

Partnership for the Assessment of Risks from Chemicals

Additional Deliverable AD5.7

Draft strategy for the development of bioinformatics tools that are needed to link omics data and other mechanistic information to human relevant disease mechanisms

WP5 – T5.3



Technical reference	
Work package	WP5 – Hazard Assessment
Task	Task 5.3 - Quantitative systems toxicology and development of new AOPs
Dissemination level ¹	PU
Lead Beneficiary/ Responsible AE	UL-LACDR
Contributing Participants	SCIENSANO
Responsible author(s)	Bob van de Water/UL-LACDR/ water_b@lacdr.leidenuniv.nl
Co-authors	Giulia Callegaro/UL-LACDR/ g.callegaro@lacdr.leidenuniv.nl Imke Bruns/UL-LACDR/ i.b.bruns@lacdr.leidenuniv.nl Birgit Mertens/SCIENSANO/ Birgit.Mertens@sciensano.be
Reviewers	Albert Braeuning / BfR / albert.braeuning@bfr.bund.de John Colbourne / UoB/ j.k.colbourne@bham.ac.uk
Due date of deliverable	30/04/2023 – postponed 31/10/2023 ¹
Actual submission date	02/02/2024

¹ PU = Public

PP = Restricted to other programme participants (including the Commission Services)

RE = Restricted to a group specified by the consortium (including the Commission Services)

CO = Confidential, only for members of the consortium (including the Commission Services)

¹ The delay is a consequence of the "workshop on systems toxicology: data requirement and strategy" being moved from December to January. This workshop was essential for gathering information and executing the AD. Its purpose was to discuss relevant datasets and case studies.

Based on the requirements of three identified case studies, there was an additional search need in the public domain for relevant omics datasets

Document history

Version	Date	Reviewer name/Institutions	Short description of changes
version 1	24-11-2023	Giulia Callegaro	Started draft
version 2	5-12-2023	Imke Bruns	Included preservation and human data approach
Version 3	7-12-2023	Giulia Callegaro	Included AOP mapping approach, final version
Version 4	30-07-2024	Giulia Callegaro	Addressing reviewers comments

“Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.”

Abstract

This additional deliverable presents a strategic framework focusing on the development of bioinformatics tools crucial for linking omics data to human-relevant disease mechanisms. The proposed strategy, centered on Weighted Gene Co-expression Network Analysis (WGCNA) and Adverse Outcome Pathway (AOP) mapping, aims to systematically align in vitro and in vivo datasets. Specifically addressing challenges in translating omics data into meaningful insights for risk assessment, the strategy involves comparing human pathology and in vitro test system datasets while integrating AOP-based gene mapping. Preliminary results showcase the application of this approach to liver diseases, demonstrating the preservation of co-expression modules across different biological contexts. The ongoing extension of this strategy to other target organs emphasizes the importance of a multi-organ framework for comprehensive risk assessment in chemical safety.

Key Words

AOPs, toxicogenomics, gene co-expression analysis, system toxicology, preservation, AOP-Wiki

Table of contents

Document history	3
Abstract	4
Key Words	4
Table of contents	5
Authors and Acknowledgements	6
Acronyms	6
1. Introduction	7
2. Strategy	7
2.1 Requirements	8
2.1.1 Large transcriptomic data of human pathology and representative <i>in vitro</i> test systems	8
2.1.1.1 Transcriptomics data of human liver pathology	8
2.1.1.2 Transcriptomics data of PHH test system	9
2.1.2 FAIR knowledge bases of the AOP framework	9
3. Results	10
3.1 Pathology association and preservation analysis for liver diseases	10
3.1.1 Gene co-expression network generation	11
3.1.2 Disease-specific networks: fatty liver diseases case study	12
3.1.3 Preservation of subnetworks across levels of complexity	15
3.2. AOP mapping	17
3.3. Gap filling for other target organs: the kidney example	18
4. Discussion & future prospects	18
5. Conclusion	20
References	20
Supplemental data	21
Supplemental tables	21

Authors and Acknowledgements

Bob van de Water, Giulia Callegaro and Imke Bruns (Leiden University)

Birgit Mertens (Sciensano)

Acronyms

AO/ AON/ AOP: Adverse Outcome/ AO Network/ AO Pathway

EG: EigenGene scores

FAIR: Findable, Accessible, Interoperable, and Reusable

FFPE: formalin-fixed paraffin-embedded

HBV: hepatitis B virus

KEs: Key Events

KER: Key Event Relationship

NAFLD: non-alcoholic fatty liver disease

NASH: non-alcoholic steatohepatitis

NF- κ B: nuclear factor kappa-light-chain-enhancer of activated B cells

PCA: Principal Component Analysis

PHH: primary human hepatocyte

PRO: Protein Ontology

RDF: Resource Description Framework

RMA: Robust Multi-array Average

WGCNA: Weighted gene correlation analysis

1. Introduction

Omics data have proven highly informative in unraveling the intricate mechanisms underlying xenobiotic exposure (Beal et al., 2022; Harrill et al., 2021; Johnson et al., 2022; Krewski et al., 2019). However, translating these identified mechanisms into a meaningful understanding of their significance with respect to ultimate endpoints or human diseases remains a challenge. Questions arise regarding the interpretation of individual gene regulations, the relevance of altered pathways or functional ontologies, and how to navigate redundancy and terms defined in diverse fields, given the inherent bias in knowledge towards specific study areas. The mere knowledge of a gene being up or down-regulated raises fundamental questions about its impact on human health. Similarly, understanding the implications of perturbed pathways or functional ontologies, such as those associated with liver toxicity, demands a comprehensive approach. Dealing with redundancy in biological systems and terms borrowed from other disciplines further complicates the task of deriving meaningful insights from omics data. Moreover, the ambiguity persists in bridging the gap between in vitro transcriptomic responses and the complexities of human pathophysiology.

To address these challenges and establish a connection between in vitro models and human pathology, in the project P5.3.1.a_Y1_SystemsToxicology_UL-LACDR we propose a strategy employing Weighted Gene Co-expression Network Analysis (WGCNA)-based preservation analysis (Callegaro et al., 2023; Langfelder et al., 2011). This strategy aims to compare omics datasets representative of human pathology with datasets from in vitro models, providing a systematic approach to assess the relevance and confidence of omics data for risk assessment. The strategy not only involves comparing omics datasets but also integrates Adverse Outcome Pathway (AOP)-based gene mapping, offering a comprehensive framework to contextualize the molecular responses observed in in vitro models with human pathophysiology (Martens et al., 2022).

In essence, this drafted strategy endeavors to bridge the gap between the wealth of omics data and their application in risk assessment. By systematically aligning in vitro and in vivo datasets and incorporating AOP-based gene mapping, we aspire to enhance the usability and confidence in utilizing omics data for a more robust understanding of the potential health implications associated with xenobiotic exposure. This approach is a crucial step towards establishing a reliable framework for informed risk assessment in the realm of chemical safety.

2. Strategy

The strategy involves two main points:

1. Systems biology comparison of omics human and in vitro datasets

Systems biology methods offer a holistic approach to understanding complex biological systems and the effects of toxic substances on living organisms. In particular, gene co-expression models allow to identify groups of co-expressed genes, or modules, which are functionally related and may play a collective role in the onset of target organ adversities. Such models can be optimized for different test systems of increasing complexity, ranging from simple monocellular in vitro 2D models to whole organ in vivo responses. Interestingly, networks constructed for different systems can be mathematically compared to identify the subnetworks, representative of co-expressed genes functionally related, that share preserved patterns across test systems. We here propose to, starting from human in vivo data, 1) construct representative molecular networks on human in vivo

adversities 2) identify which subnetworks are associated with the occurrence of adversity in humans
3) identify which subnetworks are preserved in in vitro models.

2. Systematic mapping of AOPs with transcriptomics data

In parallel and complementary, we propose to start from a knowledge-rich source, the AOP-Wiki, which includes several molecular and biological descriptions of events underlying the occurrence of adverse events. Several efforts are ongoing to FAIRify the AOP-Wiki by facilitating its programmatic access and incorporating ontologies (Martens et al., 2022; Pittman et al., 2018). We here plan to further contribute to the translation of the AOPs into the gene space by systematically mapping the key events (KEs) with the above defined co-expressed modules. Co-expressed modules are inherently associated with processes and have a high likelihood to represent complex KEs rather than individual genes. In addition, co-expression modules derived from methods such as WGCNA are easily quantifiable by the eigengene score (EGs), mathematically the first principal component of the expression matrix of the corresponding module, representing a summarized score of the entire group of genes, therefore providing a ready-to-use method to estimate the activity of KEs within an AOP.

For both strategies, we here present some preliminary results, and an outlook on the next steps.

In summary, this approach will establish a connection between the AOP framework and the transcriptomic and co-expression spaces to support a consensus, causally directed interpretation of toxicogenomics.

2.1 Requirements

There are several crucial requirements to achieve the objectives defined above: large transcriptomic data of human pathology and of representative in vitro test systems, and FAIR knowledge bases for the AOP framework.

2.1.1 Large transcriptomic data of human pathology and representative *in vitro* test systems

2.1.1.1 Transcriptomics data of human liver pathology

As proof of concept, we obtained microarray-derived human liver pathology samples from the BioStudies database (Supplemental Table 1). As a prerequisite, studies were only selected if the batch contained at least one sample without a pathology (normal sample), which can later be used to determine log₂ fold change values of differential gene expression. Additionally, only adult human whole liver samples were selected, and the abundance of disease classes was averaged to prevent an overrepresentation of one specific pathology. Lastly, all hepatic carcinoma samples and all treatment conditions were excluded from the data due to their dissimilarity from the other pathologies. Ultimately, we collected 145 microarray samples from different studies, which comprised 14 different liver-related pathologies. The most prevalent diseases in this cohort were advanced non-alcoholic fatty liver disease (n=32), hepatitis B virus (HBV) infection (n=30), and HBV-associated acute liver failure (n=17) (Figure 1). All samples were Robust Multi-array Average (RMA) normalized in separate batches. Additionally, batch correction was achieved within the DESeq design (controls vs individual patients) using the DESeq2 package.

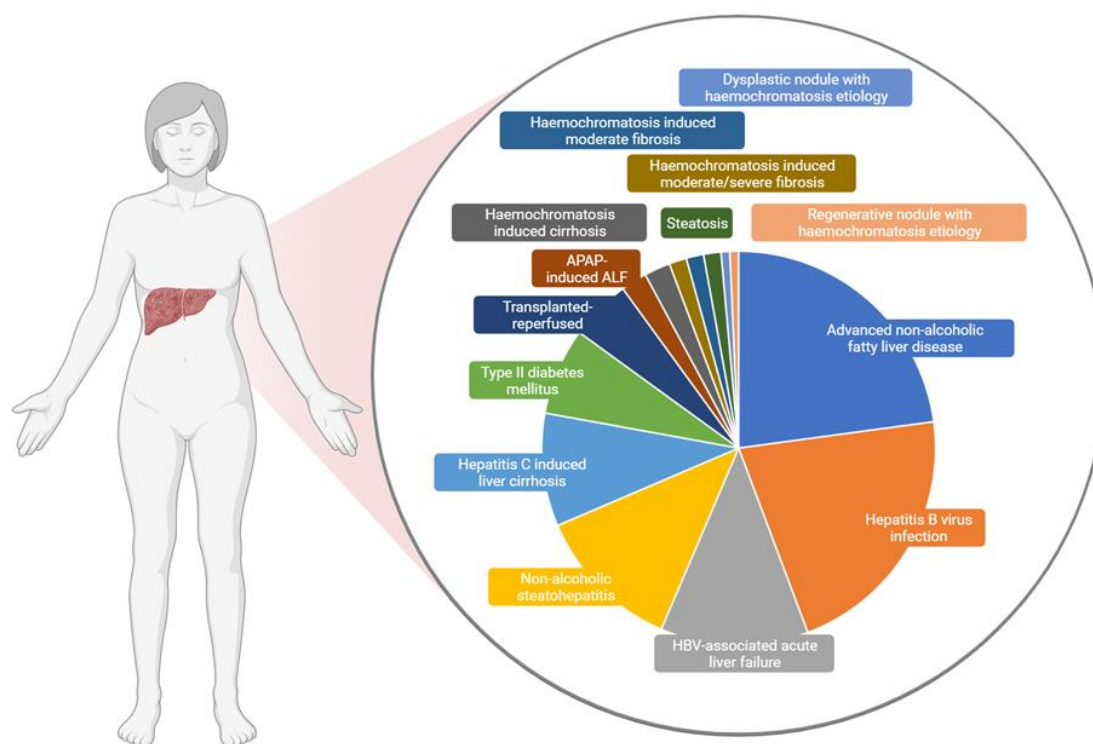


Figure 1: Included GEO samples. Gene expression data of 145 human liver biopsy samples were collected from different studies in the GEO database. These studies comprised 14 different pathologies. The most prevalent diseases in this cohort were advanced non-alcoholic fatty liver disease (n=32), hepatitis B virus infection (n=30) and HBV-associated acute liver failure (n=17).

2.1.1.2 Transcriptomics data of PHH test system

Our objective is to systematically align human in vivo data with in vitro datasets, aiming to improve the usability and confidence in employing omics data in risk assessment. For this purpose, we employ transcriptomics data derived from a primary human hepatocyte (PHH) cell line obtained from the TG-GATEs repository. This dataset contains 941 PHH treatments, where each treatment is defined as a combination of a compound, concentration and time. Callegaro et al. previously described this dataset and used it to perform a WGCNA analysis (Callegaro et al., 2023). As a result, these data are well-suited for directly comparing the modules formed in this in vitro model to those in the human liver pathology-based in vivo model. This enables us to provide an initial assessment of its human relevance.

2.1.2 FAIR knowledge bases of the AOP framework

The AOP Wiki is an online, collaborative platform designed to facilitate the sharing and development of Adverse Outcome Pathways (AOPs). However, for the purpose defined above, a programmatic access is required. Two resources are continuously developed in this regard and ready to be used in this proposed framework. The AOP-Wiki database is a comprehensive platform that compiles and organizes Adverse Outcome Pathway (AOP) information, providing a structured repository for the collective knowledge of biological pathways leading to adverse effects (Pittman et al., 2018). On the other hand, the AOP-Wiki RDF (Resource Description Framework) extends this functionality by representing AOP information in a machine-readable format, facilitating interoperability and integration with other databases and tools in the domain of toxicology and risk assessment (Martens et al., 2022).

3. Results

3.1 Pathology association and preservation analysis for liver diseases

An illustration of the overview of the first strategy point (Systems biology comparison of omics human and in vitro datasets) is provided in Figure 2 and exemplified further in the following subsection for liver diseases. First, a large dataset of transcriptomic samples of patients representing various liver disease is collected, in order to define the landscape of molecular pathways triggered in liver diseases. From this dataset, WGCNA is applied to derive co-expression networks representing individual (sub)pathways, that are differentially activated or repressed in each different liver disease. This specific gene co-expression network analysis is applied to transcriptomic data in order to cluster co-expressed set in a completely data-driven approach and not dependent in the first instance on the pathological classes (unsupervised). Then, in order to identify the molecular signature of each disease, statistical modelling is performed to associate the modules' score to a particular disease group. Lastly, human liver disease co-expression modules are evaluated for their preservation towards an in vitro hepatic model, PHH: networks topology can be quantified and compared with a different dataset, and the overlap between the modules in the two different systems measured to identify common modules and nodes. Network topology is evaluated with a combined score (Z-summary) including density and connectivity patterns of the genes in each given network (Langfelder et al., 2011). More details in the section 3.1.3 Preservation of subnetworks across levels of complexity.

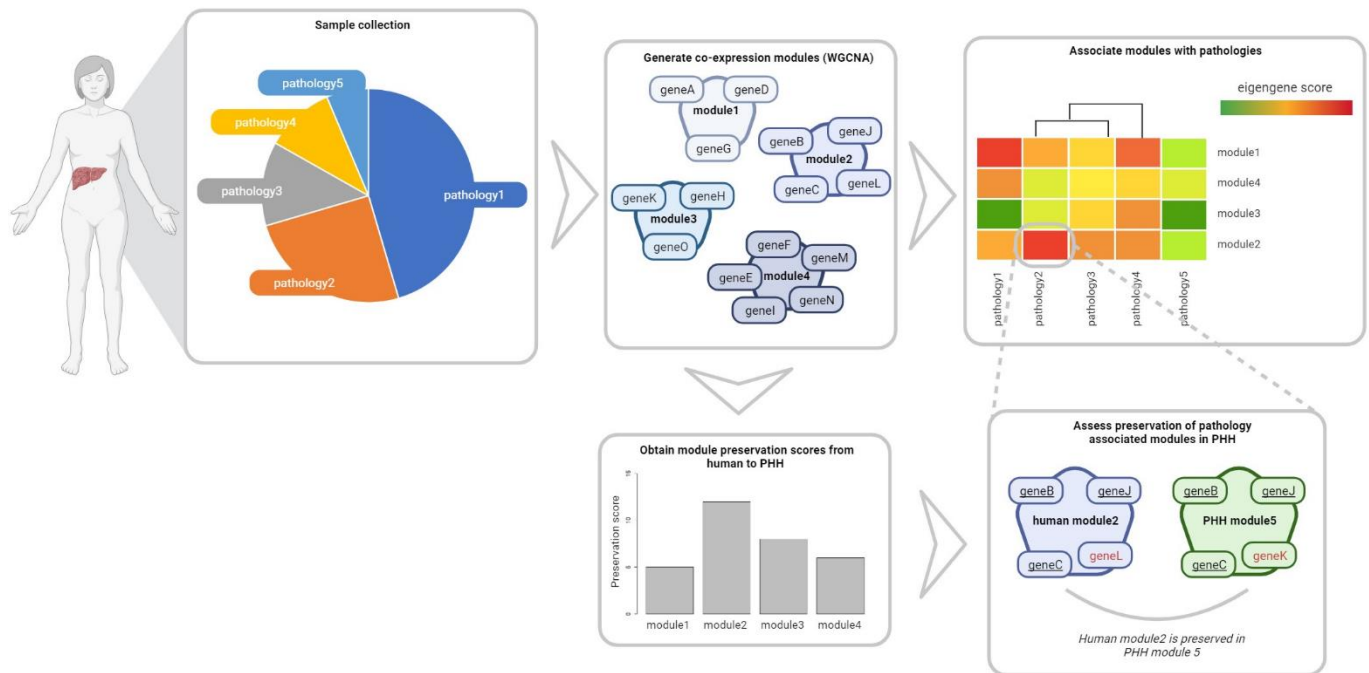


Figure 2: Overview of the suggested workflow. First, the collected human liver pathology samples are analyzed to establish co-regulated gene networks. For each sample, the eigengene score (EGs) is then calculated to summarize the collective log₂ fold change its constituent genes. These modules scores are then used to quantify the activation or repression induced by a given pathology and give an indication of which of the established co-expression networks is underlying the liver pathologies (traits) represented in the analyzed cohorts. Lastly, the preservation of the established gene co-expression modules will be evaluated in the PHH cells.

3.1.1 Gene co-expression network generation

The human liver pathology transcriptomics data were analyzed to establish co-regulated gene networks by performing a weighted gene co-expression network analysis (WGCNA). We created unsigned (absolute value of Pearson correlation score) gene modules, meaning that both co-induced and co-repressed genes are grouped together, and selected the optimal soft-power parameter. The criteria used to determine the most suitable soft-power value involved evaluating networks based on two key considerations: 1) scale free topology: a scale-free network is characterized by a few highly connected nodes and many nodes with lower connectivity, which is often observed in biological systems. We therefore sought a soft-power value that would promote the emergence of a scale-free structure in the network. 2) mean connectivity: while striving for a scale-free topology, it is important to maintain a reasonable level of mean connectivity. Excessive connectivity may lead to a fully connected or overly dense network, which can obscure meaningful biological relationships. Conversely, insufficient connectivity might result in a network that lacks integration. The soft-power value should therefore strike a balance to ensure a network with appropriate mean connectivity levels that reflect the biological complexity being studied. To obtain an optimal balance between scale-free topology and preservation of connectivity patterns, we systematically evaluated various soft-power values. After exploration, a soft-power threshold of 10 was identified as most suitable. Additionally, we set the minimal module size on 5 genes, to make sure to obtain biologically relevant networks. We based this choice on the observation that annotated pathways in well-established databases such as WikiPathways, Reactome, and Hallmarks Pathways typically consist of gene sets that surpass the threshold of five genes. By aligning our analysis parameters with this observation, we aimed to focus on constructing biologically meaningful modules that are more likely to correspond to functional units within cellular processes. With these settings, we ultimately obtained a total of 231 modules, Enrichment of the modules was examined using GO-term and pathway annotation databases including BioCarta, KEGG, WikiPathways and Reactome in the enrichR package.

For each sample, we computed the eigengene score (EGs) to summarize the collective log₂ fold change magnitudes of its constituent genes. Briefly, this involves the application of a Principal Component Analysis (PCA) on the log₂ fold change gene matrix associated with each module, where the first principal component then corresponds to the EGs. These module scores are then used to get an indication of the level of activation or repression induced by a given pathology and give an indication of which of the established co-expression networks is underlying the liver pathologies (traits) represented in the analyzed cohorts. Here, we focus fatty liver diseases including three subgroups: I.e. non-alcoholic fatty liver disease (NAFLD), non-alcoholic steatohepatitis (NASH) and steatosis.

3.1.2 Disease-specific networks: fatty liver diseases case study

Logistic regression was used to model the statistical association between module and trait by including the module as a predictor variable in the model. The model's coefficients were then used to determine the strength and direction of the module-trait correlation. Logistic regression was solely applied to pathologies with a sample size of at least five, leading to the exclusion of the subgroup steatosis. This threshold was employed to ensure a sufficient representation within each pathology, enhancing statistical robustness and reliability of the analysis (Figure 3A). Additionally, the Cohen's D effect size was calculated. This method provides a quantification of the magnitude of the module's influence on the trait, by expressing the module's impact in terms of effect size (Figure 3B). We have ranked the modules on their module/trait association based on the adjusted p-value obtained from the logistic regression model. The top 10 ranked modules per pathology are shown in Table 1.

The findings from this analysis reveal an association between numerous networks linked to immune responses and all different forms of steatosis pathologies (HL_GEO_82, HL_GEO_8, HL_GEO_220, HL_GEO_115). It has been described that the innate immune system is involved in steatosis by activating the resident Kupffer cells through recruitment of immune cells including neutrophils (HL_GEO_8 and HL_GEO_115), monocytes and natural killer cells to the liver (Moayedfarid et al., 2022). Furthermore, there emerges a distinct downregulation of module HL_GEO_217, which is associated with fatty acid transport. Fatty liver diseases can result from an imbalance between the uptake, synthesis, utilization and export of fatty acids. If the activity of fatty acid transporters is downregulated, it can lead to reduced uptake of fatty acids into hepatocytes. This can result in an impaired ability of the liver to process and utilize the incoming fatty acids. As a consequence, the fatty acids that are not effectively used may accumulate within the liver cells, contributing to the development of fatty liver diseases (Li et al., 2022). Additionally, module HL_GEO_161, enriched with genes involved in cholesterol homeostasis shows a positive association with non-alcoholic steatohepatitis. Previous studies found that an increased accumulation of free cholesterol in the liver can lead to toxic effects eventually contributing to the development of fatty liver diseases (Malhotra et al., 2020).

Table 1: Strongest module/trait associations for pathologies associated with steatosis (i.e. non-alcoholic steatohepatitis, advanced non-alcoholic fatty liver disease and steatosis). Only pathologies encompassing more than 5 samples were subjected to logistic regression, leading to the exclusion of steatosis (n=3). The Cohen's D effect size has been included in this table, but should be interpreted with caution.

Phenotype	Module	Cohen's D	Log reg	p _{adj}	Annotation	Transcription factor
Non-alcoholic steatohepatitis	HL_GEO_5	-1.38	-7.42	1.89e-9	Extracellular matrix organization	ETS1, FOXM1, NFIC, NFKB1, RFX5, SMAD3, SPI1, TP53, TWIST1
Non-alcoholic steatohepatitis	HL_GEO_82	-1.36	-2.11	9.82e-9	Cellular response to interleukin-7	STAT6
Non-alcoholic steatohepatitis	HL_GEO_59	-1.36	-4.54	2.33e-8	ROS degradation	
Non-alcoholic steatohepatitis	HL_GEO_143	-1.50	-1.74	3.28e-8	Regulation of MAP kinase pathways	ATF2
Non-alcoholic steatohepatitis	HL_GEO_197	-1.25	-4.79	5.57e-8	Glycolysis	
Non-alcoholic steatohepatitis	HL_GEO_217	-1.30	-3.16	8.44e-8	Fatty acid transporter activity	
Non-alcoholic steatohepatitis	HL_GEO_158	-0.919	-2.80	9.26e-8		AR, HIF1A, LEF1, TCF7L2
Non-alcoholic steatohepatitis	HL_GEO_161	1.58	1.84	1.10e-7	Cholesterol biosynthesis	SREBF1, YY1
Non-alcoholic steatohepatitis	HL_GEO_67	-1.42	-3.61	6.28e-7	Organic anion/cation transport	
Non-alcoholic steatohepatitis	HL_GEO_229	-1.45	-1.93	1.03e-6	RNA transcription	
Advanced non-alcoholic fatty liver disease	HL_GEO_163	-0.708	-2.50	2.15e-8		
Advanced non-alcoholic fatty liver disease	HL_GEO_56	-0.726	-2.42	2.33e-7	Sialic acid metabolism	
Advanced non-alcoholic fatty liver disease	HL_GEO_8	-0.984	-3.61	2.48e-7	Neutrophil activation	ERG, NFKB1, RELA, RFX5, RUNX1, SP1, SPI1, TP53
Advanced non-alcoholic fatty liver disease	HL_GEO_117	-0.520	-1.72	3.79e-7		STAT6
Advanced non-alcoholic fatty liver disease	HL_GEO_189	-0.841	-2.30	5.97e-7		
Advanced non-alcoholic fatty liver disease	HL_GEO_182	0.679	1.78	1.88e-6		
Advanced non-alcoholic fatty liver disease	HL_GEO_228	1.06	1.75	4.62e-6		
Advanced non-alcoholic fatty liver disease	HL_GEO_135	-0.522	-1.47	7.84e-6		
Advanced non-alcoholic fatty liver disease	HL_GEO_155	0.980	1.45	1.02e-5		
Advanced non-alcoholic fatty liver disease	HL_GEO_32	0.919	1.03	1.07e-5		
Steatosis	HL_GEO_101	2.95	N/A	N/A		ESR1, FOSL2, JUND, SRF
Steatosis	HL_GEO_165	2.93	N/A	N/A	Specific granule membrane	
Steatosis	HL_GEO_220	2.73	N/A	N/A	NFkB signaling	E2F1
Steatosis	HL_GEO_191	2.30	N/A	N/A		SREBF2
Steatosis	HL_GEO_115	2.27	N/A	N/A	Neutrophil activation	CEBPA, SPI1
Steatosis	HL_GEO_150	2.01	N/A	N/A		ARNTL, EPAS1, VDR
Steatosis	HL_GEO_159	1.79	N/A	N/A	Response to metal ions	
Steatosis	HL_GEO_230	-1.79	N/A	N/A		
Steatosis	HL_GEO_148	1.72	N/A	N/A	Transcriptional activator activity	
Steatosis	HL_GEO_171	1.69	N/A	N/A		

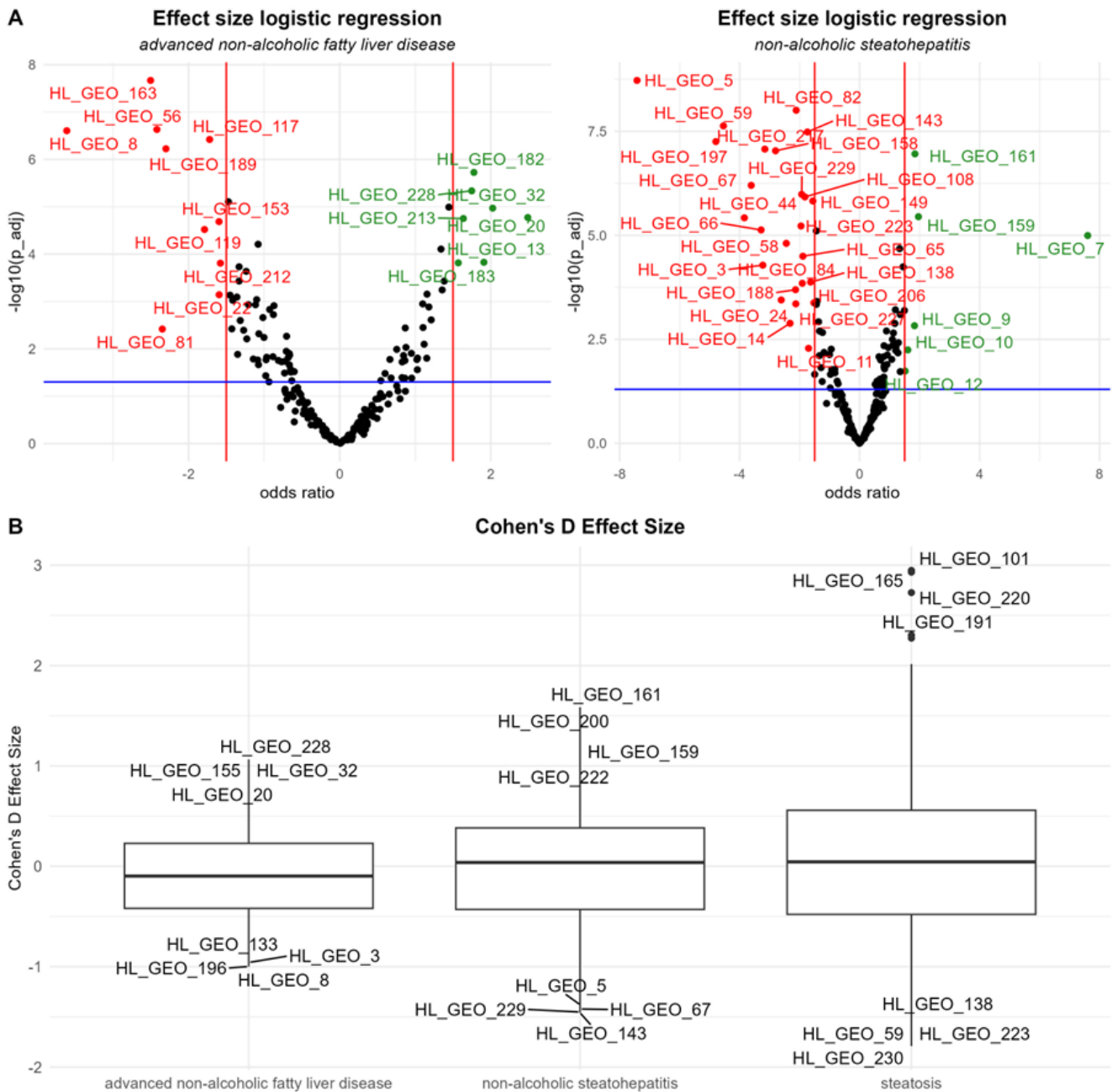


Figure 3: Logistic regression and Cohen's D effect size results. Logistic regression was solely calculated for pathologies that encompassed a minimum of 5 samples, thus excluding steatosis due to its sample size falling below this threshold. The modules shown in green signify a positive association with their corresponding pathology (trait), while those in red have a negative module/trait association (A). Within the boxplots illustrating Cohen's D effect sizes, the 4 modules with the strongest positive and negative module/trait association have been labeled (B).

3.1.3 Preservation of subnetworks across levels of complexity

After generating the gene co-expression networks based on human liver samples, we evaluated their preservation across distinct biological contexts. Here, we focused on the assessment of module preservation in primary human hepatocytes (PHH). The concept of module preservation encapsulates the extent to which gene co-expression patterns remain consistent across different systems (I.e. human in vivo and human in vitro). This analysis provides insights into the robustness of the modules, and sheds light on the biological human relevance of the in vitro models used for preclinical safety assessment. The modules derived from human liver data were scrutinized for their preservation in PHH cells to elucidate the degree to which the co-expression networks established in the human in vivo context persisted in this cellular test system.

To perform a preservation analysis, we used the modulePreservation function from the WGCNA package. This function calculates the module preservation statistics between independent data sets and yields a Z-summary. A Z-summary compares the observed preservation of a module in the validation datasets to its preservation under random circumstances.

The Z-summary offers a quantitative measure that indicates the degree of module preservation across different biological systems. A Z-summary of >2 signifies a moderate preservation of a module, whereas a Z-summary of >10 indicates a high preservation. In the context of the PHH cell line, approximately 48% of the formed modules within the GEO human liver dataset are moderately preserved in PHH cells, while a subset of $\sim 6\%$ demonstrates a high degree of preservation (Figure 4 and Table 2). These findings indicate that a substantial number of modules found in the human setting can also be found in PHH cells. Additional analyses are required to gain insights into the translation of human gene co-expression patterns in other human in vitro models.

WGCNA Preservation Statistics
GEO human liver to PHH

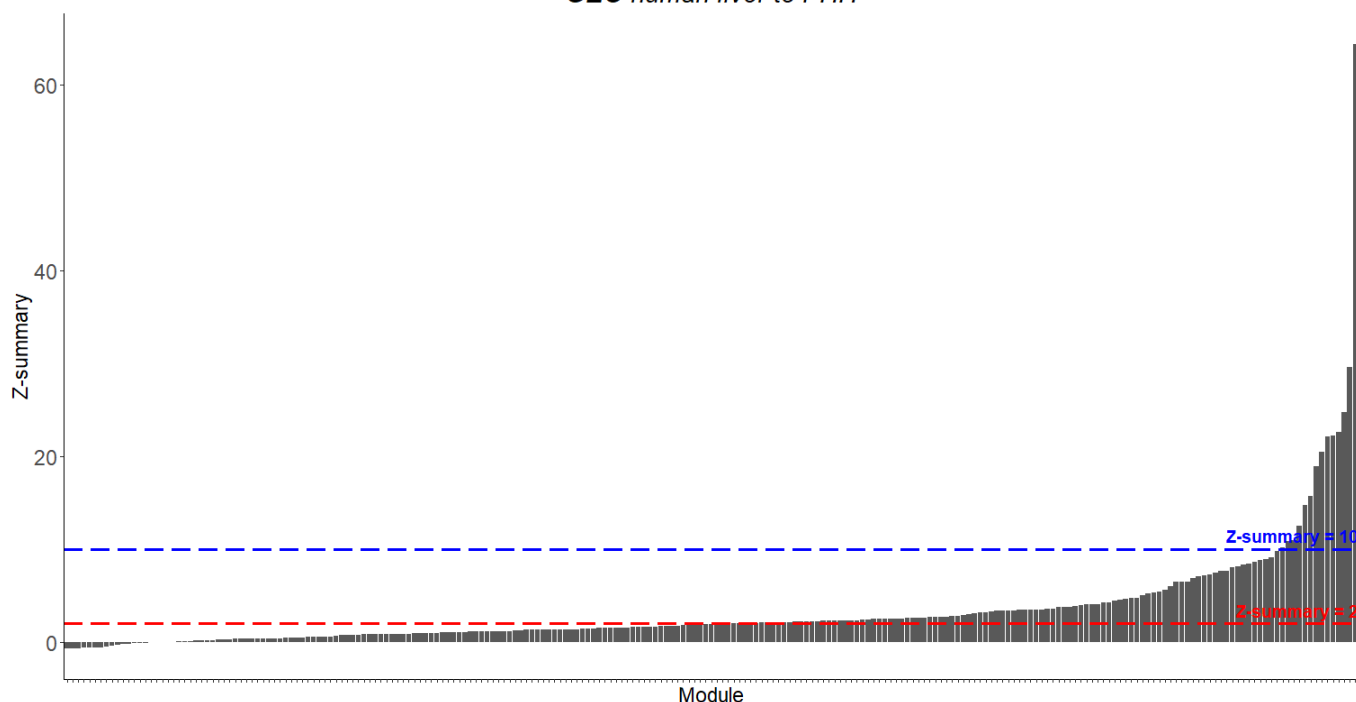


Figure 4: Z-summary of WGCNA module preservation. A Z-summary of >2 , indicated by the red dotted line, signifies a moderate preservation of a module, whereas a Z-summary of >10 , illustrated by the blue dotted line, signifies a high preservation. In the PHH cell line, $\sim 48\%$ of the formed modules in the GEO human liver data are moderately preserved, whereas $\sim 6\%$ exhibit a high degree of preservation.

Subsequently, we specifically investigated the preservation of the modules that we found to be associated with fatty liver diseases (I.e. NAFLD, NASH, and steatosis) (Table 2). We found that module HL_GEO_143, HL_GEO_161, HL_GEO_67, HL_GEO_56, HL_GEO_165, HL_GEO_220, HL_GEO_159, HL_GEO_230, and HL_GEO_148 are moderately preserved in PHH. These modules are associated with the regulation of MAP kinase pathways, cholesterol biosynthesis, organic anion/cation transport, sialic acid metabolism, specific granule membrane, NF- κ B signaling, response to metal ions and transcriptional activator activity, respectively.

Table 2: Ranked preservation statistics of modules associated with fatty liver diseases. A Z-summary of >2 indicates a moderate preservation of a module, whereas a Z-summary of >10 signifies a high preservation. The preservation of the modules generated on the GEO human liver data were tested in PHH cells. Module HL_GEO_143, HL_GEO_161, HL_GEO_67, HL_GEO_56, HL_GEO_32, HL_GEO_165, HL_GEO_220, HL_GEO_159, HL_GEO_230, and HL_GEO_148 are moderately preserved in PHH.

Pathology	Module	Annotation	Z-summary tested towards PHH
Non-alcoholic steatohepatitis	HL_GEO_161	Cholesterol biosynthesis	8.8
Steatosis	HL_GEO_159	Response to metal ions	7.52
Advanced non-alcoholic fatty liver disease	HL_GEO_32		6.48
Steatosis	HL_GEO_230		5.41
Non-alcoholic steatohepatitis	HL_GEO_143	Regulation of MAP kinase pathways	4.29
Advanced non-alcoholic fatty liver disease	HL_GEO_56	Sialic acid metabolism	2.54
Non-alcoholic steatohepatitis	HL_GEO_67	Organic anion/cation transport	2.25
Steatosis	HL_GEO_148	Transcriptional activator activity	2.22
Steatosis	HL_GEO_165	Specific granule membrane	2.09
Steatosis	HL_GEO_220	NF κ B signaling	2.03
Steatosis	HL_GEO_101		1.96
Advanced non-alcoholic fatty liver disease	HL_GEO_135		1.82
Advanced non-alcoholic fatty liver disease	HL_GEO_189		1.71
Steatosis	HL_GEO_150		1.58
Non-alcoholic steatohepatitis	HL_GEO_229	RNA transcription	1.53
Steatosis	HL_GEO_115	Neutrophil activation	1.39
Advanced non-alcoholic fatty liver disease	HL_GEO_117		1
Steatosis	HL_GEO_171		0.822
Advanced non-alcoholic fatty liver disease	HL_GEO_155		0.768
Non-alcoholic steatohepatitis	HL_GEO_158		0.55
Non-alcoholic steatohepatitis	HL_GEO_217	Fatty acid transporter activity	0.543
Advanced non-alcoholic fatty liver disease	HL_GEO_182		0.376
Non-alcoholic steatohepatitis	HL_GEO_197	Glycolysis	0.363
Advanced non-alcoholic fatty liver disease	HL_GEO_228		0.163
Non-alcoholic steatohepatitis	HL_GEO_82	Cellular response to interleukin-7	-0.106
Steatosis	HL_GEO_191		-0.136
Advanced non-alcoholic fatty liver disease	HL_GEO_163		-0.322
Non-alcoholic steatohepatitis	HL_GEO_59	ROS degradation	-0.585
Non-alcoholic steatohepatitis	HL_GEO_5	Extracellular matrix organization	-0.641
Advanced non-alcoholic fatty liver disease	HL_GEO_8	Neutrophil activation	-0.721

3.2. AOP mapping

As a first proof of concept, we evaluated the overlap between the gene space of AOP-Wiki as provided by the AOP-Wiki RDF (Martens et al., 2022), and the PHH co-expression modules defined with the dataset explained in section 2.1.1.2.

In the XML-to-RDF conversion process of the AOP-Wiki, two distinct methods are employed for mapping gene and protein identifiers. The first method relies on existing Biological Object annotations utilizing Protein Ontology (PRO) terms in the AOP-Wiki, then converted to NCBI Gene and UniProt, as well as symbols from HGNC, using the relevant PR mapping file. The second method involves textual gene identifier mapping for genes in Key Events (KEs) and Key Event Relationships (KERs): the text string of the genes HGNC identifiers have been matched in the available text describing the KEs and KERs, leading to a possible identification of genes involved.

All genes obtained via the first method suffer from a mapping issue, related to the use of an outdated mapping file between the PR terms and the other gene identifiers, resulting in genes not correctly matched; therefore, the first methods has not yet been included in this preliminary analysis. We therefore focused on the second mapping method, extracting all genes from all represented KEs and AOPs, then matched the list with the gene membership of the PHH co-expression model. Several AOPs, KEs and PHH co-expression modules are represented (Figure 5A). Modules are matching with KE for up to 40% of their members, with smaller modules having higher likelihood to overlap with a higher percentage (Figure 5B). When testing for significant overlap (hypergeometric test with Benjamini-Hochberg multiple test correction), ~60% of the pairs are found significant (Figure 5C). Interestingly, several AOPs have more than one KE significantly overlapping with a PHH co-expression module, indicating the concrete possibility to describe and quantify a considerable portion of an AOP with interpretable gene expression data (Figure 5D).

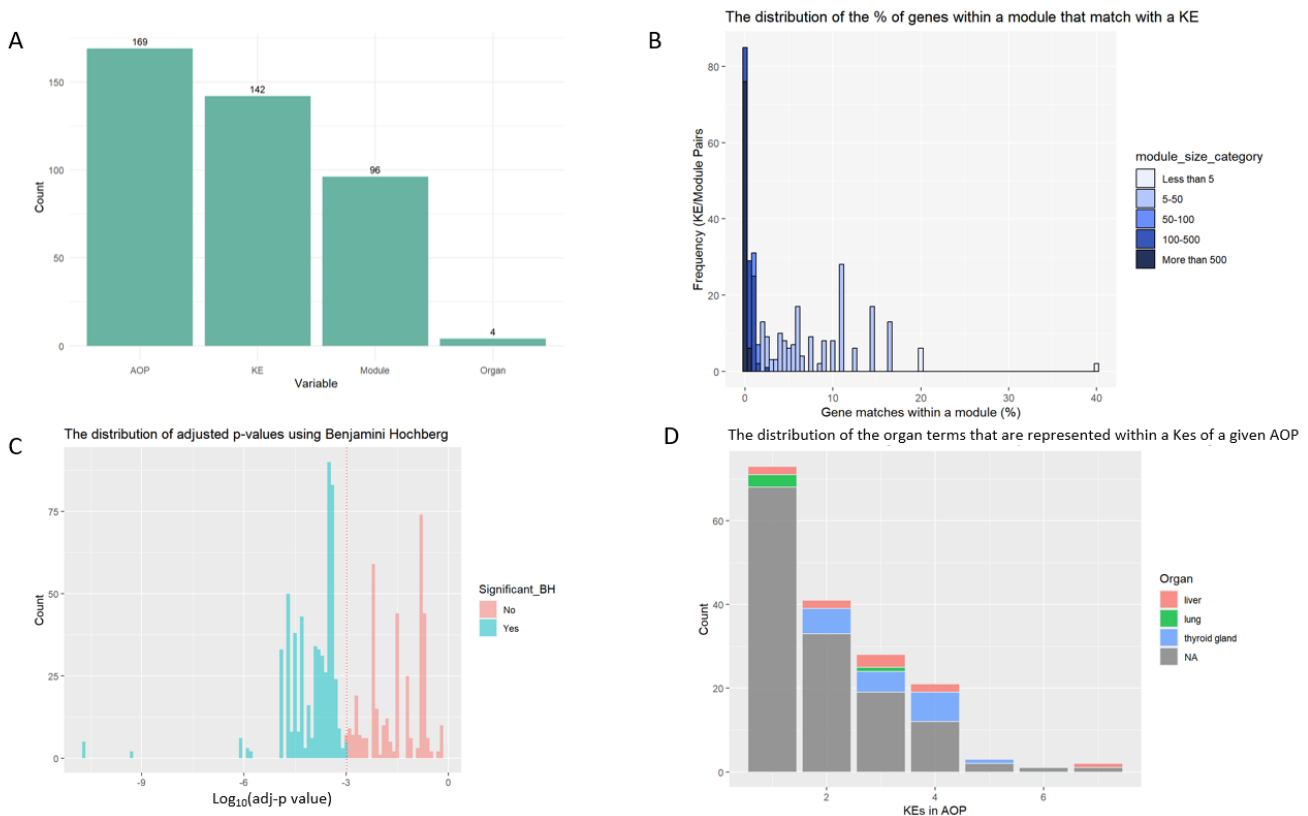


Figure 5: Preliminary results on the mapping between the AOP-Wiki RDF and the PHH gene co-expression modules. A) statistics of the resulting dataset after gene matching between the two datasets. B) Frequency of module-KE pairs overlap expressed in percentage, stratified by the modules' size. C) The distribution of the adjusted p-values of the overlap between KEs and modules. The dashed line correspond to adj p-value = 0.001 D) The distribution of the significant KE-modules pairs per AOP, stratified by the organ terms, when available.

3.3. Gap filling for other target organs: the kidney example

We are currently following a similar strategy to elucidate genes' role in kidney pathophysiology. To start, we have collected a comprehensive set of 678 human kidney allograft biopsy specimens sourced from biobanks affiliated with Dutch academic hospitals, specifically the Leiden University Medical Centre (LUMC) and the Academic Medical Center Amsterdam (AMC). Collected between 2017 and 2022, all biopsy samples were initially formalin-fixed paraffin-embedded (FFPE) before preservation in the biobank. To acquire transcriptomic data, 10 μ m slices were extracted from the FFPE tissue and subjected to high-throughput targeted RNA sequencing using TempO-Seq technology. The collection represents a variety of kidney pathologies, as indicated in Figure 6.

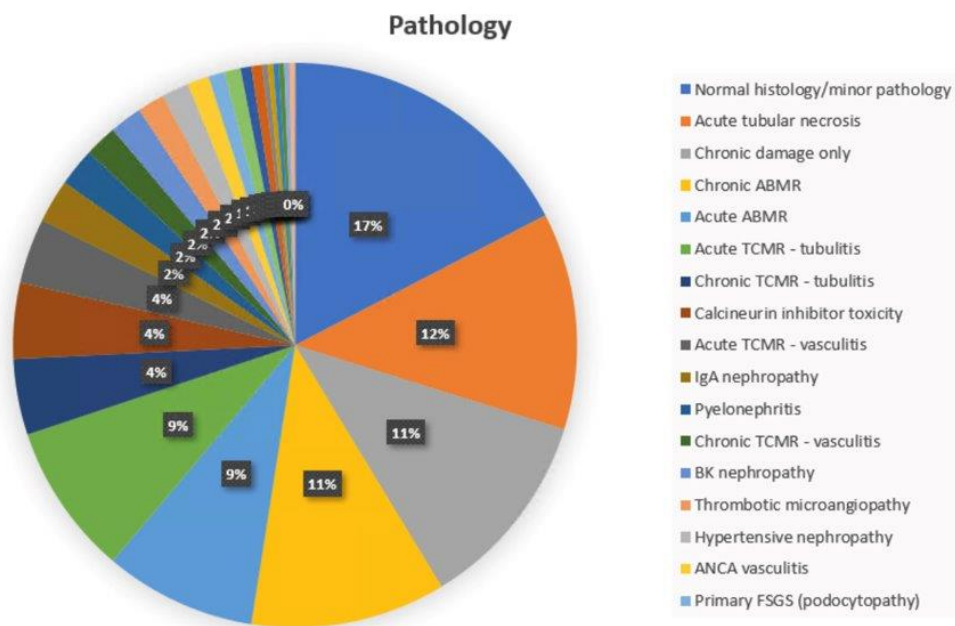


Figure 6: Distribution of kidney pathologies in the collected samples for TempOseq analysis. ABMR: Antibody-Mediated Rejection; TCMR: T-Cell Mediated Rejection; IgA: immunoglobulin A; BK: BK polyomavirus; ANCA: Anti-neutrophil cytoplasmic autoantibody; FSGS: Focal segmental glomerulosclerosis

We are currently analyzing the data and constructing the WGCNA model.

Simultaneously, we have generated a dataset of RPTEC-TERT1 cells exposed to 43 compounds including nephrotoxics, growth factors and adaptive stress pathway activators at multiple time points and concentrations to obtain the relative *in vitro* WGCNA model.

Next, we are going to determine which subnetworks are associated with diseases phenotype, and identify which subnetworks are preserved in the *in vitro* model.

4. Discussion & future prospects

We here outlined a strategy to bridge the gap between omics data, particularly transcriptomics, and their application in risk assessment, with a focus on understanding the implications of xenobiotic exposure. The proposed strategy involves two main components: systems biology comparison of

omics human and in vitro datasets and systematic mapping of Adverse Outcome Pathways (AOPs) with transcriptomics data.

Systems biology approaches aim to construct representative molecular networks for human in vivo adversities, identify associated subnetworks, and assess their preservation in in vitro models. The co-expression networks, generated through Weighted Gene Co-expression Network Analysis (WGCNA), allow the identification of functionally related gene modules. The preservation analysis evaluates the consistency of these modules across different biological contexts, providing insights into the relevance of in vitro models for preclinical safety assessment. Preliminary results show the moderate to high preservation of certain modules associated with liver diseases in primary human hepatocytes (PHH) cells.

The second component involves mapping AOPs with transcriptomics data, utilizing the AOP-Wiki database and its RDF representation. The mapping aims to associate Key Events (KEs) from AOPs with the co-expression modules defined in the transcriptomics data. Preliminary results demonstrate significant overlap between KEs and modules, supporting the feasibility of describing a considerable portion of an AOP with interpretable gene expression data.

The current proposed framework focuses on transcriptomics data, being the most mature and information rich omics technology available at the moment. However, other omics layers, such as metabolomics and proteomics could provide unique information not represented by RNA abundancies, and would be particularly suited to define accessible and stable biomarkers associated with human pathologies.

Finally, it is critical to extend this approach to all the target organs that are relevant to risk assessment in various domains. Ideally, one would need to collect transcriptomic data from human patient cohorts representative of several target organs (e.g. liver, kidney, lung and neuronal systems) and their corresponding in vitro models to allow for translation of gene association in an in vitro setting. Simultaneously, the AOP mapping would allow to anchor the gene co-expression modules to the recognized framework of description of molecular events in toxicology, allowing a qualitative and quantitative assessment of the potential adverse effects (**Figure 7**).

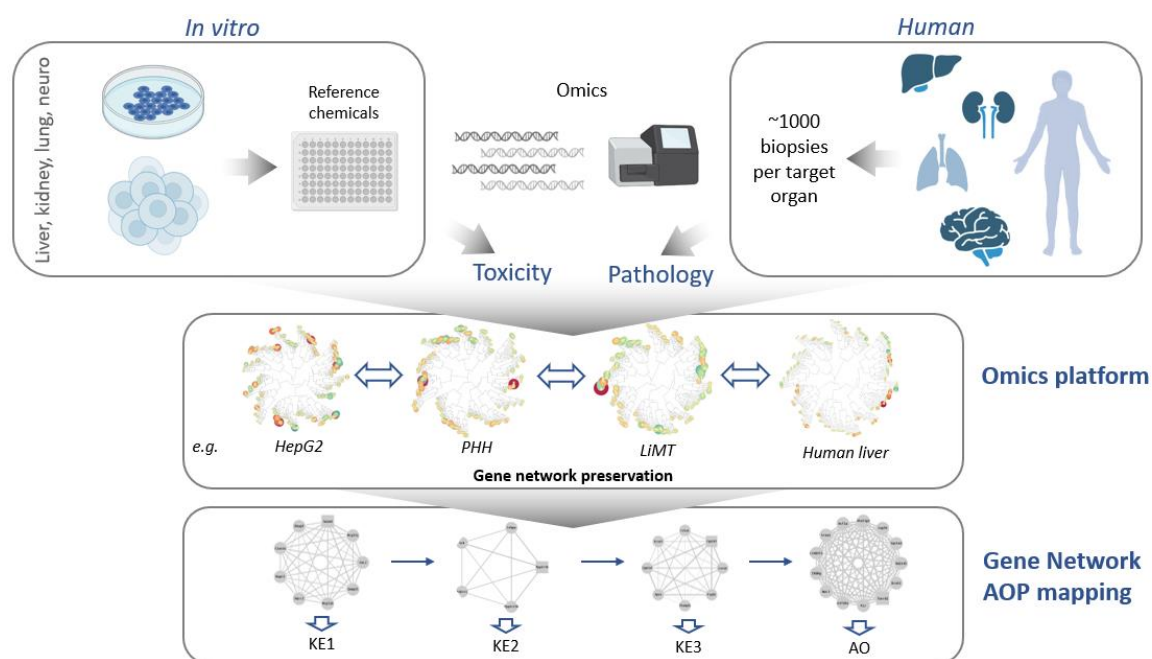


Figure 7: Envisioned strategy. HepG2: hepatocellular carcinoma cell line; PHH: primary human hepatocytes; LiMT: liver micro tissues

5. Conclusion

In this deliverable, we have presented a comprehensive strategy to bridge the gap between omics data and risk assessment in chemical safety. Our proposed approach involves the systematic comparison of human and in vitro datasets through systems biology methods and the mapping of Adverse Outcome Pathways (AOPs) with transcriptomics data. By employing Weighted Gene Co-expression Network Analysis (WGCNA) and AOP-Wiki, we aim to enhance the usability and confidence in utilizing omics data for a more robust understanding of the potential health implications associated with xenobiotic exposure. This integrated strategy is a crucial step forward in establishing a reliable framework for informed risk assessment. As we outline our strategy for liver diseases, we recognize the need to extend this approach to all target organs relevant to risk assessment. Our vision includes collecting transcriptomic data from human patients representing various target organs and their corresponding in vitro models. Simultaneously, systematic AOP mapping would anchor gene co-expression modules to the recognized framework of molecular events in toxicology. This holistic vision is essential for a qualitative and quantitative assessment of potential adverse effects across multiple organ systems, ultimately advancing the field of chemical safety assessment.

References

- Beal, M. A., Gagne, M., Kulkarni, S. A., Patlewicz, G., Thomas, R. S., & Barton-Maclaren, T. S. (2022). Implementing *In Vitro* Bioactivity Data to Modernize Priority Setting of Chemical Inventories. *Altex*, 39(1), 123–139. <https://doi.org/10.14573/ALTEX.2106171>
- Callegaro, G., Schimming, J. P., Piñero González, J., Kunnen, S. J., Wijaya, L., Trairatphisan, P., van den Berk, L., Beetsma, K., Furlong, L. I., Sutherland, J. J., Mollon, J., Stevens, J. L., & van de Water, B. (2023). Identifying multiscale translational safety biomarkers using a network-based systems approach. *Science*, 26(3), 106094. <https://doi.org/10.1016/J.ISCI.2023.106094>
- Harrill, J. A., Viant, M. R., Yauk, C. L., Sachana, M., Gant, T. W., Auerbach, S. S., Beger, R. D., Bouhifd, M., O'Brien, J., Burgoon, L., Caiment, F., Carpi, D., Chen, T., Chorley, B. N., Colbourne, J., Corvi, R., Debrauwer, L., O'Donovan, C., Ebbels, T. M. D., ... Whelan, M. (2021). Progress towards an OECD reporting framework for transcriptomics and metabolomics in regulatory toxicology. *Regulatory Toxicology and Pharmacology*, 125, 105020. <https://doi.org/10.1016/J.YRTPH.2021.105020>
- Johnson, K. J., Auerbach, S. S., Stevens, T., Barton-Maclaren, T. S., Costa, E., Currie, R. A., Dalmas Wilk, D., Haq, S., Rager, J. E., Reardon, A. J. F., Wehmas, L., Williams, A., O'Brien, J., Yauk, C., Larocca, J. L., & Pettit, S. (2022). A Transformative Vision for an Omics-Based Regulatory Chemical Testing Paradigm. *Toxicological Sciences*, 190(2), 127–132. <https://doi.org/10.1093/TOXSCI/KFAC097>
- Krewski, D., Andersen, M. E., Tyshenko, M. G., Krishnan, K., Hartung, T., Boekelheide, K., Wambaugh, J. F., Jones, D., Whelan, M., Thomas, R., Yauk, C., Barton-Maclaren, T., & Cote, I. (2019). Toxicity testing in the 21st century: progress in the past decade and future perspectives. *Archives of Toxicology*. <https://doi.org/10.1007/s00204-019-02613-4>
- Langfelder, P., Luo, R., Oldham, M. C., & Horvath, S. (2011). Is my network module preserved and reproducible? *PLoS Computational Biology*, 7(1), 1001057. <https://doi.org/10.1371/journal.pcbi.1001057>
- Li, H., Herrmann, T., Seeßle, J., Liebisch, G., Merle, U., Stremmel, W., & Chamulitrat, W. (2022). Role of fatty acid transport protein 4 in metabolic tissues: insights into obesity and fatty liver disease. *Bioscience Reports*, 42(6). <https://doi.org/10.1042/BSR20211854/231317>

- Malhotra, P., Gill, R. K., Saksena, S., & Alrefai, W. A. (2020). Disturbances in Cholesterol Homeostasis and Non-alcoholic Fatty Liver Diseases. *Frontiers in Medicine*, 7. <https://doi.org/10.3389/FMED.2020.00467>
- Martens, M., Evelo, C. T., & Willighagen, E. L. (2022). Providing Adverse Outcome Pathways from the AOP-Wiki in a Semantic Web Format to Increase Usability and Accessibility of the Content. *Applied In Vitro Toxicology*, 8(1), 2–13. https://doi.org/10.1089/AIVT.2021.0010/ASSET/IMAGES/LARGE/AIVT.2021.0010_FIGURE2.JPEG
- Moayedfard, Z., Sani, F., Alizadeh, A., Bagheri Lankarani, K., Zarei, M., & Azarpira, N. (2022). The role of the immune system in the pathogenesis of NAFLD and potential therapeutic impacts of mesenchymal stem cell-derived extracellular vesicles. *Stem Cell Research & Therapy*, 13(1). <https://doi.org/10.1186/S13287-022-02929-6>
- Pittman, M. E., Edwards, S. W., Ives, C., & Mortensen, H. M. (2018). AOP-DB: A database resource for the exploration of Adverse Outcome Pathways through integrated association networks. *Toxicology and Applied Pharmacology*, 343, 71–83. <https://doi.org/10.1016/J.TAAP.2018.02.006>

Supplemental data

Supplemental tables

Supplemental Table 1: Biostudies accession numbers. APAP: acetaminophen; ALF: acute liver failure; HVB: hepatitis B virus; HCV: hepatitis C virus

Pathology	Accession number
Transplanted-reperfused	E-GEOD-14951
Type II diabetes mellitus	E-GEOD-23343
Non-alcoholic steatohepatitis	E-GEOD-37031
Advanced non-alcoholic fatty liver disease	E-GEOD-49541
Non-alcoholic steatohepatitis, steatosis	E-GEOD-63067
APAP-induced ALF	E-GEOD-74000
HBV-infected liver disease	E-GEOD-83148
HBV-infected liver disease	E-GEOD-96851
HBV, HCV and haemochromatosis background cirrosis	E-MTAB-950
HCV-infected liver disease	E-GEOD-6764