



Funded by the Seventh Framework  
Programme of the European Union



Project full title:  
**Hepatic and Cardiac Toxicity Systems modelling**

Project acronym:  
**HeCaTos**

Collaborative project  
HEALTH.2013.1.3.-1:  
Modelling toxic response in case studies for predictive human safety assessment

**FP7-HEALTH-2013-INNOVATION-1-602156-HeCaTos**

**Deliverable Report D2.3**  
**Report on the modelling and machine**  
**learning pipeline**

*Edited by*  
*Chris Bauer, Ralf Herwig, Johannes Schuchhardt, Arif Malik*  
Work package 2

Due date of deliverable: M18  
Actual submission date: M22

Start date of project: October, 2013

Duration: 60 months

**Maastricht University (UM)**

Project co-funded by the European Commission within the 7th Framework Programme (2013-2018)		
Dissemination Level		
<b>PU</b>	Public	X
<b>PP</b>	Restricted to other programme participants (including the Commission Services)	
<b>RE</b>	Restricted to a group specified by the consortium (including the Commission Services)	
<b>CO</b>	Confidential, only for members of the consortium (including the Commission Services)	

## Contributions to deliverable - Internal review procedure

Deliverable produced by:	Date:
Arif Malik - Partner MD	10.06.2105, 17.07.2015, 09.09.2015
Chris Bauer - Partner MD	22.06.15
Ralf Herwig - Partner MPIMG	10.07.2015, 09.09.2015
Johannes Schuchhardt - Partner MD	17.07.2015
Deliverable internally reviewed by:	Date:
Jos Kleinjans - Partner UM	09.09.2015

## Contents

Publishable Summary.....	3
Objectives .....	3
Introduction .....	4
1. Results .....	6
1.1 Literature mining .....	6
1.2 Differential Expression .....	6
1.3 Multivariate Statistics .....	9
1.3.1 Cross-omics case study: Analysis of RNA-seq and ChIP-seq data for drug-dependent liver regulatory elements.....	9
1.3.2 Analysis of time series and multiple dosage regimens .....	13
1.4 Network modules computed from protein-protein interactions (PPIs) .....	14
1.5 Network modules computed from pathways .....	15
Outlook .....	17
Difficulties .....	17
References .....	18

## PUBLISHABLE SUMMARY

We have developed a computational pipeline that allows the identification of networks and pathways that are predictive for drug toxicity. The pipeline consists of different steps ranging from primary data analysis (e.g. transcriptome and proteome analysis), to the detection of disease associations (e.g. literature mining), multivariate statistical methods for analyzing time series and dose variations, as well as network and pathway analyses. The input of the pipeline consists of case-control drug treatment experiments at several doses and time points and delivers as output modules of interacting proteins and of connected pathways that are mostly affected by the treatment. Such modules can be used as:

- i) The basis for building molecular models;
- ii) Molecular patterns for drug toxicity prediction;
- iii) Functional links to associate the drug effects to cellular phenotypes and thus for construction of adverse outcome pathways (AOPs).

## OBJECTIVES

Aim of the work is to develop a computational pipeline that utilizes omics data and other sources of information derived from drug treatments in order to deliver network modules that are predictive of molecular changes induced by the drug. These network modules are either derived from interaction data or pre-defined molecular pathway concepts and provide information on individual response proteins and their functional interactions. Once such modules are derived they can serve as basis for designing computational molecular models that are associated with known themes of drug toxicity in liver and heart such as mitochondrial dysfunction, apoptosis or immune responses. WP2 is tightly connected with WP7, WP11 and WP12.

Benchmark data sets built in WP12 of chemically treated and untreated samples from (sub-)cellular studies analyzed in WP7 and WP8, will be used to identify a broad spectrum of predictive features. This goal will be achieved in cooperation between partner 10 (MPIMG) and 13 (MD) who will comprise different technologies suitable for the specific goal of modelling.

In this report, we describe the different parts of the pipeline and demonstrate its usage with selected data sets.

## INTRODUCTION

The computational modelling pipeline is illustrated in Figure 1. Starting with a compound and available omics data (e.g. transcriptome, proteome, DNA-protein binding) parallel steps are performed for identifying associated biomolecules. These steps are:

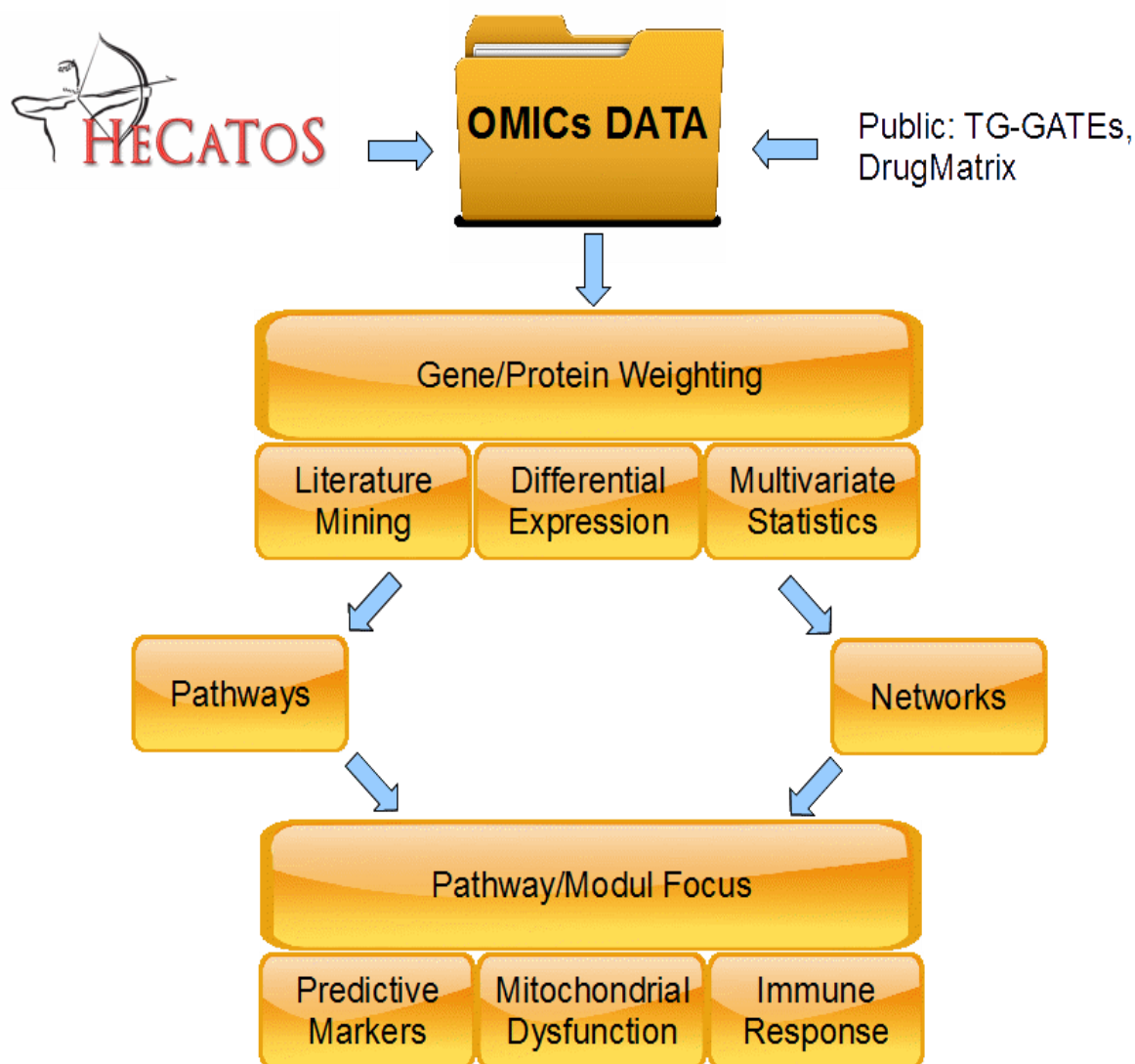
- The identification of differentially expressed genes based on case-control transcriptome data which is the major source of information in toxicogenomics;
- The identification of disease associations using literature mining;
- The application of cross-omics or more complex data analysis based on multivariate statistics.

These strategies are essentially used for assigning each biomolecule under study a weight that corresponds to the level of alteration induced by the drug.

In a further step the identified biomolecules are associated with interaction data in order to describe the compound effects at the level of biological networks. This is done either with pre-defined molecular pathway concepts or with protein-protein interactions integrated from MPIMG's ConsensusPathDB resource (<http://consensuspathdb.org>). The results from the omics analyses are used to weight the components of these interaction networks. Graph algorithms are applied in order to identify those parts of the networks that agglomerate many of the induced changes (network modules). The network modules can be used as:

- The basis for building molecular models;
- Molecular patterns for drug toxicity prediction;
- Functional links to associate the drug effects to cellular phenotypes and thus for construction of adverse outcome pathways (AOPs).

Details on how the different steps are performed are presented below.



**Figure 1:** Different parts of the computational modelling pipeline. Goal of the pipeline is to derive pathways/network modules for toxicity prediction from omics data. These modules will be used for two purposes: *i)* to derive predictive markers; we have processed ca. 1,100 different compound experiments from TG-GATES and DrugMatrix and generated a catalogue of network modules and response pathways characteristic for different compound classes (cf. deliverable report D12.1). These features serve as a training set and machine learning algorithms (support vector machines) are used to assess prediction power and to classify compounds with 'unknown' toxicity. *ii)* to build computational models based on ordinary differential equation systems that are based on specific molecular foci such as mitochondrial dysfunction and immune responses.

## 1. RESULTS

### 1.1 Literature mining

The goal of this part of the pipeline to extract initial biomarker lists and preliminary concepts from public resources and validate the information generated in HeCaTos with additional data.

Literature mining as initial part of the modelling pipeline has already been described in Deliverable D2.1 and was verified with the analysis on public data for five different compounds of primary interest to the consortium as defined by the project board of HeCaTos. The approach leads to compound specific lists of associated biomolecules. In order to select more specific biomolecules the results were compared with a reference set using Fisher's enrichment statistic. For each biomolecule a p-value using Fisher's test statistic was calculated. The p-value is a reliable measure of the strength and specificity of the association of the gene or protein in the context of liver toxicity. For example, among the top hits for rifampin different members from CYP protein-family were detected. Similar encouraging results were found for acetaminophen.

With the most significant genes (genes with p-values < 1E-05) from literature mining a pathway enrichment analysis was performed. The most significant KEGG pathways detected were: Drug metabolism, Metabolism of xenobiotics by cytochrome P450 and Chemical carcinogenesis (with p-value 1E-15). Text mining was performed for all five compounds and respective lists of associations have been made available for the consortium through the project web site.

### 1.2 Differential Expression

The goal of this part of the pipeline is the identification of differentially expressed genes (DEGs) as a first and necessary step to determine which parts of the transcriptome are affected by the drug.

This part of the pipeline has already been described in Deliverable D2.1 and was exemplified with analysis performed with public data for five different compounds of primary interest to the consortium. Differential expression is the straightforward solution if one is interested in experiments done at specific time points resp. with specific dosage regimens and leads to prioritization of genes based on a suitable statistical testing procedure that compares the treated vs. the control experiments. Weights for network analysis are then given with the p-values and fold-changes derived from the test (cf. Deliverable Report D2.1). We generated such analysis initially for five compounds (liver: acetaminophen, rifampin, azathioprine; heart: cyclophosphamide, doxorubicin) in the first reporting period and uploaded respective tables to the project repository for usage in the other work packages. In the second reporting period we extended the analysis for all drugs tested in different tissues that are available from the public benchmark data sets DrugMatrix and TG-GATES (ca. 1,100 compounds).

#### Discovering Drug Patterns

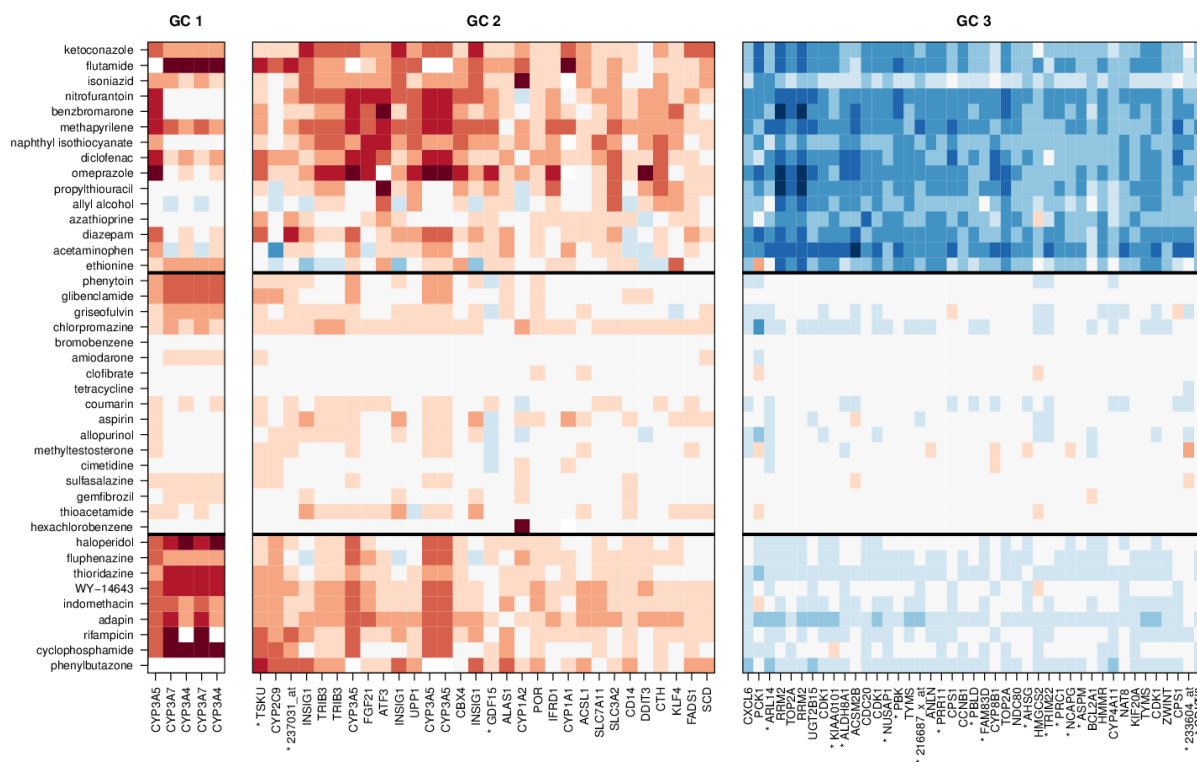
In order to challenge this part of the pipeline, partners elaborated whether it is necessary to discover common hepatotoxic mechanisms from DEGs induced by multiple drugs. We investigated a large dataset from TG-GATES for human hepatocytes with 113 toxic compounds, with up to 4 different doses (control, low, middle, high doses) and 3 different time points (2h, 8h and 24h).

On the level of differential gene expression (highest dose and latest time point) we see 3 groups of genes (see three heat maps below in Figure 2):

1. Genes with strong up-regulation by many compounds. This cluster contains Cyp-genes which have been reported as top-scoring microarray results for rifampin in the previous report. The

proteins are mostly responsible for detoxification and degradation of xenobiotics. The drugs show different patterns of Cyp-activation. Drugs like rifampicin, haloperidol or cyclophosphamide strongly induce Cyps while compounds like tetracycline or aspirin show almost no up-regulation;

2. Genes with medium up-regulation by many compounds. This cluster contains also some Cyp-genes and other genes from drug metabolism pathways;
3. Genes with down-regulation by many compounds. These genes are involved in fatty acid degradation and arachidonic acid metabolism. Down-regulation of these pathways strongly hints for mitochondria dysfunction.



**Figure 2:** Heat maps for differential gene expression (highest dose and latest time point) identify 3 groups of genes.

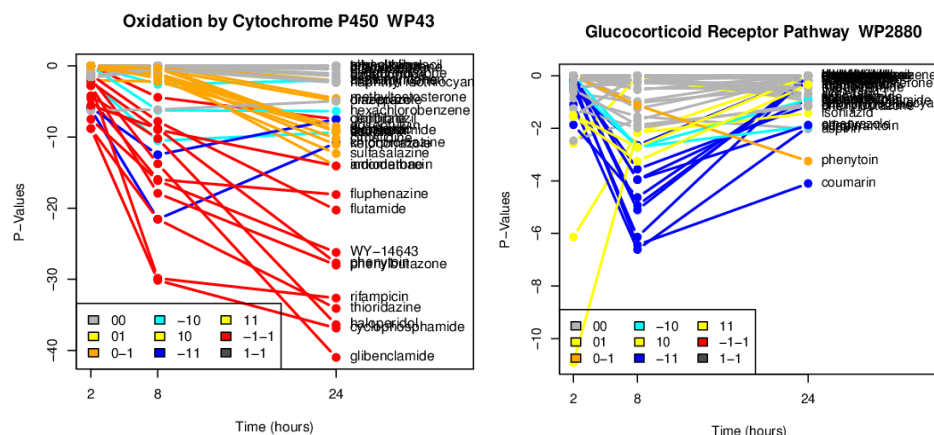
We can also see three different clusters of compounds:

1. Drugs in the lower part of the heat map (haloperidol, fluphenazine, thioridazine, indomethacin, adapin, rifampicin and cyclophosphamide) strongly induce cluster 1 and show only weak down-regulation in cluster 3.
2. Drugs in the upper part of the heat map (e.g. isoniazid, diclofenac, diazepam, etc.) show a common down-regulation of the genes in cluster 3.
3. Drugs in the middle (e.g. aspirin or tetracycline) part show almost no up-regulation or down-regulation of any of these genes.

### Time course of pathways

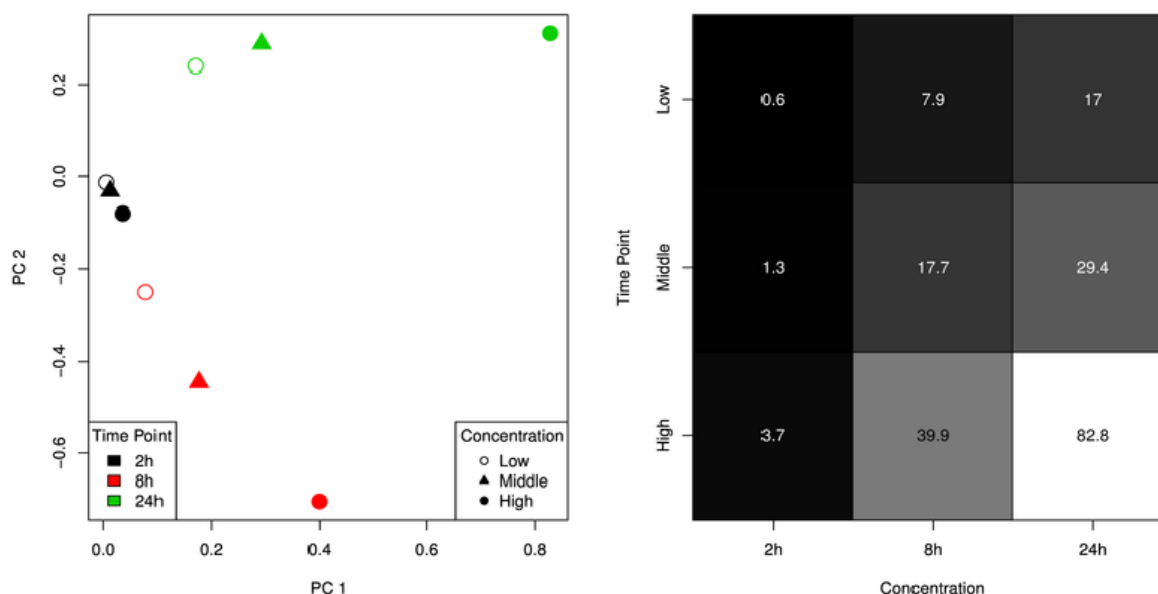
Time-resolved measurements are an important aspect to model hepatotoxicity. Some changes in gene expression may happen very early while others happen later. On the level of pathways we see that some pathways, e.g. glucocorticoid receptor pathway are induced rather early after 8 hours by several drugs. The most significant results are obtained after 8 hours. For some compounds (isoniazid and acetaminophen) even after 2 hours.

Other pathways, especially drug degradation pathways are increasing in significance by time.



**Figure 3:** Differential analysis was based on the assumption that the latest time point with the highest concentrations shows the strongest effects. By looking at time courses of pathways we have seen that the other time points are also important for modelling hepatotoxicity.

To more comprehensively investigate this effect we perform a principle component analysis (PCA) using time points and concentrations. The PCA is shown in Figure 4. The first principle component shows a very high correlation to the latest time point and highest concentration (Pearson Correlation Coefficient = 0.93). The first principle component accounts for more than 25% of the overall variability. But the other concentrations and the other time points are also important. Therefore in order to achieve maximal predictive power all doses and all time points should be incorporated into a multivariate model.



**Figure 4:** Principle component analysis (PCA) using time points and concentrations



### 1.3 Multivariate Statistics

The goal of this part of the pipeline is to deliver methods for more complex set-ups of toxicogenomics experiments in particular for:

- i) cross-omics analysis and
- ii) time- and dose-resolved omics data sets.

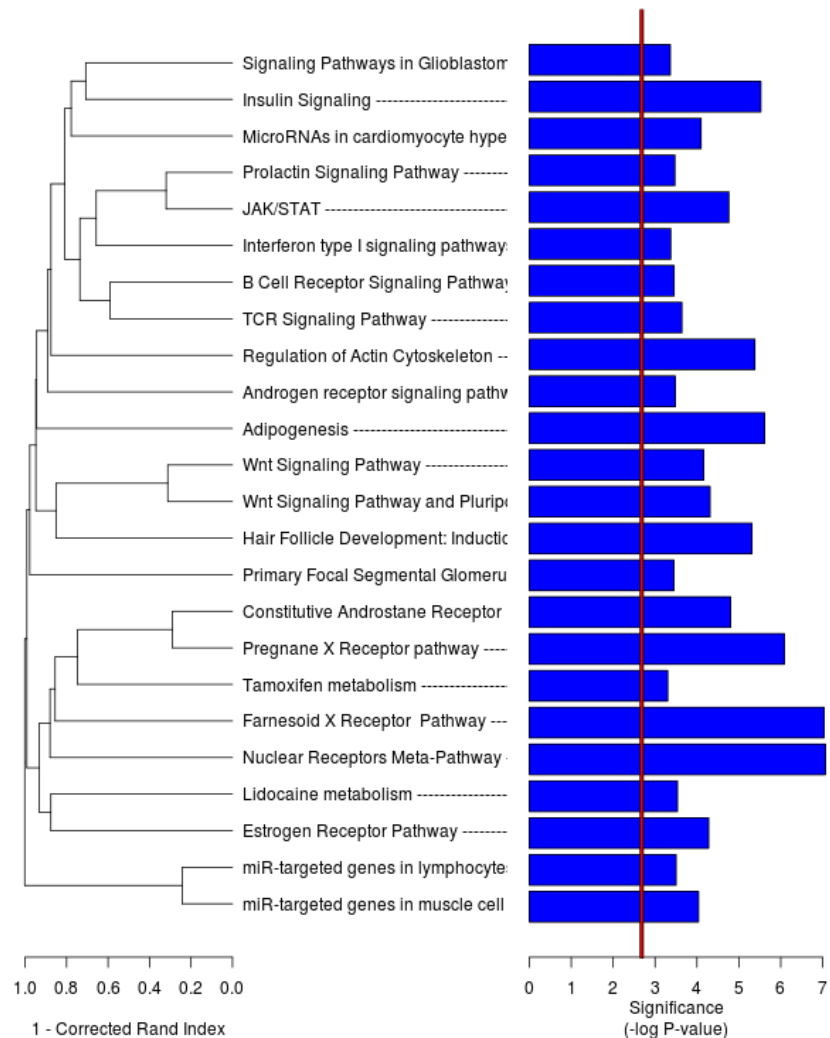
#### *1.3.1 Cross-omics case study: Analysis of RNA-seq and ChIP-seq data for drug-dependent liver regulatory elements*

In their study, Smith et al. (PLoS Genet. 2014 Oct 2; doi: 10.1371/journal.pgen.1004648) analyzed human hepatocytes for the influence of rifampin treatment. ChIP-seq was performed using antibodies against pregnane X receptor (PXR), p300, and the histone marks H3K27ac and H3K4me1 on primary human hepatocytes treated with rifampin or DMSO as control. Rifampin and PXR were chosen since they are part of the CYP3A4 pathway which accounts for the metabolism of more than 50% of all prescribed drugs. Both the RNA-seq and ChIP-seq data are available from Sequence Read Archive (SRA). We identified target genes of PXR under rifampin treatment and compared it to the differentially expressed genes occurring in the RNA-seq analysis.

##### A) ChIP-seq

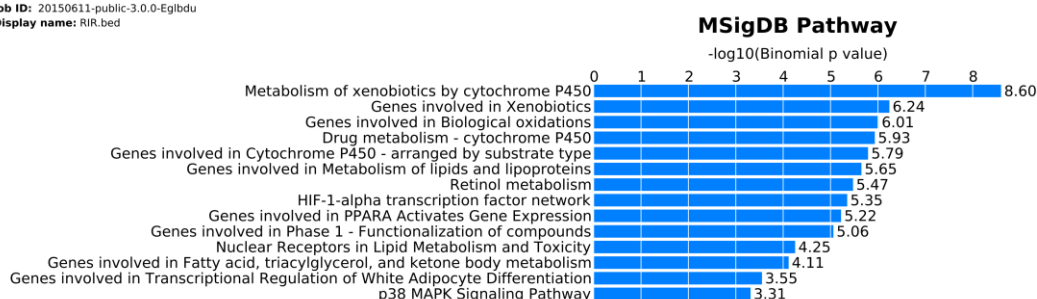
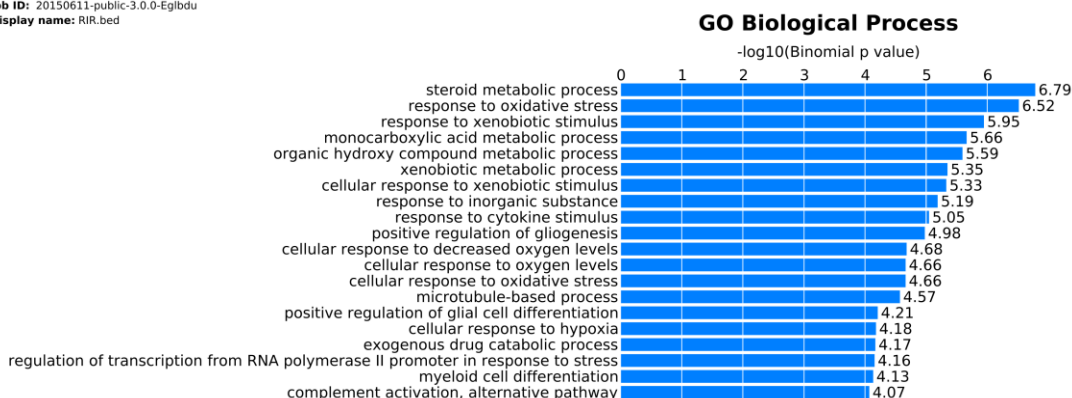
After application of our basic ChIP-seq pipeline (QC of raw data, trimming, mapping, duplicate removal), transcription factor binding sites and regulatory regions were determined by peak calling with MACS2.

First, we determined PXR binding sites occurring only under rifampin treatment and not in the control. The binding sites were annotated with the nearest gene and these were subjected to a pathway analysis using WikiPathways as resource. Here, nuclear receptor pathways were among the top hits (Figure 5).



**Figure 5:** Pathway analysis of genes associated with ChIP-seq binding sites. WikiPathways was used as resource.

For a more comprehensive view, genomic regions for both treatments were identified where all 4 marks were present. Then, regions that occur only within rifampin treatment were defined as rifampin-induced regions (RIR) and were of further interest. These regions were annotated using the Genomic Regions Enrichment of Annotations Tool (GREAT) which also incorporates long-range regulations. The enriched gene ontology biological processes and MSigDB pathways are shown below. Here, oxidative stress and xenobiotics pathways were among the most prominent hits (Figure 6).



**Figure 6:** Genomic Regions Enrichment of Annotations Tool (GREAT) analysis of rifampin-induced regions.

## B) RNA-seq

The study contained RNA-seq data from two replicates each of DMSO and rifampin treated hepatocytes. After basic processing (QC, trimming, mapping with TopHat2), a differential analysis was performed using DESeq2. After applying a p-value cut-off of 0.05 (after correction for multiple testing), 151 genes were identified as differentially expressed. Of these, 97 were up-regulated in rifampin treatment and 54 were down-regulated (Figure 7).



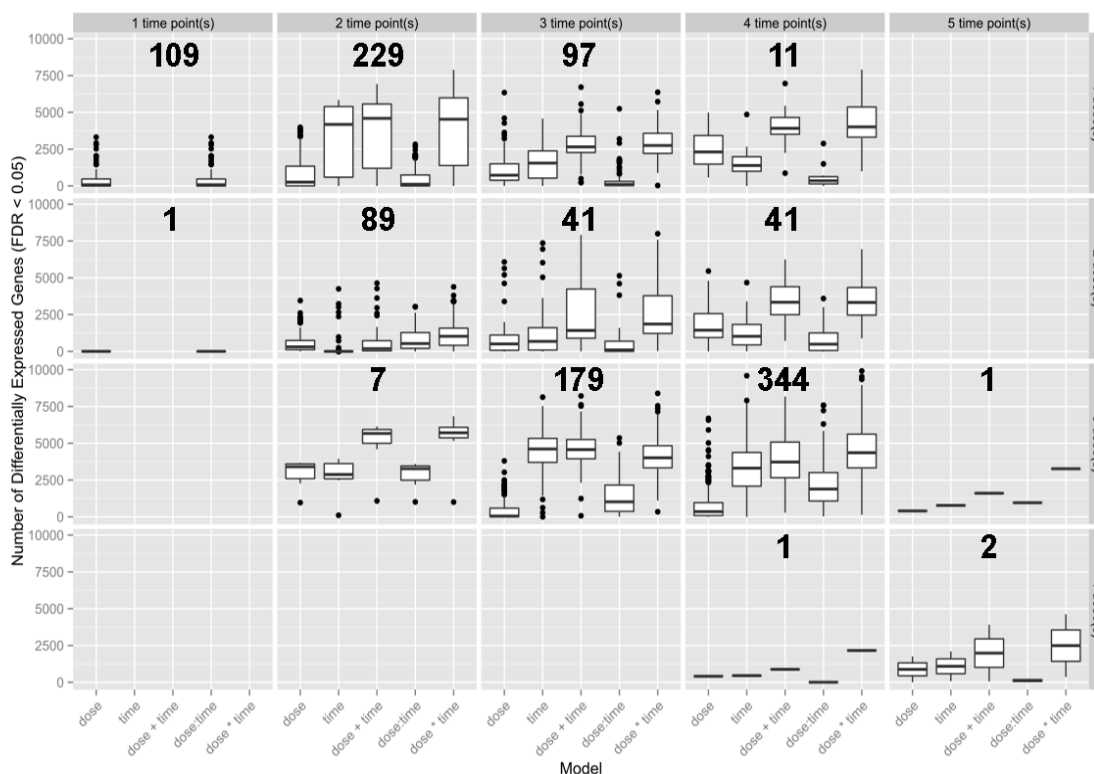
### 1.3.2 Analysis of time series and multiple dosage regimens

A typical toxicogenomics experiment consists of time-resolved drug treatments along with variations of the dose. In contrast to DEG analysis described in point 1.2 above the “*multivariate statistics*” part of the pipeline evaluates data derived from the same drug treatment at multiple time points and with several doses simultaneously by building linear models with respective definitions of co-variates. We implemented, tested and applied such analysis intensively in the second reporting period with the R-tool *limma*.

MPIMG analyzed all data sets that were made available from DrugMatrix and TG-GATES studies (1,152 drug treatment data sets). We have explored several linear models in order to identify time- and dose-dependent changes of expression. Five statistical models were considered including parameters for:

1. Time dependency only;
2. Dosage dependency only;
3. Time and dosage dependency;
4. Time-dosage interaction factor dependency;
5. Full model dependency consisting of time, dosage and time-dosage interaction factor.

Figure 9 displays the different numbers of significantly regulated genes (Y-axis; FDR<0.05) according to the different models tested (X-axis). The number of differentially regulated genes can be used as an indicator of the sensitivity of the various models and is highest with respect to the full model.



**Figure 9:** 1,152 expression data sets were analyzed with different numbers of time points and dosages (e.g. 179 drugs in TG-GATES and DrugMatrix have 3 doses and 3 time points). X-axis corresponds to the different implemented models (dose: only dose as factor; time: only time; dose + time: dose and time as factors; dose:time: only the dose-time interaction factor; dose \* time: dose, time and interaction factor). Box plots show the number of significant DEGs.

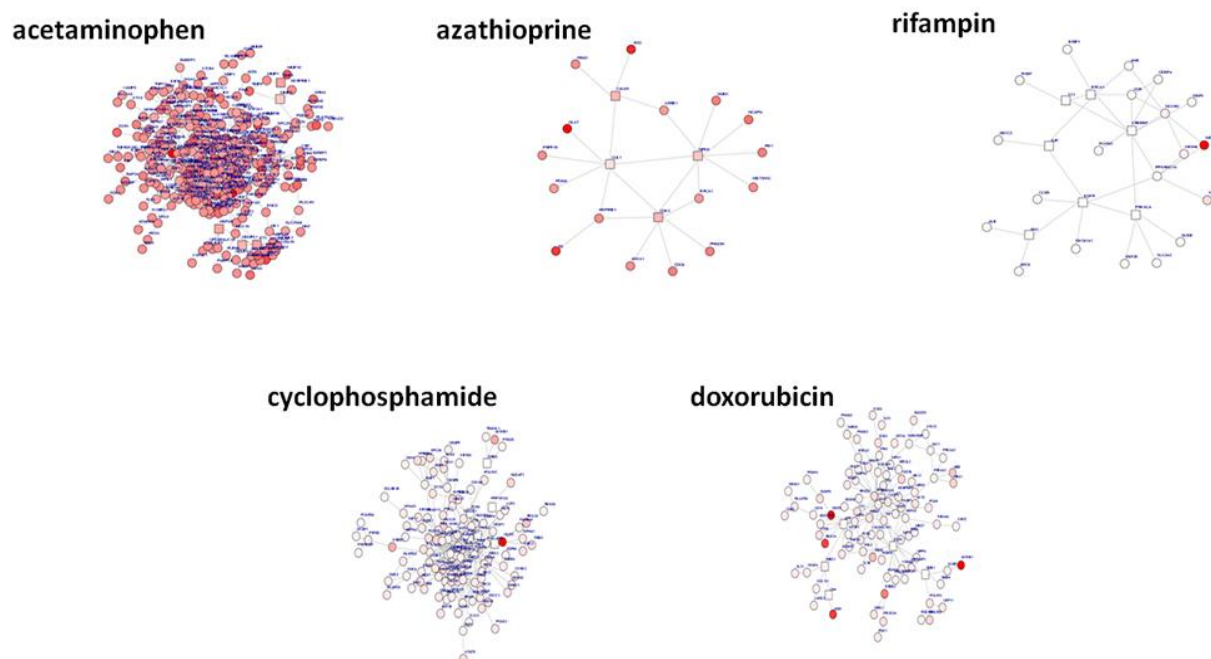
The investigation shows that the full model is the most sensitive and thus, deviation and p-value per gene derived from this model was used to weight the gene in the subsequent network analyses.

### 1.4 Network modules computed from protein-protein interactions (PPIs)

The goal of this part of the pipeline is to map the found gene expression changes, either by differential analysis, literature mining or multivariate time-dose modelling, onto PPI networks in order to identify network modules that are predictive of drug toxicity.

This part of the pipeline has already been described in Deliverable 2.1. In short, as the basis we use the protein-protein interaction network agglomerated in the ConsensusPathDB. This network consists of initially 238,766 protein-protein interactions collected from 19 different public resources (e.g. IntAct, BIND, HPRD among others). After merging the different resources into a large network we have performed quality assessment for each interaction in order to reduce false positive error and to derive a reduced high quality interaction network. As quality assessment we used our IntScore method (Kamburov et al., 2012; <http://intscore.molgen.mpg.de>) which applies six different confidence measures and computes an aggregate score for each interaction that judges its significance. These measures are either based on published topological criteria (geometric embedding, common neighbors, CAPPIC) or on annotation-based criteria (GO semantic similarity, literature evidence, pathway co-occurrence). The aggregate score is then computed from these six measures and serves as an indicator for the confidence in the individual interactions. Using only interactions with confidence ( $>0.5$ ) yielded a PPI of 9,533 proteins and 80,422 interactions.

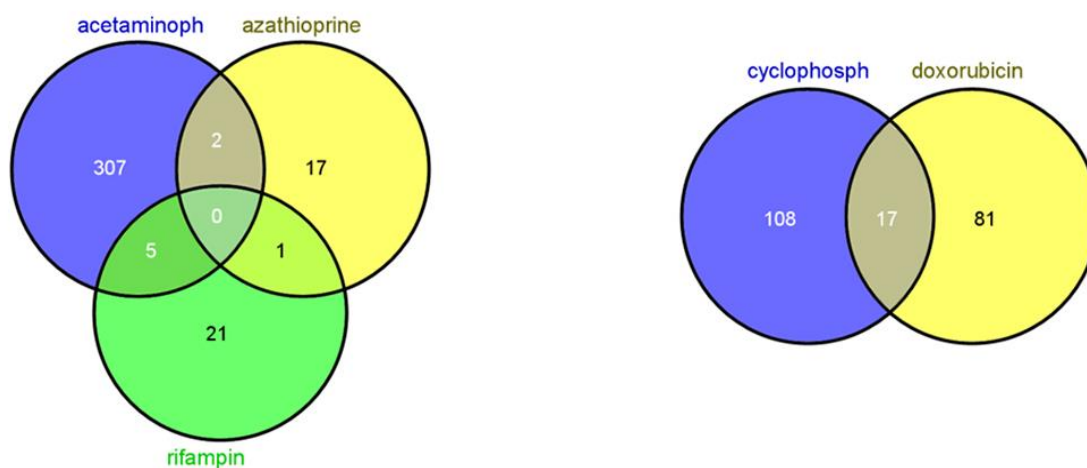
In the first reporting period we have described the identification of network modules based on an integer linear programming approach, called BioNet, published by Beisser et al. (2007). The method is described in Deliverable Report D2.1. This algorithm yielded for each drug a network module computed from the weighted PPI network (Figure 10) that can be interpreted as the most affected part of the network for each drug and thus serves as a network-based predictor of drug toxicity.



**Figure 10:** Network modules computed from the first five drugs.

It has been exemplified in reporting period 1 that, in principle, such a procedure yields reasonable network modules for each drug that could be related to biological function and describe key events that determine phenotypic outcomes. Network modules for the drugs of primary interest have been uploaded to the project repository and are subject to further validation in other WPs.

However, a fundamental problem of this approach is that network modules computed from different drug treated expression data are fairly separated so that it is hardly possible to observe overlap in proteins when screening several drugs. This makes it hard to determine a network that is predictive for a larger number of drugs (cf. Deliverable D12.1) e.g. by identifying commonalities between these drugs. Figure 11 below demonstrates the limited overlap of the three hepato- and the two cardio-toxic drugs of primary interest.



**Figure 11:** Overlap of proteins from the computed PPI network modules. Left: hepatotoxic drugs; Right: cardiotoxic drugs.

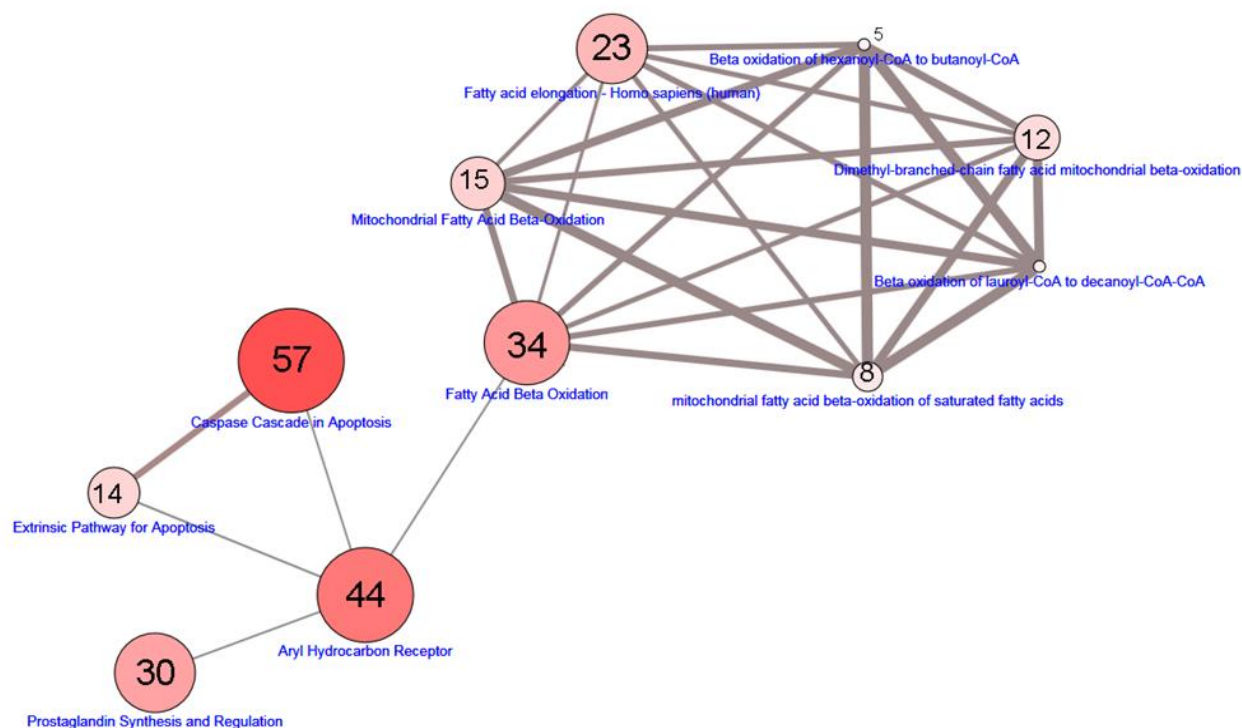
In order to improve this situation, we added a pathway-based concept for network analysis in reporting period 2 which is described in the next subsection.

### 1.5 Network modules computed from pathways

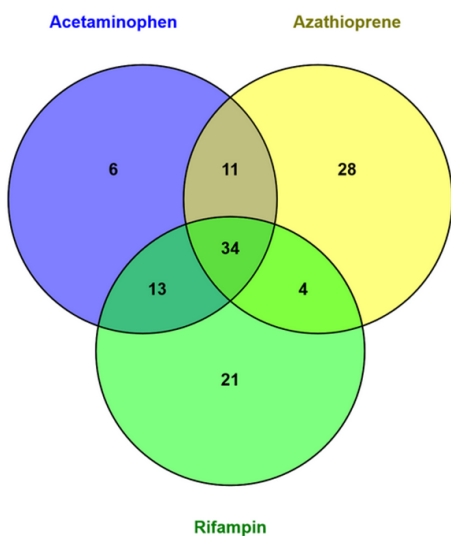
The goal of this part of the pipeline is to improve generality of toxicity prediction. Previously, we have already observed and shown that interpreting toxicogenomics data at the pathway level enhances toxicity prediction, for example in the context of carcinogenicity (Yildirimman et al., *Toxicol Sci* 124:278-290, 2011; Doktorova et al., *Carcinogenesis* 34:193-1402, 2013). Thus, in order to gain a higher overlap in network modules derived from different drugs we have replaced the original PPI network (where each node corresponds to a protein) by a network where nodes represent entire pathways. Here, we used the contents of the ConsensusPathDB (<http://consensuspathdb.org>) which agglomerates 4,349 such pathway concepts from 11 public resources such as KEGG, Reactome, WikiPathways, among others.

In order to derive a network structure we draw an edge between two pathway nodes if there is significant overlap between the gene sets corresponding to the pathways. We require that the minimal node size is 5 (i.e. we consider only pathway concepts with at least 5 associated genes), and draw an edge between two nodes if the  $FDR < 1.0E-06$ , where FDR is derived from Fisher's test with Bonferroni corrections for multiple testing. This network consists of 3,199 pathway nodes and 86,571 edges among them. Figure 12 displays a part of the network consisting of redundant pathways annotated for "mitochondrial fatty acid beta oxidation" which exerts a dense substructure of similar annotated pathway concepts.



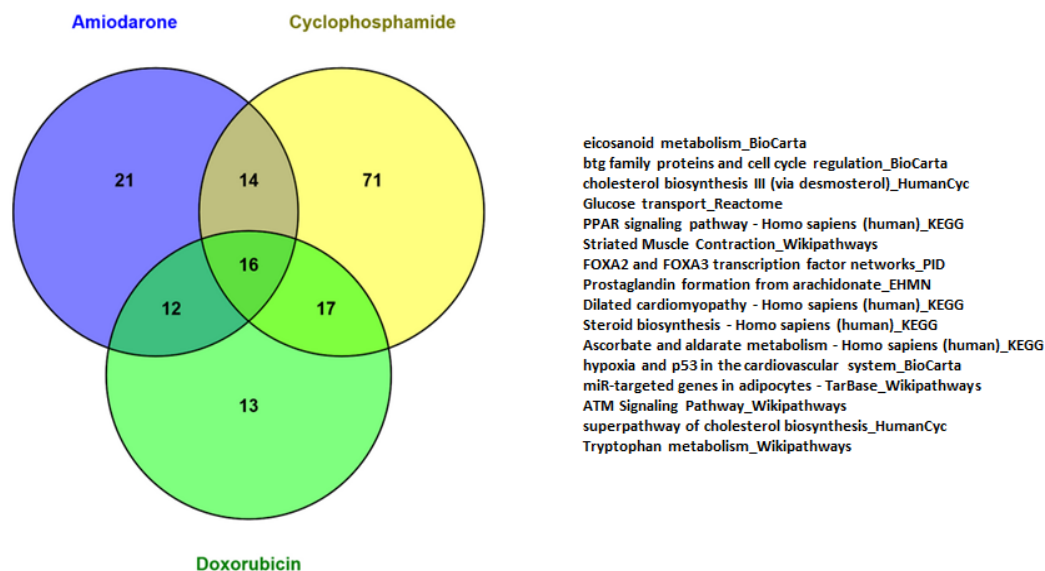


**Figure 12:** Example of pathway overlap. Several pathway concepts for “mitochondrial fatty acid beta oxidation” derived from different database sources.



Arylhydrocarbonreceptor (AhR) signaling pathway\_Wikipathways  
 CDC6 association with the ORC:origin complex\_Reactome  
 role of ran in mitotic spindle regulation\_BioCarta  
 Basigin interactions\_Reactome  
 Mismatch repair\_Wikipathways  
 G1/S Transition\_Reactome  
 retinol biosynthesis\_HumanCyc  
 Fluoropyrimidine Activity\_Wikipathways  
 akap95 role in mitosis and chromosome dynamics\_BioCarta  
 FOXA2 and FOXA3 transcription factor networks\_PID  
 Assembly of the pre-replicative complex\_Reactome  
 Tryptophan catabolism\_Reactome  
 nicotine degradation III\_HumanCyc  
 lectin induced complement pathway\_BioCarta  
 serotonin degradation\_HumanCyc  
 SCF-beta-TrCP mediated degradation of Emi1\_Reactome  
 Leading Strand Synthesis\_Reactome  
 Repair synthesis of patch ~27-30 bases long by DNA  
 polymerase\_Reactome  
 Activation of APC/C and APC/C:Cdc20 mediated degradation of mitotic  
 proteins\_Reactome  
 Chagas disease (American trypanosomiasis) - Homo sapiens (human)\_KEGG  
 Ascorbate and aldarate metabolism - Homo sapiens (human)\_KEGG  
 Regulation of DNA replication\_Reactome  
 Complement Activation, Classical Pathway\_Wikipathways  
 TNF signaling pathway - Homo sapiens (human)\_KEGG  
 Trans-sulfuration and one carbon metabolism\_Wikipathways  
 hypoxia and p53 in the cardiovascular system\_BioCarta  
 Glutathione metabolism\_Wikipathways  
 nicotine degradation IV\_HumanCyc  
 G2 Phase\_Reactome  
 Aurora A signaling\_PID  
 DNA strand elongation\_Reactome  
 Tryptophan metabolism\_Wikipathways  
 Aryl Hydrocarbon Receptor\_Wikipathways





**Figure 13:** Common pathway concepts from the hepatotoxic (upper panel) and cardiotoxic (lower panel) drugs identified with network analysis.

Indeed, it could be observed that repeating the network analysis on the graph composed of pathway patterns resulted in a much better overlap (Figure 13) so that more consistent network modules could be identified for many drugs. Actual results of this analysis are described in Deliverable Report D12.1.

## OUTLOOK

The pipeline developed in WP2 will be used in the next reporting period for classifying new compound data generated by WP7 in order to understand toxicity mechanisms induced by these compounds in heart and liver tissue. Furthermore, the network module approach in WP2 has direct implications for WP1 (computational chemistry) and WP3 (physiological modelling). WP2 partners have already received lists of molecules that present known drug targets and potential drug targets from computational chemistry predictions (WP1) which can be directly assessed with the pipeline; additionally, WP2 partners have received a list of molecules that are forming the cardiac physiological model in WP3 which will be evaluated across the different omics data sets. In a first approach we will process these gene lists as novel “pathway concepts” through our pipeline and evaluate the molecular themes identified. These attempts will allow a better interpretability of the different modelling approaches and open possibilities to include the results of the different approaches into a common AOP description.

## DIFFICULTIES

We experienced a delay of 4 months with this deliverable due to problems arising during the testing phase of the pipeline. In order to test the pipeline properly and to challenge it intensively with omics data WP2 partners decided to extend the analysis from the drugs of the priority list (ca. 30) to all available drugs (ca. 423) which took some extra time. Additionally, partners screened all available data from TG-GATES and DrugMatrix public studies which included also tissues different from heart and liver (e.g. kidney). Reason for this was to build a compendium of pathways/networks relevant for as many drugs in as many tissues as possible which will help to assess and understand toxicity mechanisms derived from the *in vitro/in vivo* samples generated in WP5 and WP6. Furthermore, we have extended

computation of network modules from PPIs to pathway-based networks in order to gain more generally applicable predictive patterns which was a point that needed some extra developing time.

## REFERENCES

**Hendrickx DM**, Aerts HJ, **Caiment F**, Clark D, Ebbels TM, Evelo CT, **Gmuender H**, **Hebels DG**, **Herwig R**, Hescheler J, **Jennen DG**, **Jetten MJ**, Kanterakis S, **Keun HC**, Matser V, Overington JP, Pilicheva E, **Sarkans U**, Segura-Lepe MP, Sotiriadou I, **Wittenberger T**, Wittwehr C, Zanzi A, **Kleinjans JC**. (2015) diXa: a data infrastructure for chemical safety assessment. *Bioinformatics*, 31(9):1505-7.

**Herwig R**. (2014) Computational modelling of drug response with application to neuroscience. *Dialogues Clin Neurosci*, 16(4):465-77.

**Hebels DG**, **Jetten MJ**, Aerts HJ, **Herwig R**, Theunissen DH, **Gaj S**, van Delft JH, **Kleinjans JC** (2014) Evaluation of database-derived pathway development for enabling biomarker discovery for hepatotoxicity. *Biomark Med*, 8(2):185-200.

Kamburov A, Stelzl U, Lehrach H, Herwig R (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res*, 41:D793-D800.

**Kamburov A**, Grossmann A, **Herwig R**, Stelzl U (2012) Cluster-based assessment of protein-protein interaction confidence. *BMC Bioinformatics*, 13:262.

**Kamburov A**, Stelzl U, **Herwig R** (2012) IntScore: a web tool for confidence scoring of biological interactions. *Nucleic Acids Res*, 40:W140-W146.