Toxicogenomics and biomarker discovery for the prediction of long term toxicity

Dr. Hans Gmuender, Scientific Consulting
Dr. Andreas Hohn, Business Development
September 2005, Eurotox 2005, Cracow, Poland

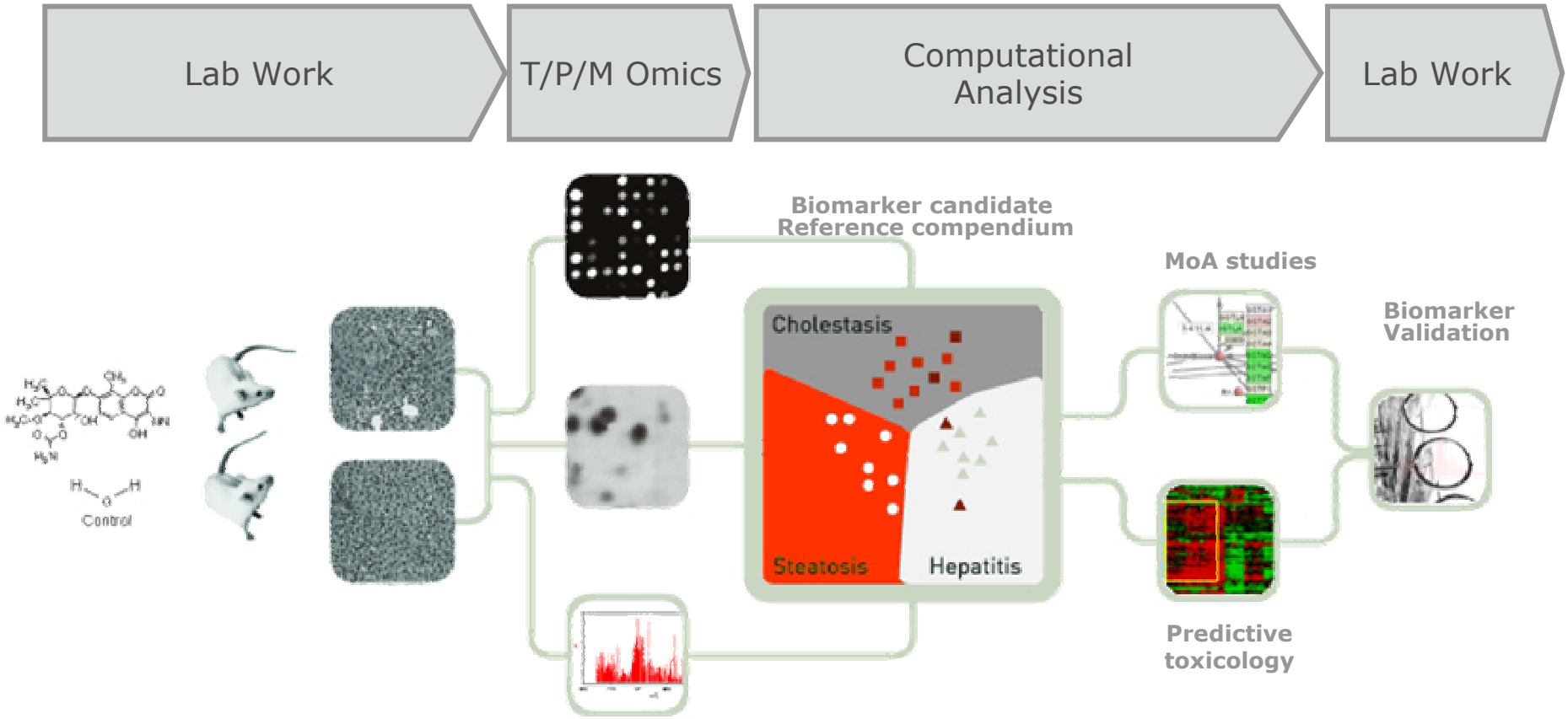
Challenges for toxicogenomics

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

- Questions to be answered by toxicogenomics:
  - Does toxicogenomics improve the prediction of long-term toxicity?
  - Does toxicogenomics lead to a better understanding of toxic effects?
  - Can cross-species biomarkers be identified and validated?

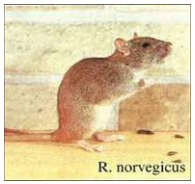
Research process



Predictive Tox Database

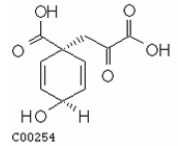
Animal Catalogue

- Species
- Strain
- Sex
- Age
- Tissue
- Histopathology



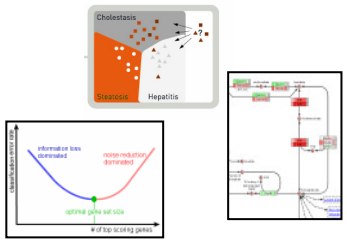
Treatment Catalogue

- Compound
- Concentration
- Treatment time
- Dosing route
- Dosing frequency
- Protocols
- Endpoints
- Histo images
- Histopathology scores
- Clinical chemistry
- Hematology



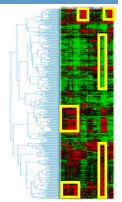
ToxResults Store

- MOA classification
- Biomarker candidates
- Tox Mechanism



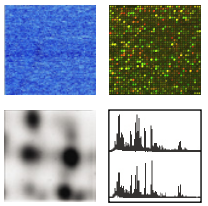
Tox Predictor

- Experimental values
- p-Values
- Gene/protein/ metabolite annotation
- Experiment Annotation

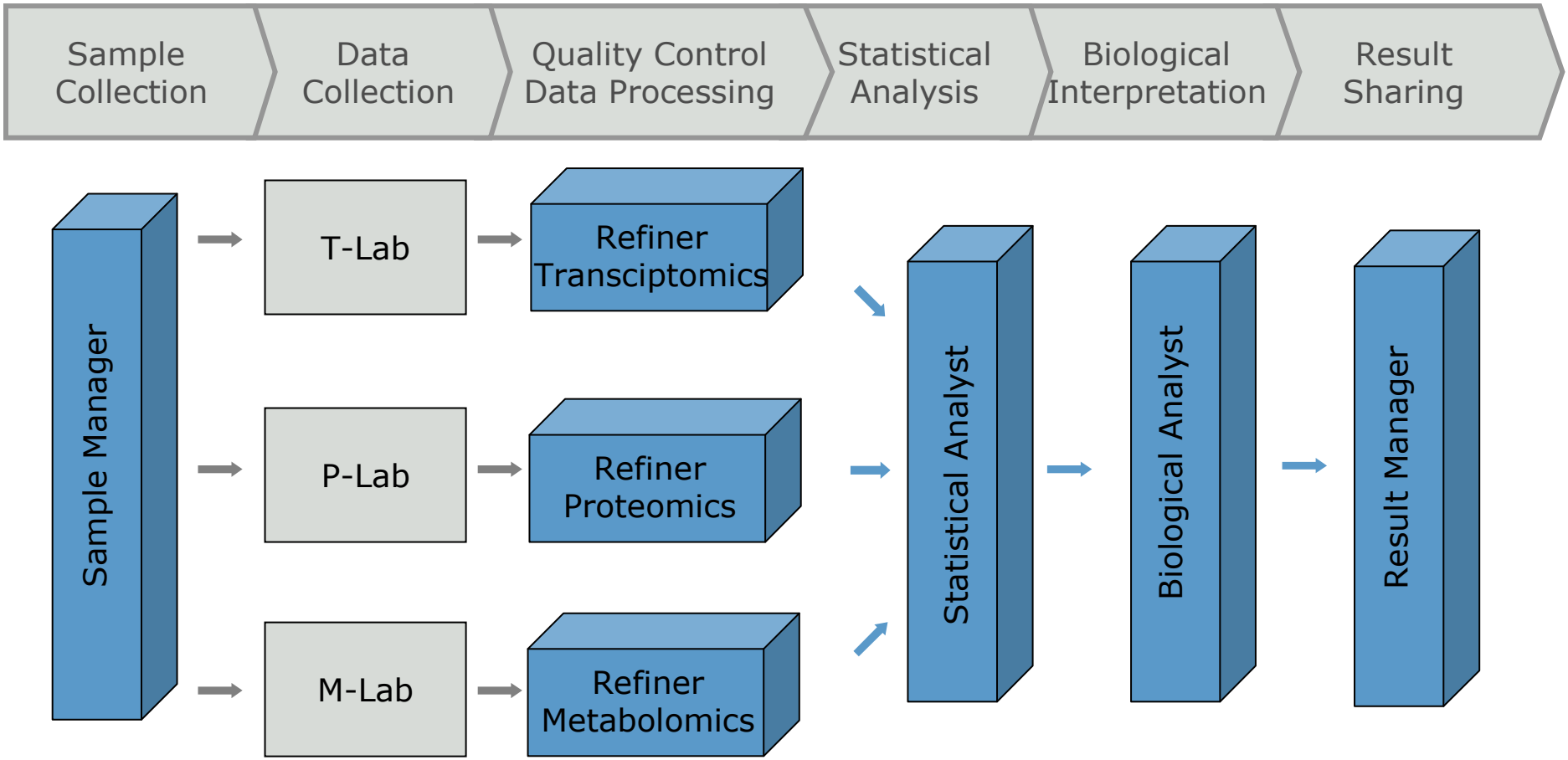


Expression Store

- Affymetrix
- 2 channel
- 2D gel
- MS
- NMR
- Recording devices
- Genes, transcripts, proteins
- Raw data



Process supported by Genedata Expressionist®



Activities in toxicogenomics

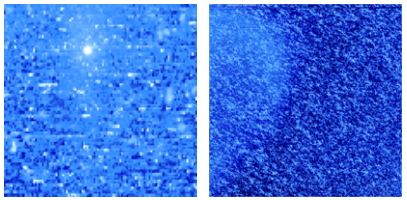
- Collaborative projects with pharmaceuticals
- Bioinformatics Partner of several EU funded Tox-related Consortia, including
  - BioCop: New technologies to screen multiple chemical contaminants in foods
  - NewGeneris: Newborns and genotoxic exposure risks
  - InnoMed: Predictive toxicology using systems biology approach
- Construction of a toxicogenomics database for -omics technologies together with conventional toxicology endpoints

Data quality control and data normalization

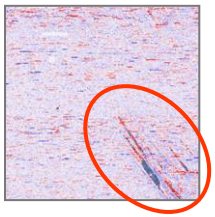
- Toxicogenomics is crucially dependent on high quality expression data

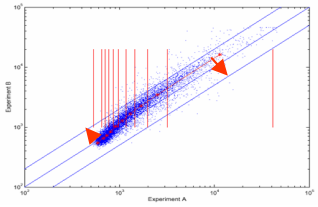
- Data quality control has to ensure:
  - Data quality assurance over large experimental series
  - High throughput analysis with standardized data processing
  - Process automation
  - Correction of site effects
  - Enable consortial work and the submission of toxicogenomics data

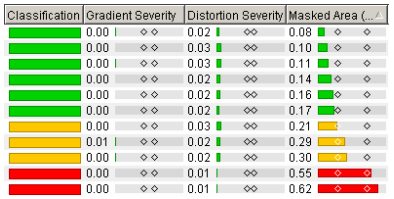
Refiner Transcriptomics



- Detection and correction of defects on microarrays
- Automated data quality control:
  - Loads uncondensed 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

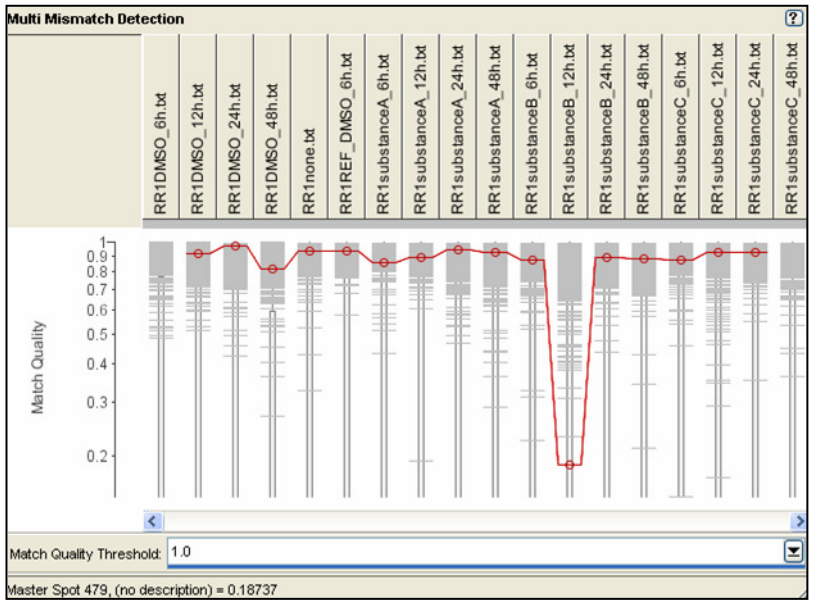




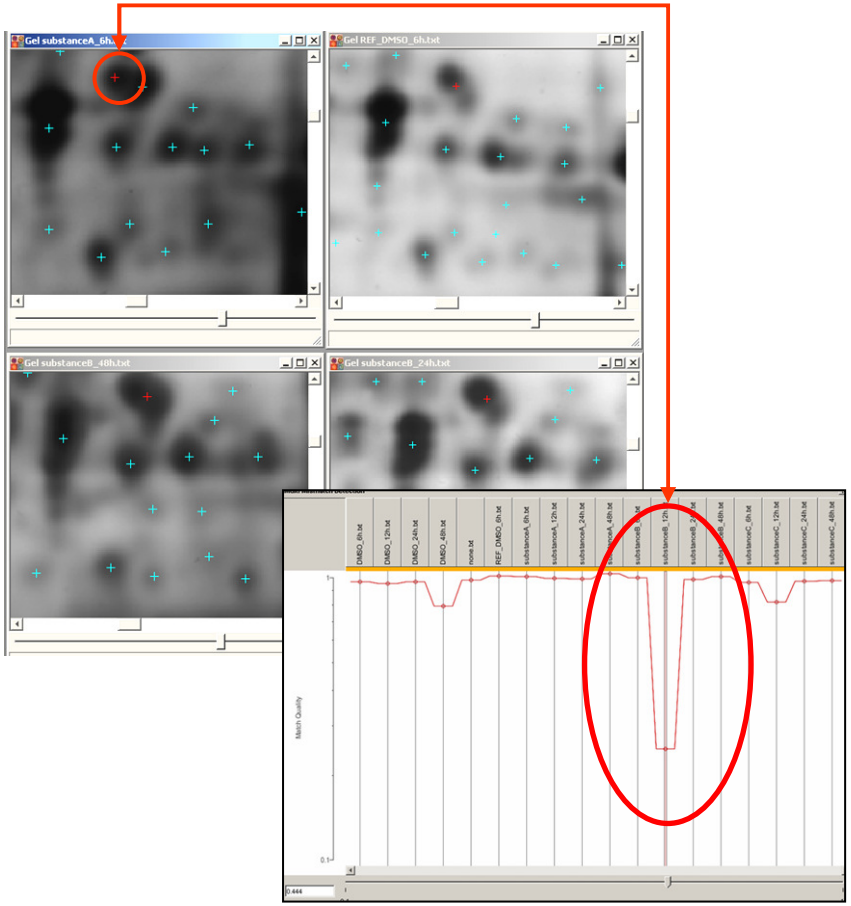


Refiner Proteomics

Compares location of spots over complete gel data set

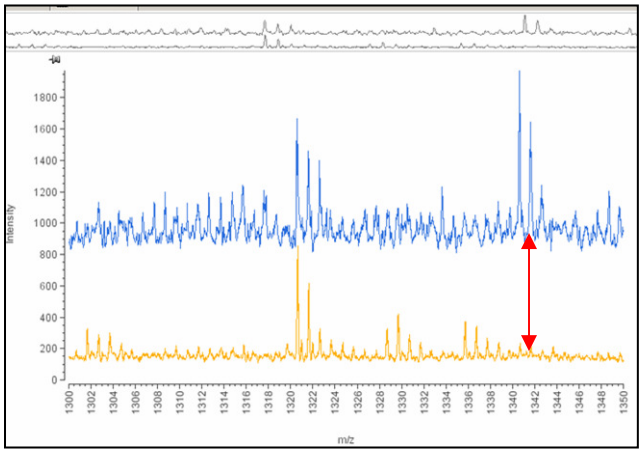


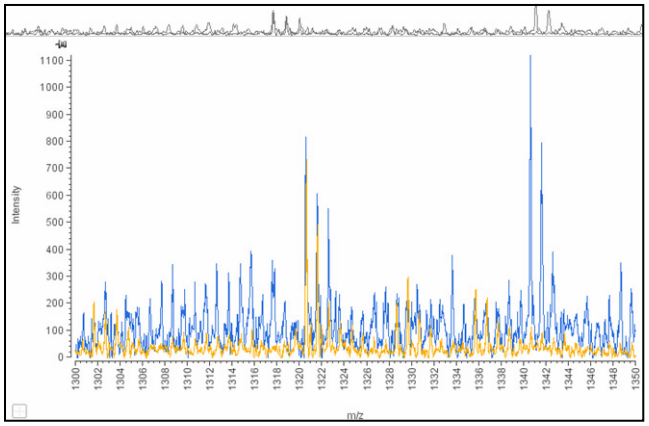
Automated mismatch detection based on calculation of standardized match scores



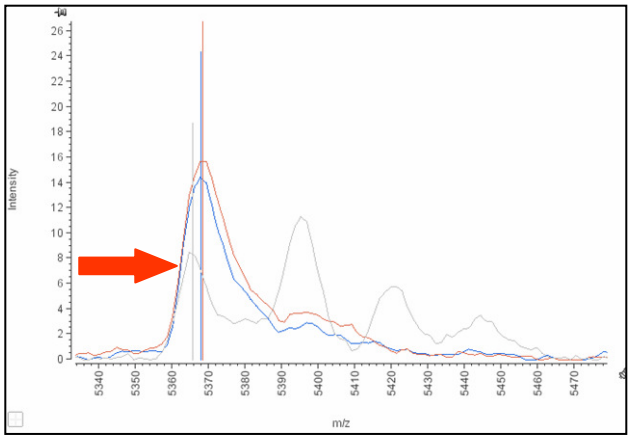
Refiner Metabolomics

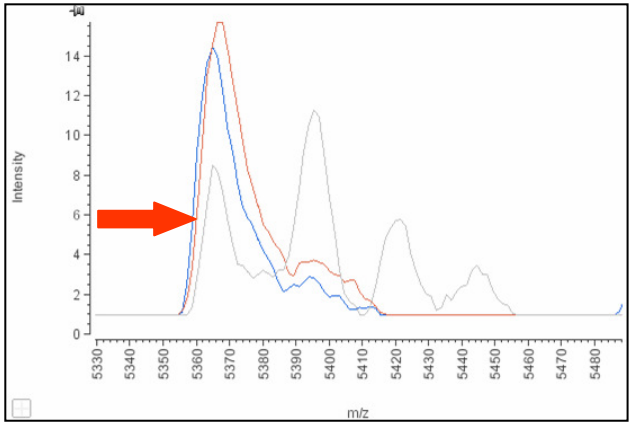
Baseline subtraction increases the comparability of spectra





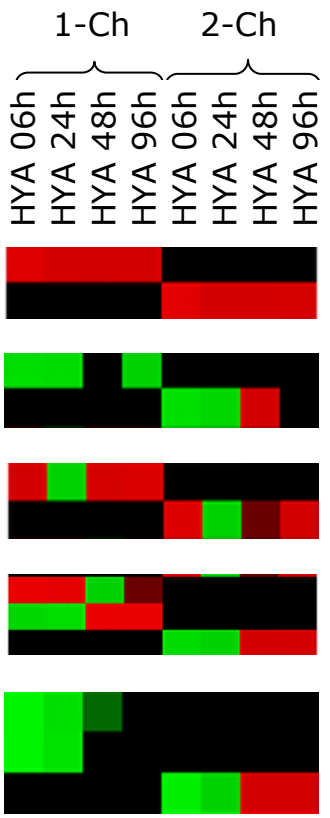
m/z alignment prevents false positives in biomarker detection





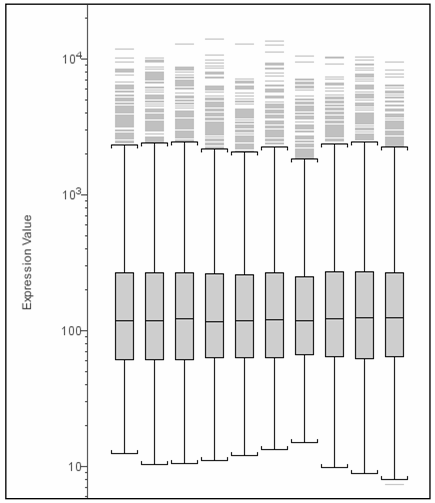
Mapping and normalization

- 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, genes, proteins, etc.)



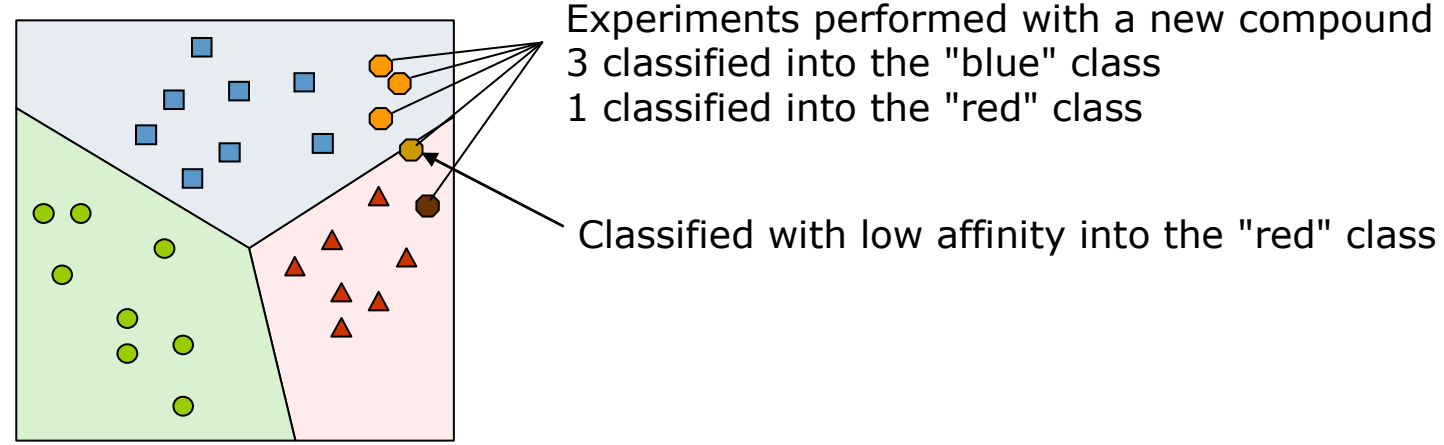
Mapping of data into a gene symbol space

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



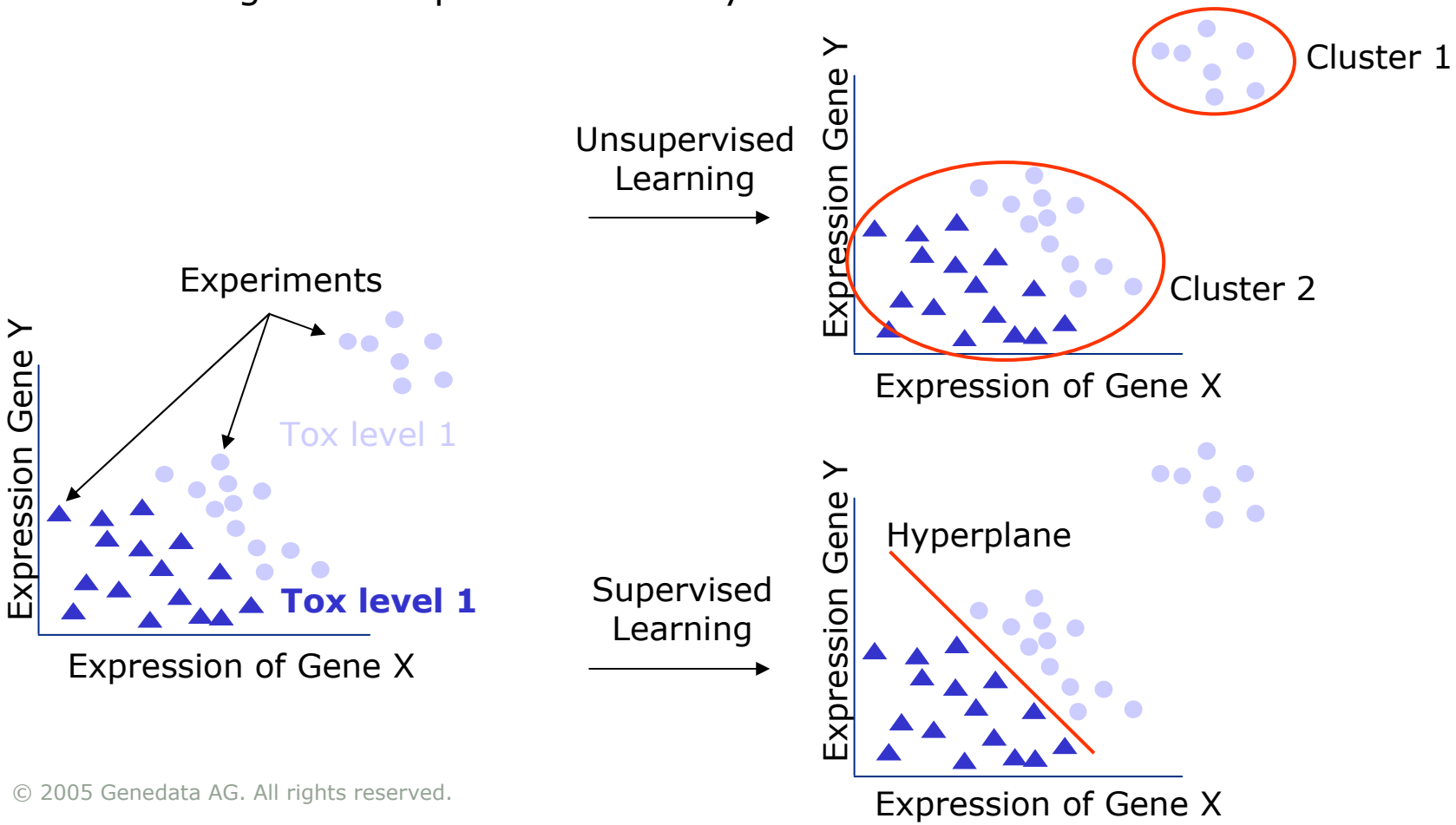
Reference compendium for toxicity prediction and biomarker identification

- 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



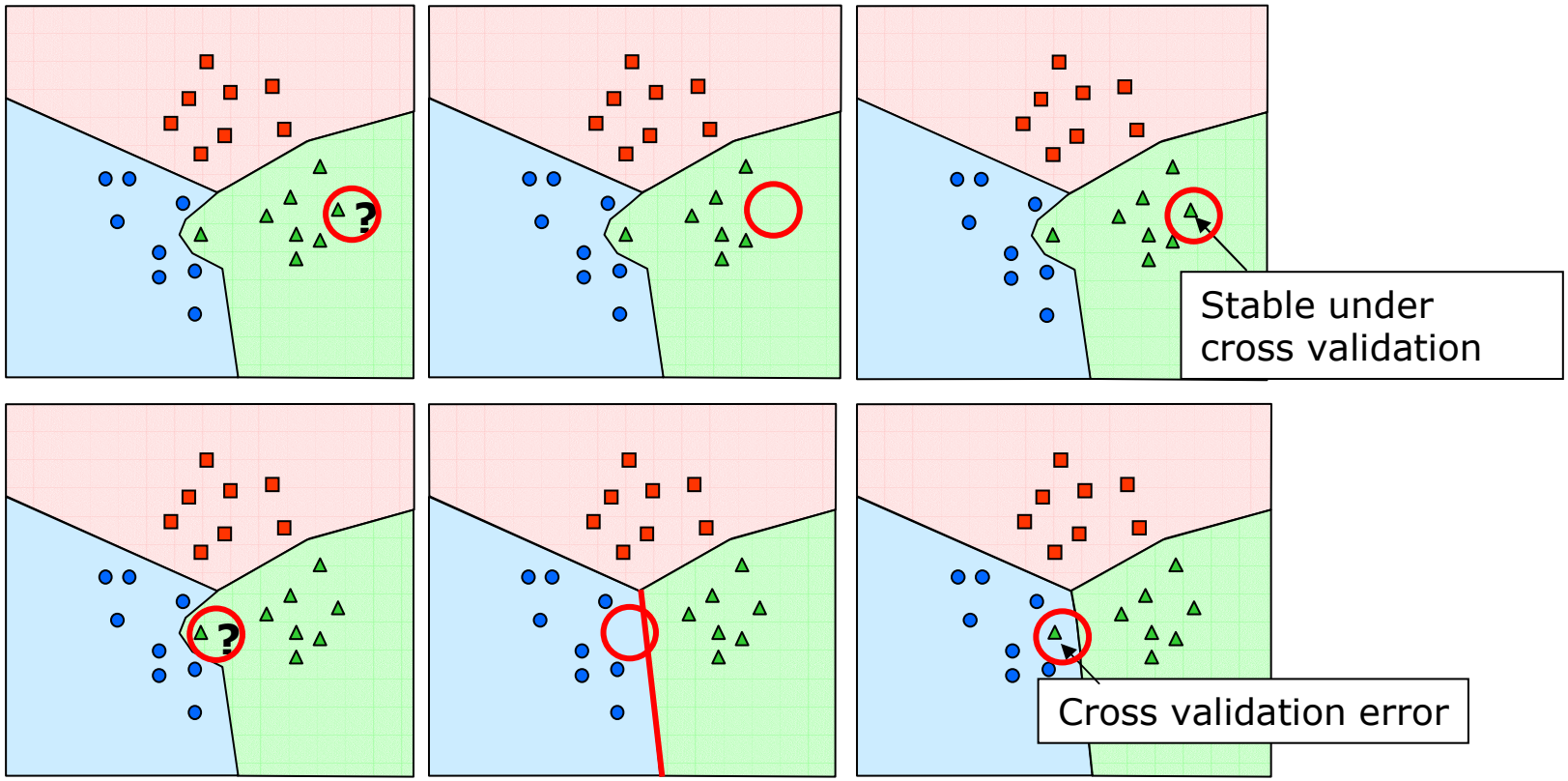
Reference compendium generation

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



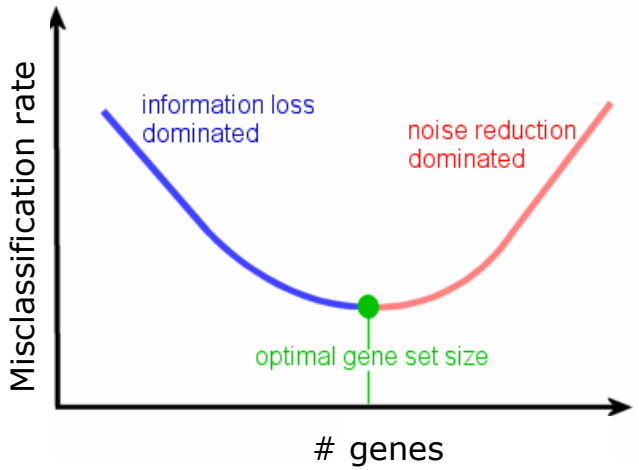
Cross validation of reference compendium

- 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



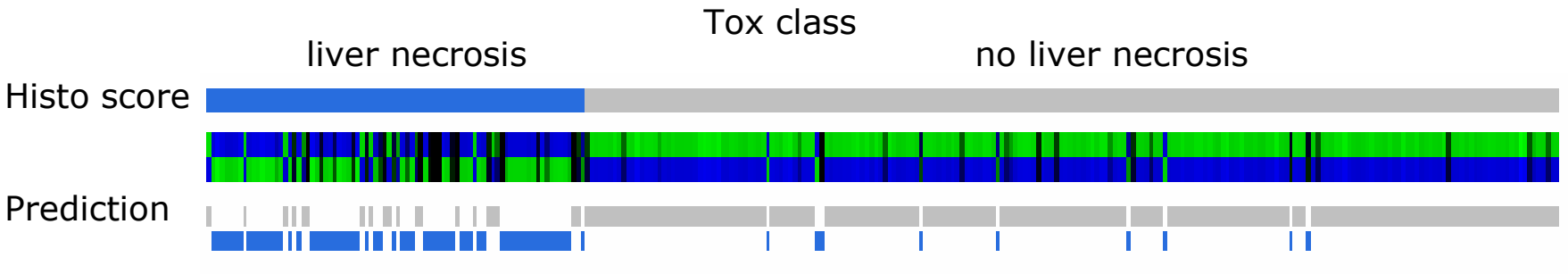
Determination of the optimal set of genes

- 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
  (Best is meant here in terms of minimizing the prediction error)
- Genes from optimal set of genes are potential biomarkers



Prediction of histopathology "liver necrosis" based on histopathological scores

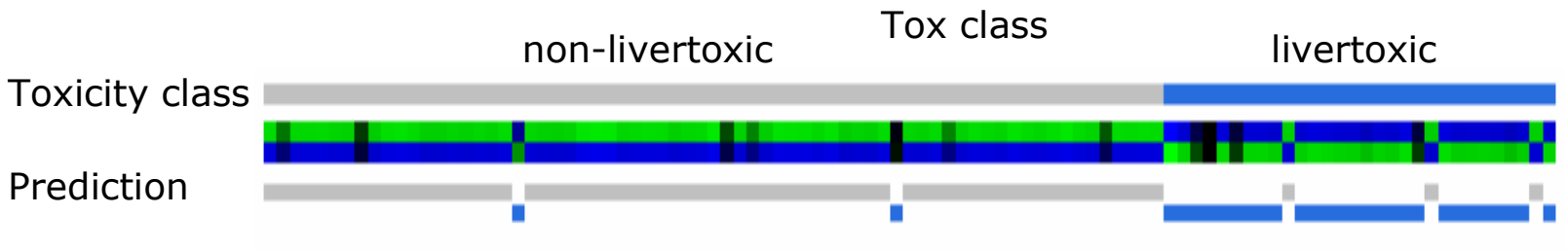
- 52 compounds tested
- Compounds applied at a low and a high concentration
- Samples taken after 6h, 24h and 72h
- Experimental data set included 1597 experiments
- Histopathological scores assigned to each experiment



- **Prediction error: ~ 10%**

Prediction of liver toxicity

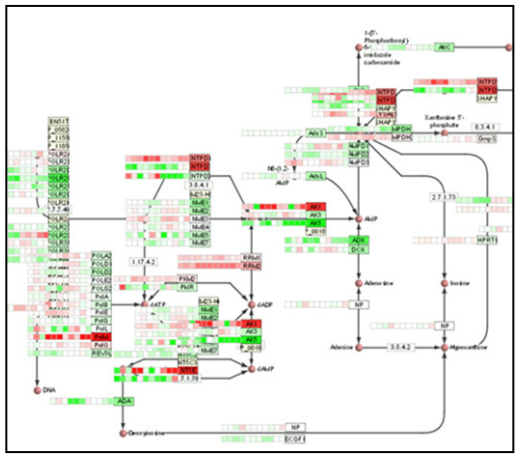
- 33 compounds
- Compounds applied at a low and a high concentration
- Samples taken after 6h, 24h and 72h
- Experimental data set included 958 experiments
- Toxicity class assigned to each experiment

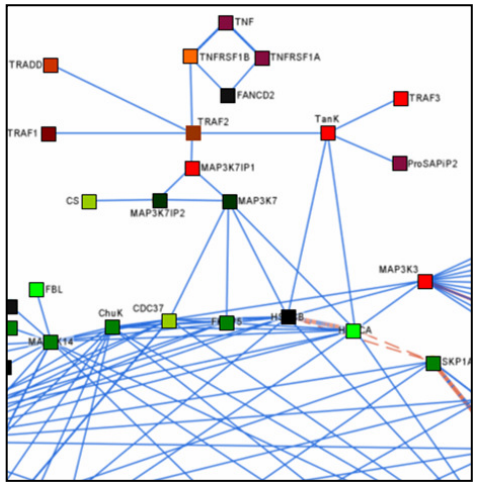


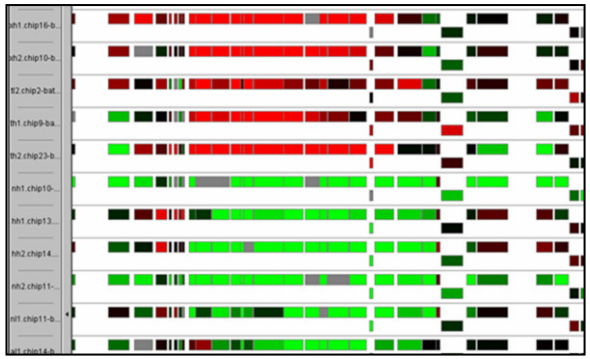
- **Prediction error: ~ 5%**

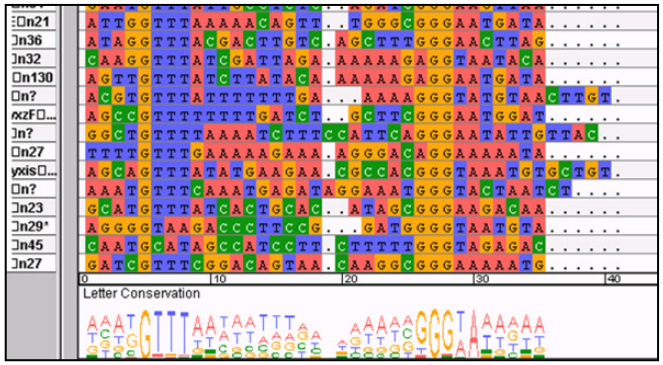
Pathway characterization and biomarker characterization

- 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.









Analysis of co-expression and co-regulation

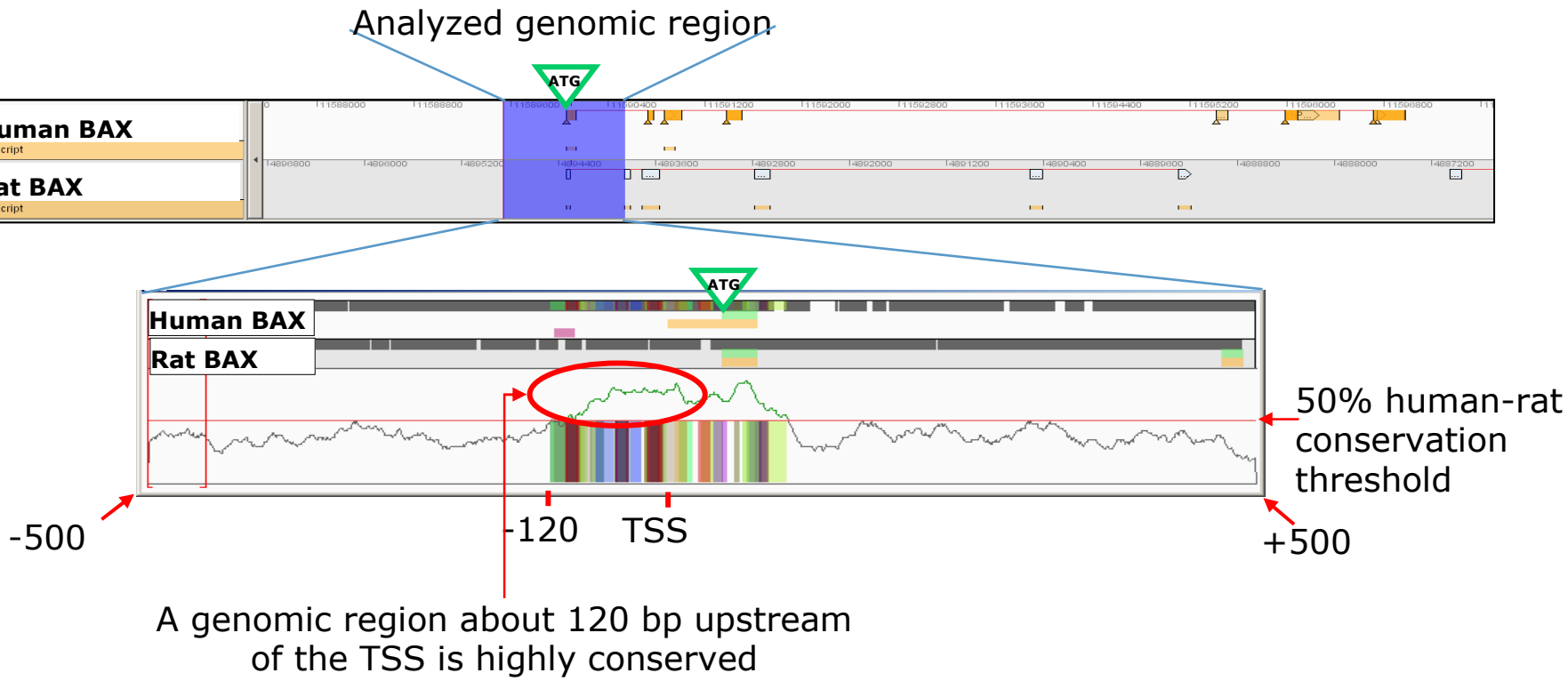
- A number of genes in rat liver were found to be de-regulated by genotoxic hepatocarcinogens [Ellinger et al. 2004]
- Transcription of a small set of genes behaves similarly (co-expression), suggesting a common molecular mechanism for gene regulation (co-regulation)
- A subset of the co-expressed genes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

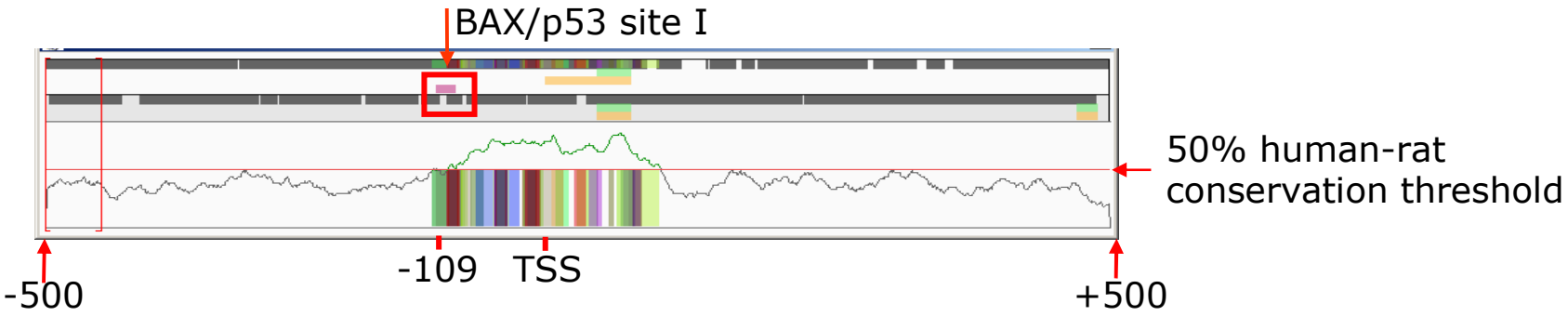
Comparison of the human and rat BAX gene and identifying conserved upstream regions

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



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)

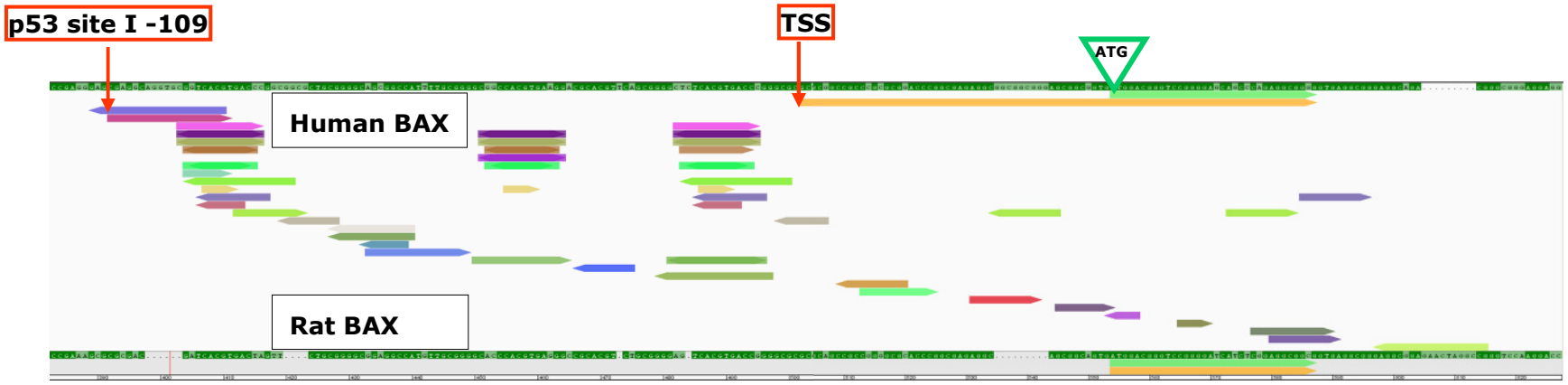


- Two other p53 binding sites can be identified in less conserved regions. One is located further upstream of the TSS (-421 bp; site II), and another in the first intron (+329 bp; site III)

- **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

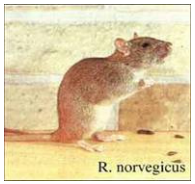


- **At least 16 TFBS sites could be found in the vicinity of p53 sites that are significantly overrepresented in the regulatory regions of genes shown to be co-expressed under genotoxic stress**
- **These factors might cooperate with p53 in the transcriptional activation caused by genotoxic hepatocarcinogens**

Predictive Tox Database

Animal Catalogue

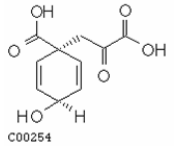
- Species
- Strain
- Sex
- Age
- Tissue
- Histopathology



MoT?
Biomarkers?
Predictability?

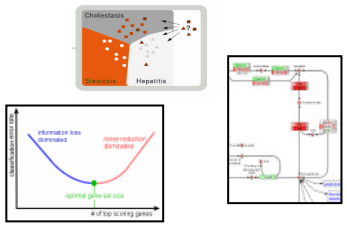
Treatment Catalogue

- Compound
- Concentration
- Treatment time
- Dosing route
- Dosing frequency
- Protocols
- Endpoints
- Histo images
- Histopathology scores
- Clinical chemistry
- Hematology



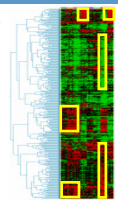
ToxResults Store

- MOA classification
- Biomarker candidates
- Tox Mechanism



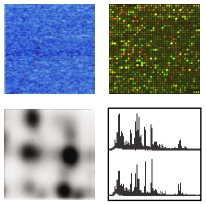
Tox Predictor

- Experimental values
- p-Values
- Gene/protein/ metabolite annotation
- Experiment Annotation



Expression Store

- Affymetrix
- 2 channel
- 2D gel
- MS
- NMR
- Recording devices
- Genes, transcripts, proteins
- Raw data



Thank you

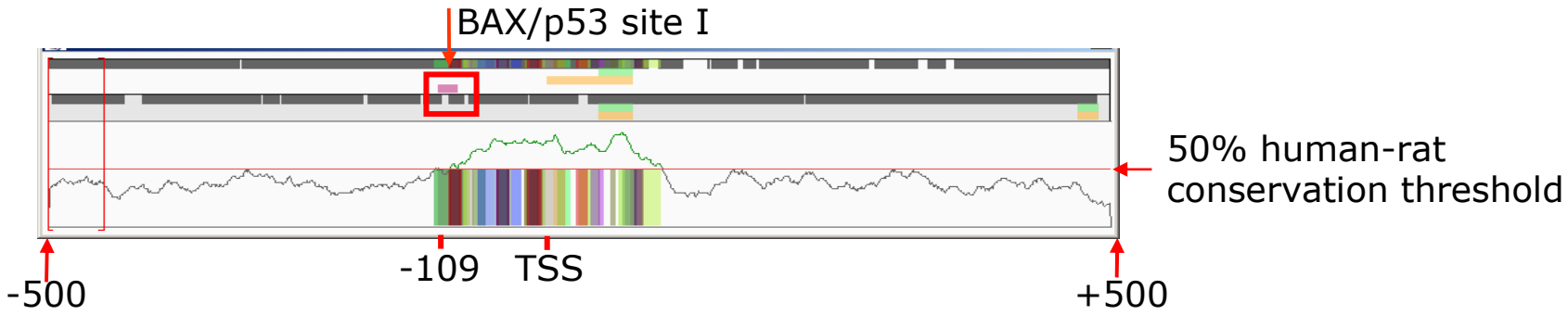
Genedata AG
Maulbeerstrasse 46
CH-4016 Basel, Switzerland
Tel +41 61 697 7651
Fax +41 61 697 7244

hans.gmuender@genedata.com

www.genedata.com

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)