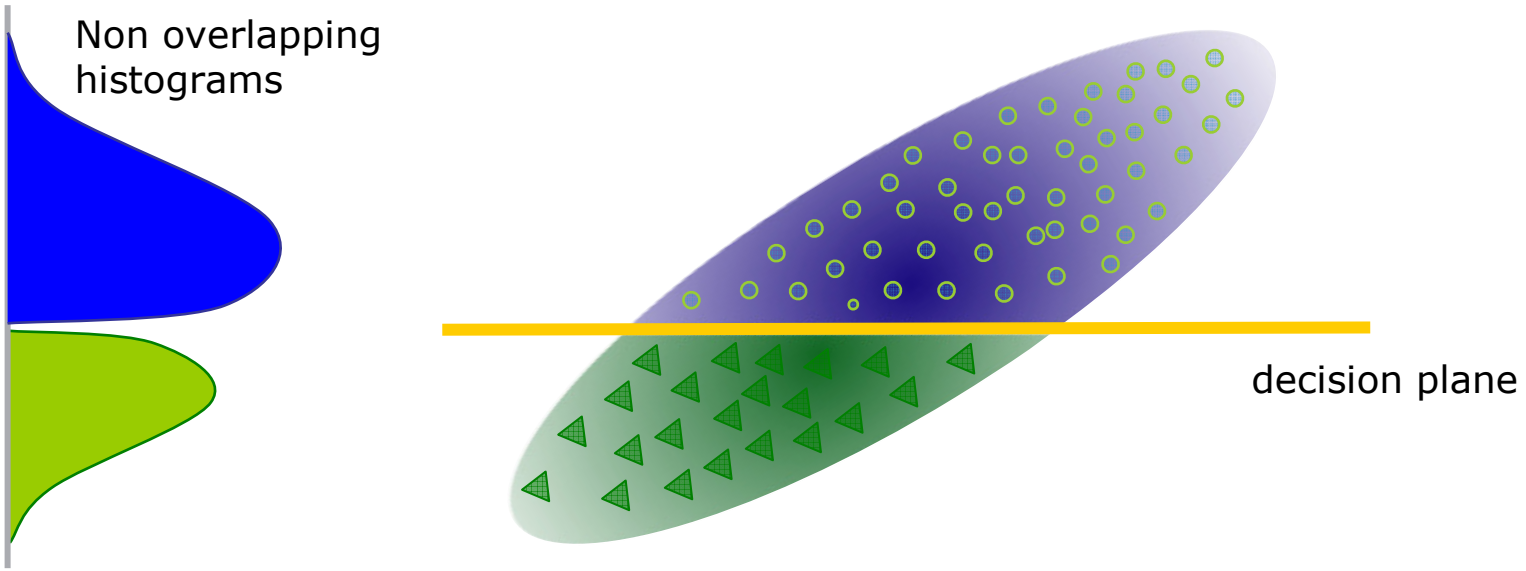
- A rotation to the eigenspace therefore does not solve the classification problem





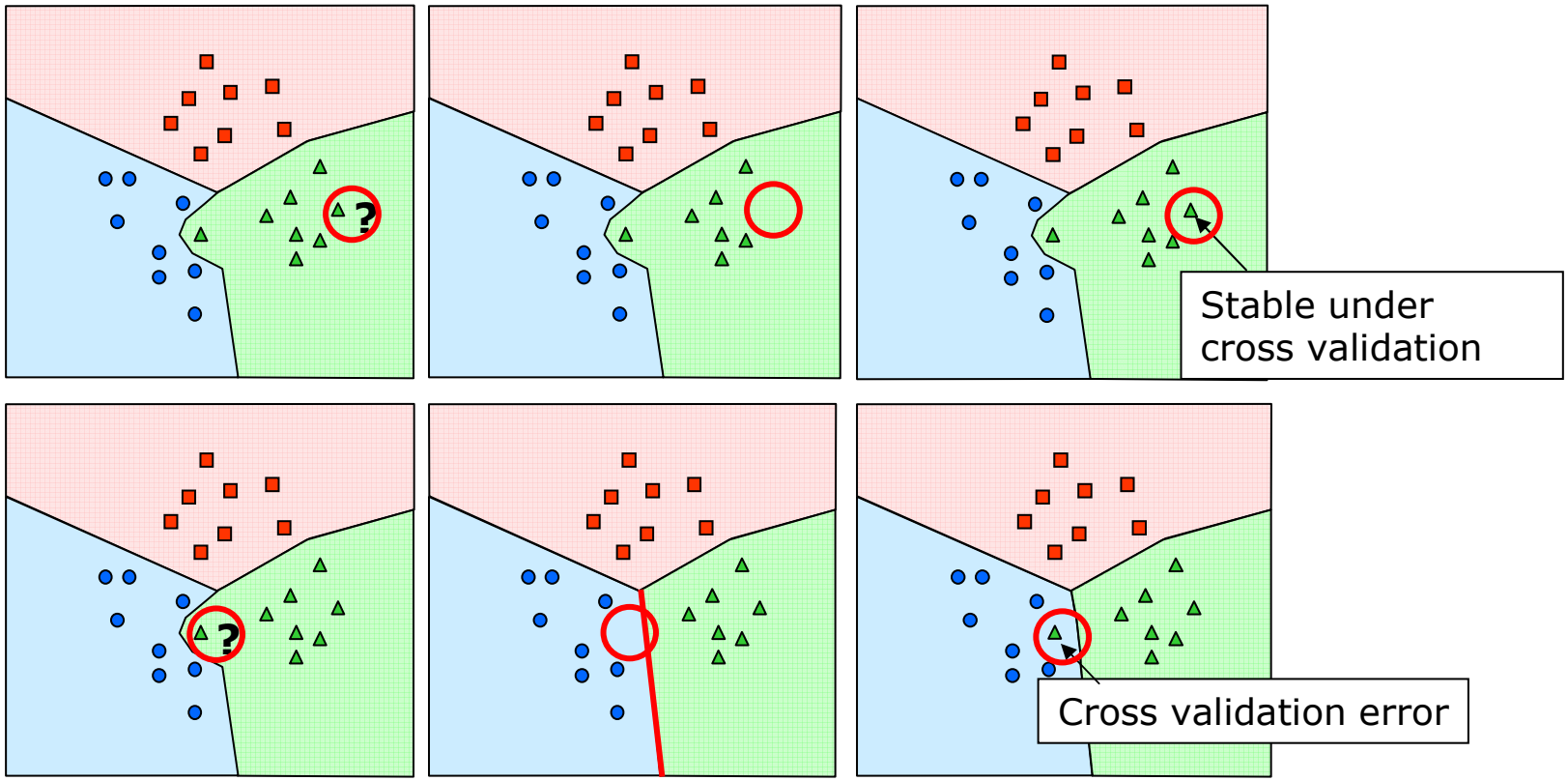
PCA vs. supervised learning

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



Cross validation of 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



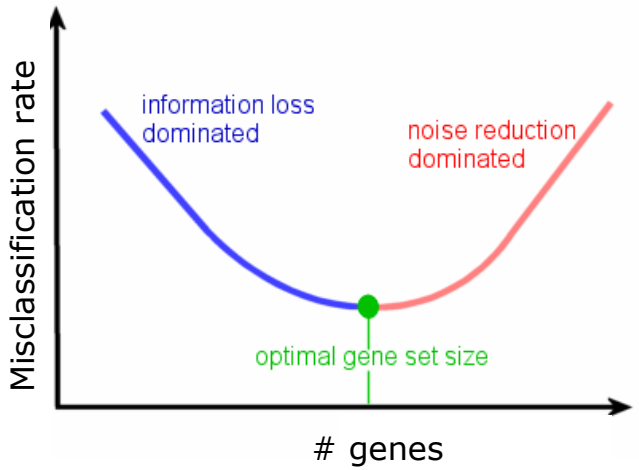
Determination of the optimal set of genes

- 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

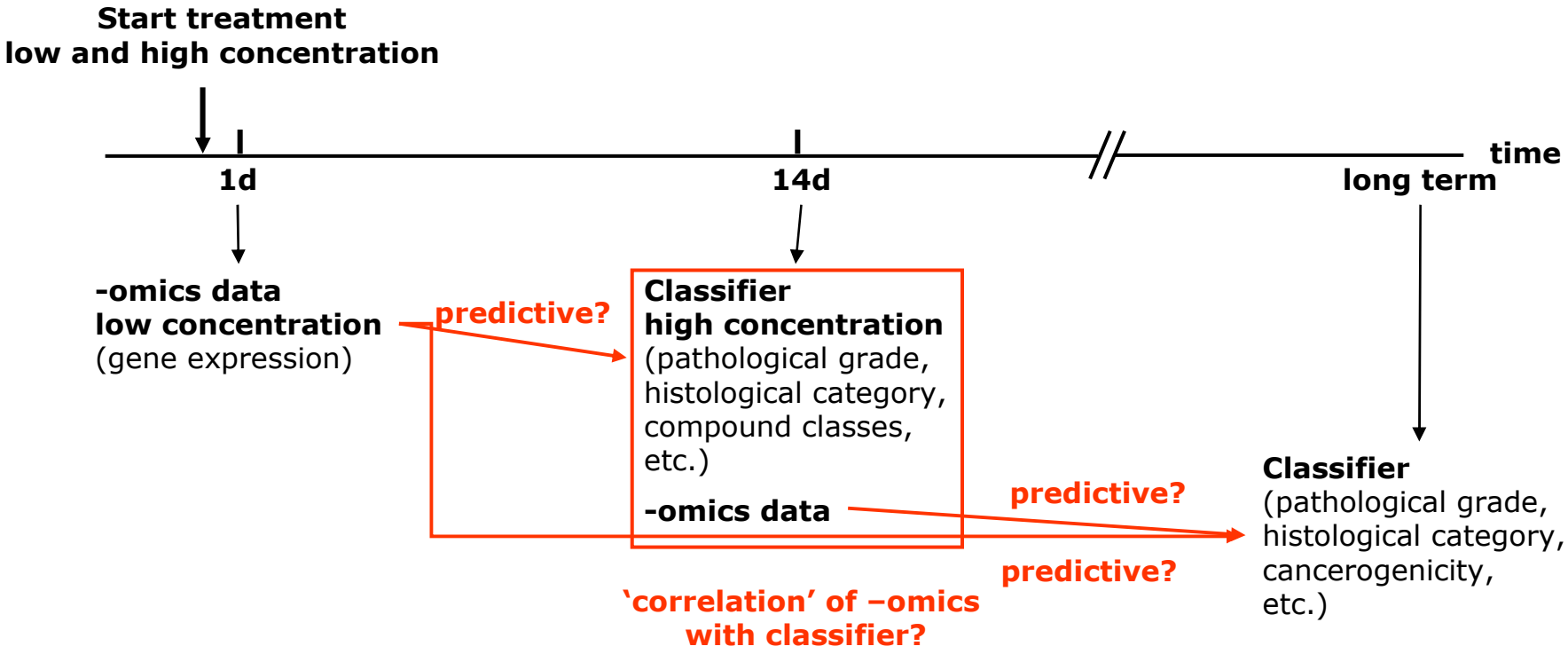
- Supervised learning
  - Support Vector Machine
  - Sparse Linear Discriminant Analysis
  - Fisher Linear Discriminant Analysis
  - K-Nearest Neighbours

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



Hypothesis-driven correlation

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

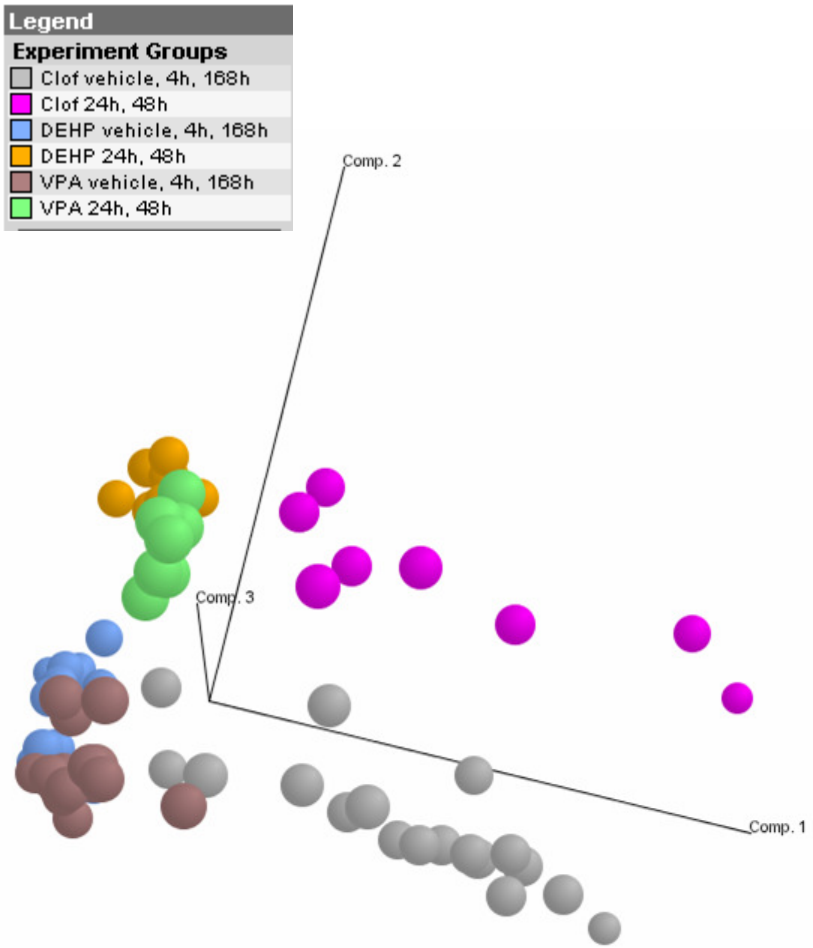


- What shall be predicted? Careful selection of hypothesis-driven classifier!

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

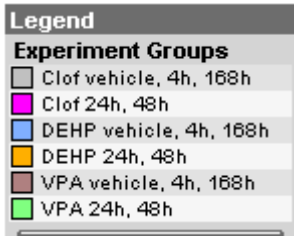


Vehicle, 4h, 168h experiments
= same colour for each compound

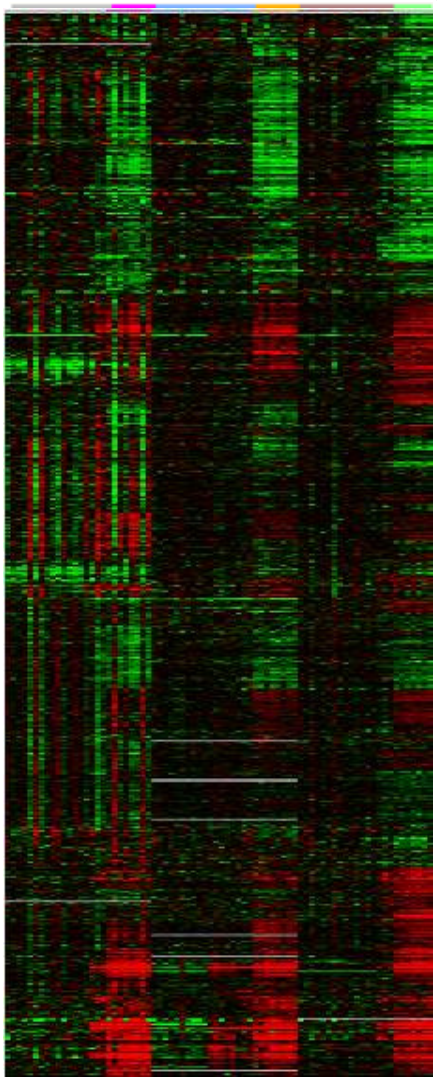
24h and 48h experiments
= same colour for each compound

- PCA with 3241 transcripts
- 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

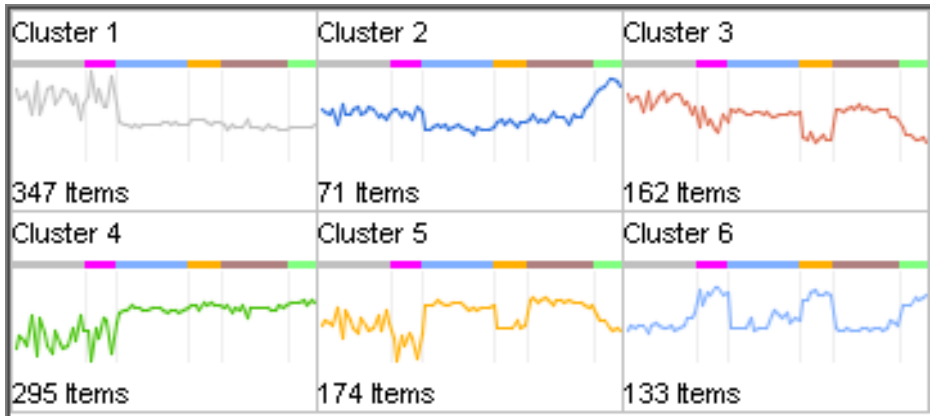


Hierarchical clustering



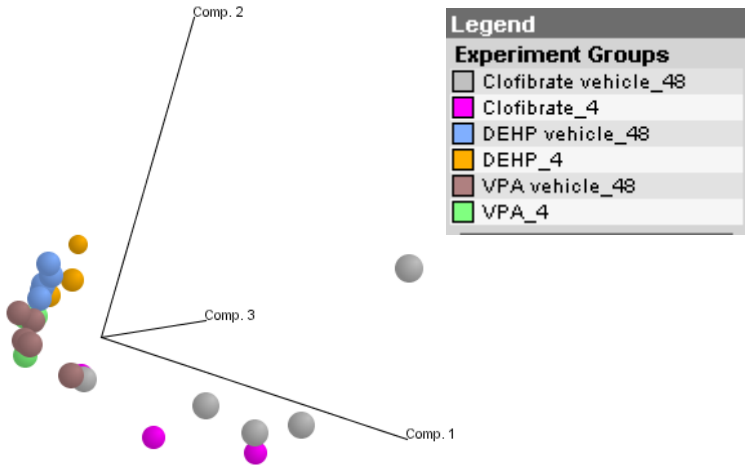
- Hierarchical clustering and K-means with 1205 transcripts after ANOVA (p-values < 0.001, fold change ≥ 2 or ≤ 0.5)
- Most up- or down-regulated transcripts after 24h and 48h
- Clofibrate experiments separated from DEHP and VPA experiments

K-means clustering



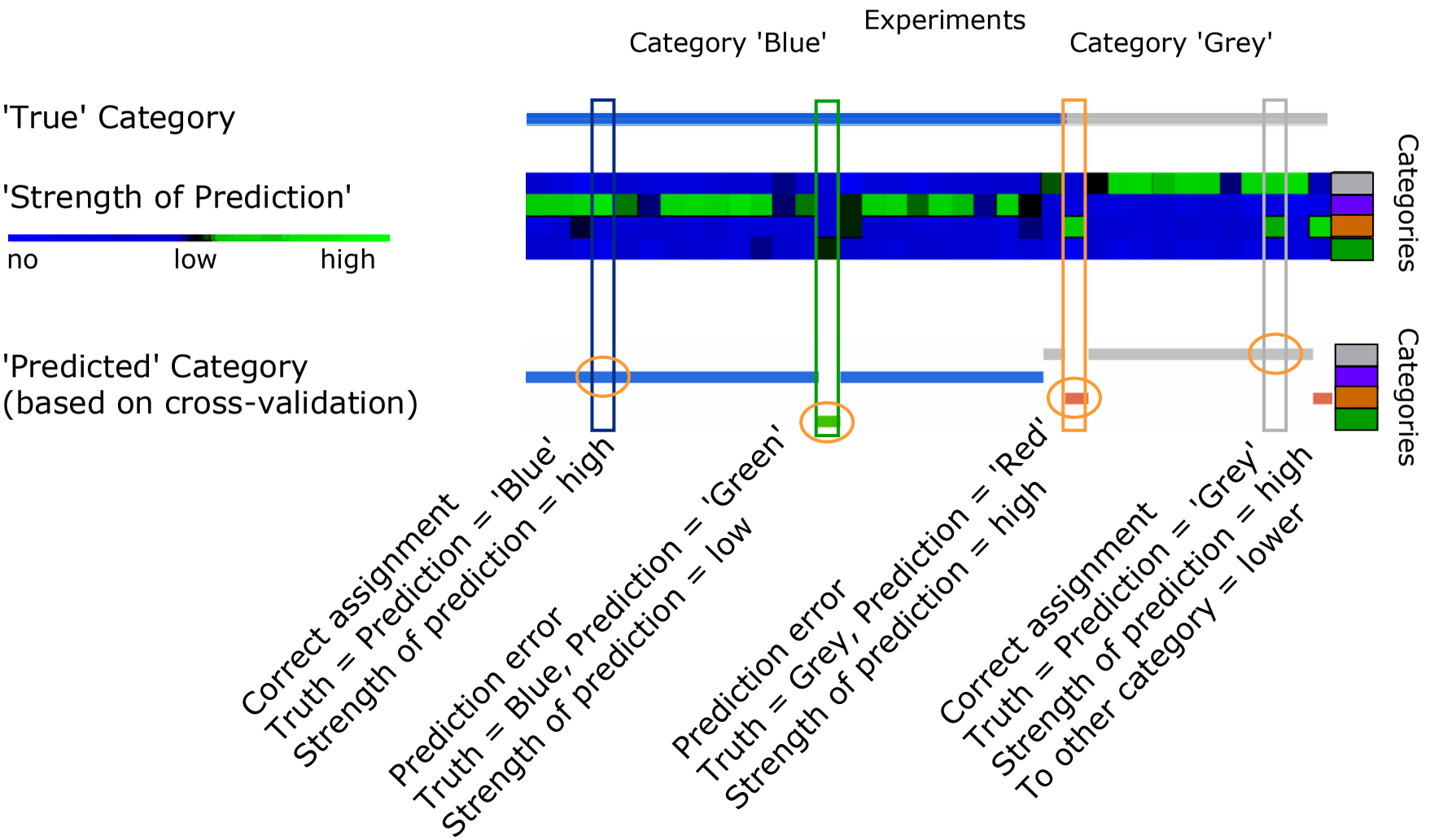
Which data should be selected to build a reference compendium?

- The 24 and the 48 hours data had the largest transcription response (no. of changes), and the most robust change in liver phenotype (liver weight increase as well as minimal changes in morphology)
- Based on gene expression the 4 hour experiments are not or barely distinguishable from the 48 hour control (vehicle) experiments
- However, the treatment with the compounds lead to the observed phenotypic changes in liver

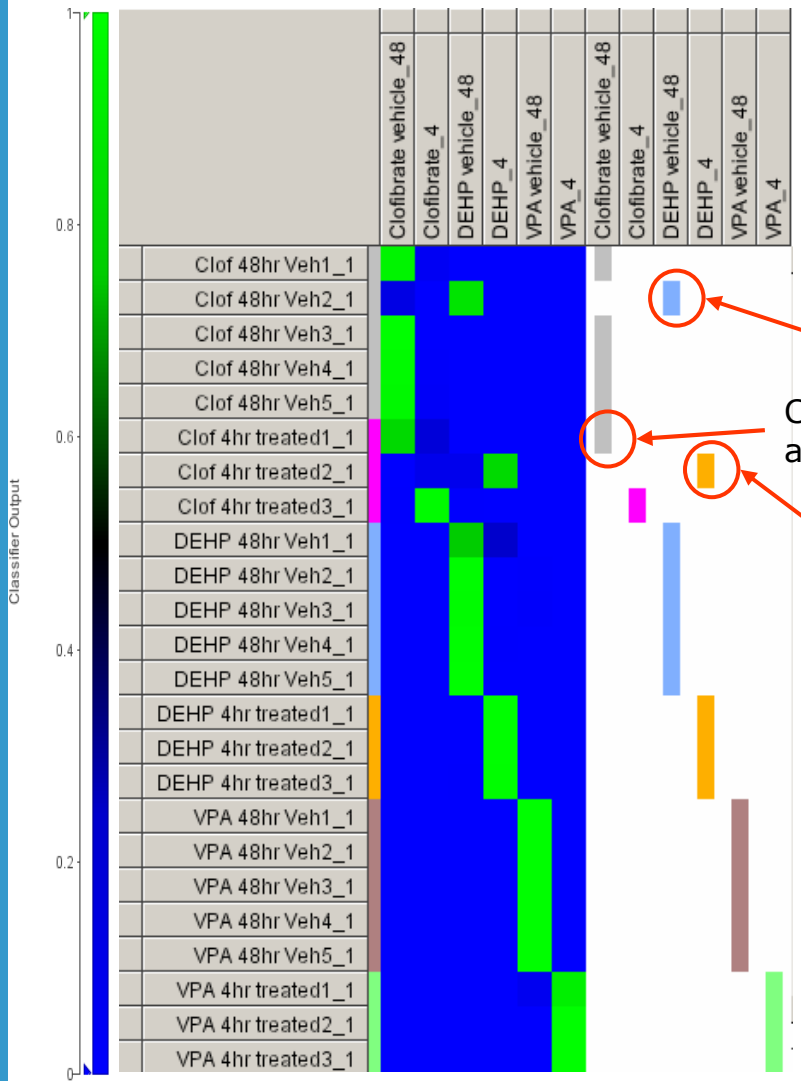


- **Questions:**
  - Is there enough information in the expression profiles of the 4h and 48h vehicle experiments to use these experiments for a reference compendium?
  - Can the gene expression profiles of these 4 hour and 48 hour control experiments be used to predict outcomes at later time points?

Explanation of Cross Validation results



Cross-validation of vehicle and 4h experiments



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

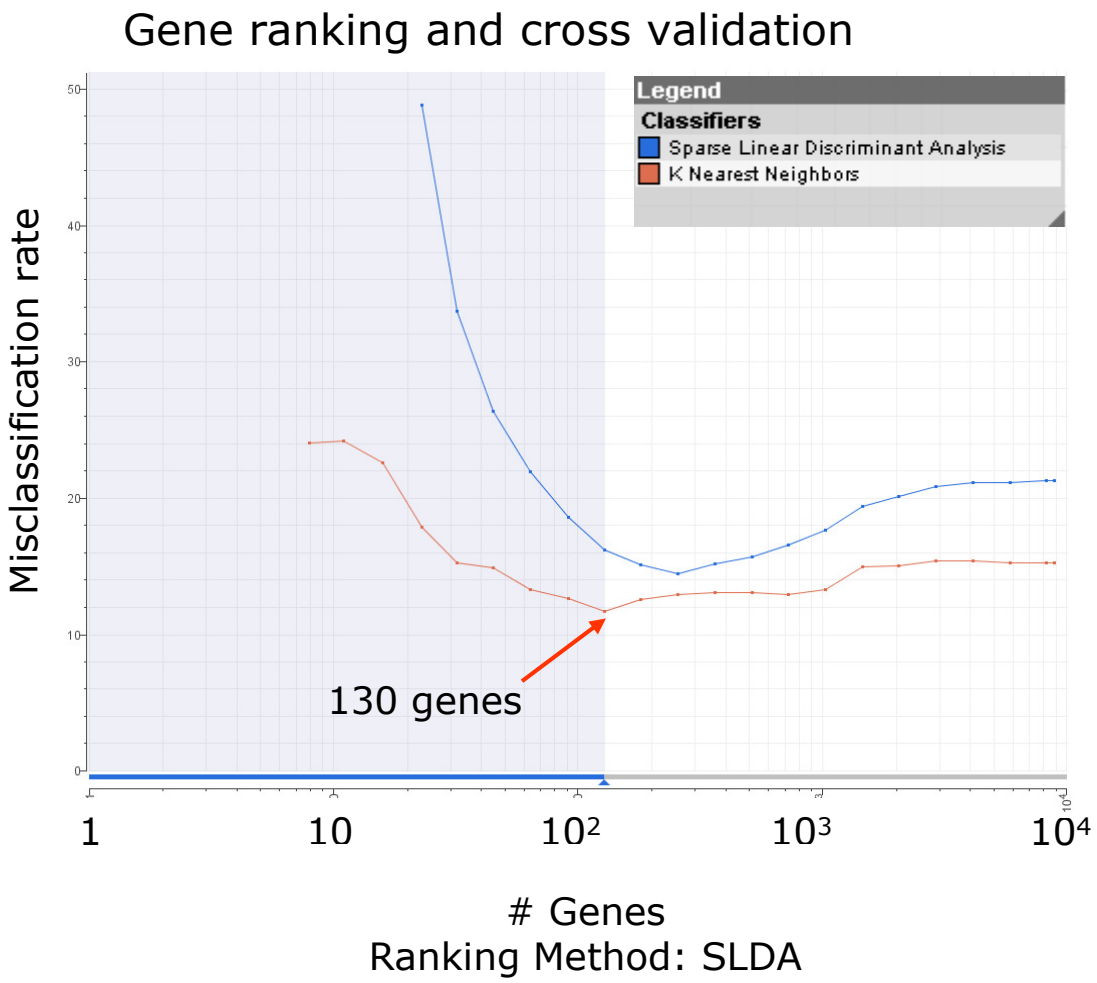
Clof 48h vehicle classified as DEHP vehicle 48h

Clof 4h classified as Clof vehicle 48h

Clof 4h classified as DHEP 4h

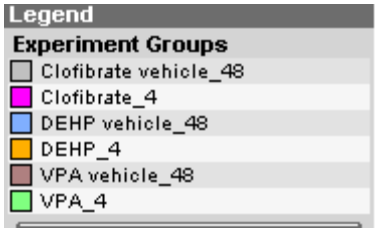
- Acceptable misclassification error in reference compendium consisting of vehicle and 4h experiments
- Reference compendium can be used for predicting the output variable of a compound based on the expression profile

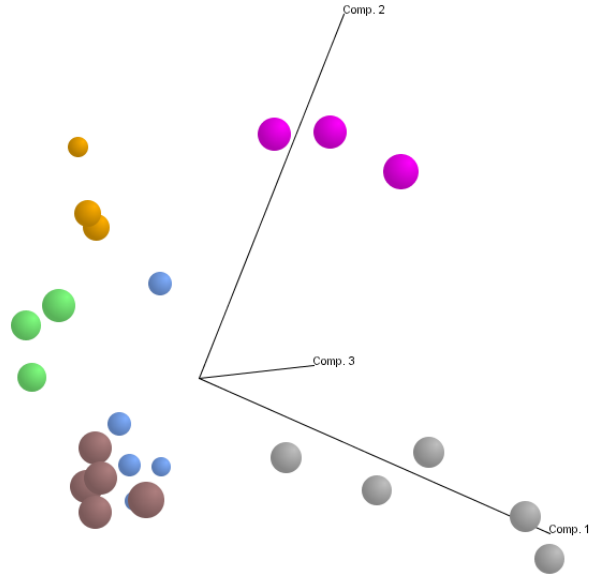
Gene ranking to define optimal marker gene set



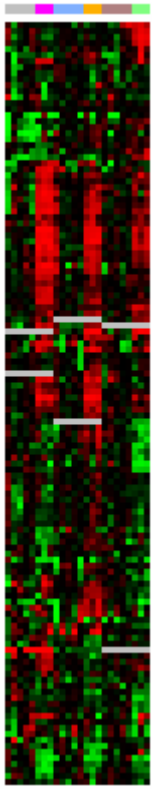
- Classifiers: KNN and SLDA
  all genes (~ 8800)
  Test Set Fraction=25
  Number of Repeats=500
  Ranking: Sparse Linear Ranking
- Optimal marker gene set consists of ~ 130 genes
- Prediction error ~ 12 %
- No significant reduction of prediction error but reduction in number of significant transcripts

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

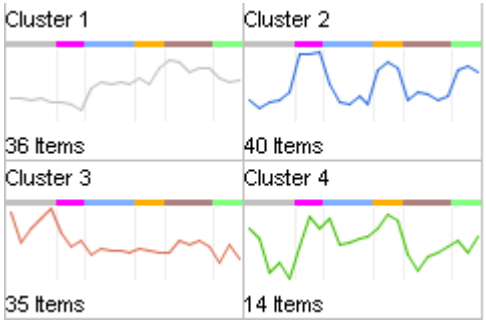




Hierarchical clustering



K-means clustering

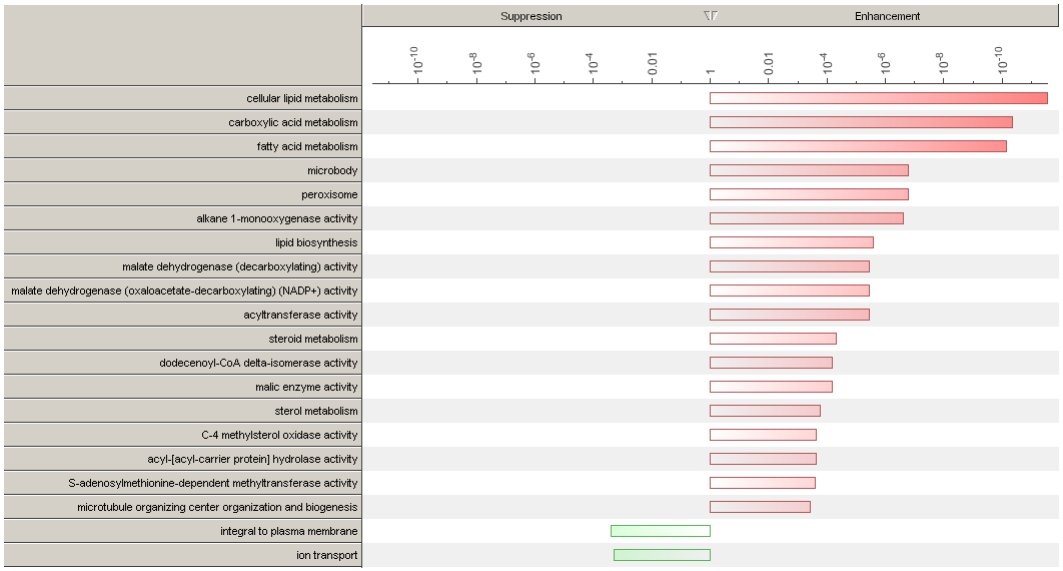


- PCA, hierarchical clustering and K-means with optimal marker gene set consisting of 130 transcripts

- There is already enough information in about 130 transcripts from the 4h experiments to predict the clinical outcome of the 24 and the 48 hours experiments (liver weight increase as well as minimal changes in morphology)

Biological interpretation

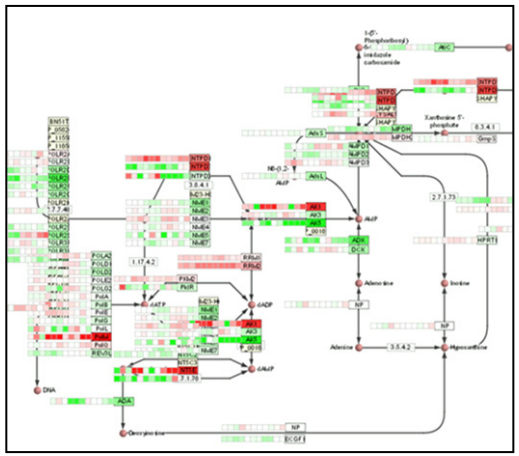
- Fisher’s exact test using GO annotation
  - Fisher's Exact Test is a counting test that assigns statistical significance to statements about the over- or under-representation of properties in a selection group when compared to a so-called universe group.
  - Example: comparison of optimal gene set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

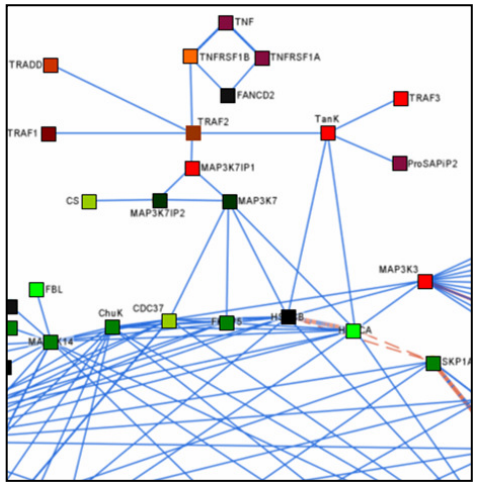


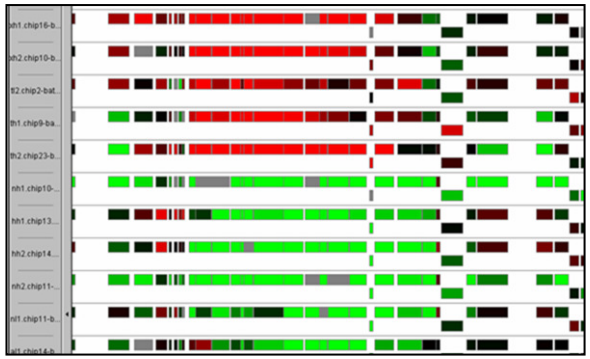
- Additional tools: Chromosomal Location; Pathway Mapping

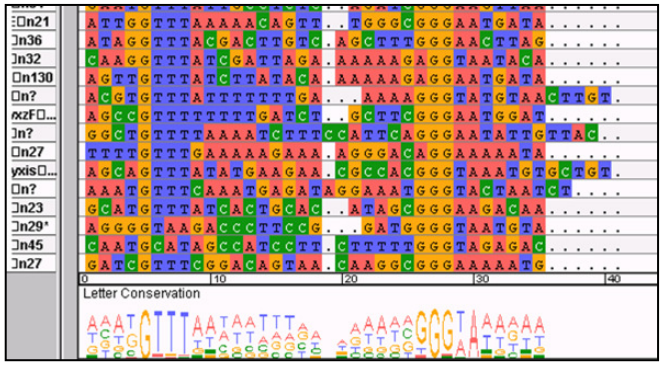
Pathway characterization and biomarker characterization

- 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.







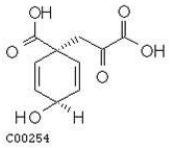


Predictive Tox Database

Genedata
solutions in-silico

Compounds

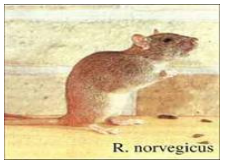
- Compound
- Compound class
- Concentration
- Treatment time
- Dosing route
- Dosing frequency
- Vehicle
- Endpoints
- …



MOA, MOT
Biomarkers
Predictability

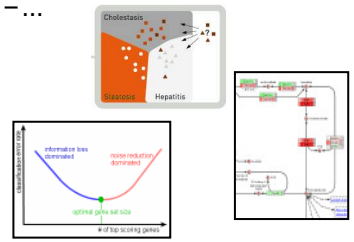
Animals

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



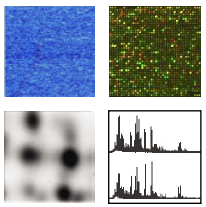
MOA, Tox Prediction

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



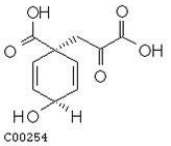
-omics data

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



Treated animals

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



Thank you

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

hans.gmuender@genedata.com